République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université de Batna 2 - (Mostefa Ben Boulaïd)
Faculté des Mathématiques et Informatique
Département d'Informatique

Thèse

En vue de l'obtention du diplôme de

Doctorat en Informatique

Présentée Par

Tarek Amine HADDAD

Une approche coopérative basée sur l'IoT pour améliorer la qualité du trafic routier

Soutenue le 21 Décembre 2023

Devant le jury composé de :

Président	Saber BENHARZALLAH	***Prof.**, Université de Batna 2*
Rapporteur	Djalal HEDJAZI	***Prof.**, Université de Batna 2*
Examinateurs	Lyamine GUEZOULI	***MCA.**, Ecole Nationale Supérieure REESD-Batna*
	Toufik Messaoud MAAROUK	***MCA.**, Université de Khenchela*
	Rafik MAHDAOUI	***MCA.**, Université de Khenchela*
Invité	Sofiane AOUAG	***Prof.**, Université de Batna 2*

Remerciements

Tout d'abord, Je suis avant tout reconnaissant au *Tout-Puissant Allah* de m'avoir accordé le courage et la persévérance, tout ce que je suis aujourd'hui est juste parce que Lui *Allah Subhana-Wa-Taala*.

Je tiens à exprimer ma profonde gratitude et mes remerciements à mon directeur de thèse, **Pr. Djalal Hedjazi** pour ses précieux conseils, ses orientations, ses encouragements, ses expertises et son soutien tout au long de mon travail doctoral. Ses commentaires perspicaces m'ont orienté dans la bonne direction.

Je voudrais également exprimer ma gratitude aux honorables membres du jury pour leur volonté de lire attentivement ma thèse et d'avoir accepté la demande de nous rejoindre et de contribuer à l'évaluation de thèse. Merci à :

- **Pr. Saber Benharzallah,** Professeur à l'université de Batna 2 ;
- **Dr. Lyamine Guezouli,** Maître de conférences -A- à l'Ecole Nationale Supérieure REESD-Batna ;
- **Dr. Toufik Messaoud Maarouk,** Maître de conférences -A- à l'université de khenchela ;
- **Dr. Rafik Mahdaoui,** Maître de conférences -A- à l'université de khenchela ;
- **Pr. Sofiane Aouag,** Professeur à l'université de Batna 2.

Je suis très reconnaissant à toute ma famille pour leurs dons et leur soutien innombrables, infinis et sans demande.

Pour être sûre de n'oublier personne, J'adresse également mes sentiments de gratitude et de remerciements à tous ceux qui ont contribué de près ou de loin, directement ou indirectement à la réalisation de ce travail.

Résumé

De nos jours, les automobiles sont devenues très utiles pour les déplacements quotidiens aussi bien des individus que des marchandises. Cependant, l'accroissement de leur nombre a généré une augmentation importante des demandes d'utilisation des réseaux routiers, ce qui peut entraîner des retards, des embouteillages et une mauvaise fluidité de la circulation du trafic, notamment dans les grandes villes, notamment les métropoles mondiales. Ce type de problème, souvent désigné par la notion de congestion routière, peut être causé par d'autres facteurs tels que les travaux routiers, les accidents, la gestion insuffisante de la circulation, etc. De nombreuses stratégies permettant de réduire la congestion routière ont été adoptées par les gouvernements et les organismes de transport. Ces stratégies comprennent la construction de nouvelles routes, la planification de voies de transport en commun supplémentaires, la gestion du trafic en temps réel, etc.

D'autre part, le contrôle des feux de circulation (*TSC* en anglais *Traffic Signal Control*) représente l'une des tendances modernes qui peut jouer un rôle primordial dans la gestion du trafic. Elle désigne également la gestion et la coordination des feux de signalisation sur les réseaux routiers afin de bien contrôler la circulation et améliorer la fluidité du trafic. Cela peut impliquer l'ajustement du temps de cycle des feux de signalisation en fonction du trafic réel pour optimiser certains paramètres de performance. Les systèmes de *TSC* modernes peuvent adopter des technologies comme les capteurs de circulation, les caméras de surveillance et les algorithmes de traitement des données pour optimiser la gestion du trafic.

L'apprentissage par renforcement (en anglais *RL : Reinforcement Learning*) constitue une approche intelligente parmi les plus adoptées par les systèmes adaptifs de *TSC* pour optimiser la gestion des feux de signalisation, en améliorant la fluidité du trafic. En effet, il est possible de former un système pour apprendre à ajuster les durées de cycle des feux de signalisation en fonction de l'état réel du trafic pour réduire la congestion routière.

Dans cette thèse, nous proposons plusieurs approches coopératives basées le *DRL* (*Deep Reinforcement Learning*) pour optimiser intelligemment la gestion des feux de signalisation dans un réseau routier à multiple intersections. Nous avons ainsi modélisé notre problème comme étant un système d'apprentissage par renforcement multi-agents (*MARL* en anglais *Multi-Agent Reinforcement Learning*). Ceci implique l'utilisation de plusieurs agents dont chacun peut apprendre à prendre des décisions en termes d'ajustement des durées de cycle des feux en fonction de la situation local du trafic. Ces décisions peuvent être synchronisées avec celles prises par les autres agents pour garantir un fonctionnement optimal de l'ensemble du système. Dans de telles approches, chaque agent peut recevoir de la part de ses voisins leurs états, actions et récompenses, en les combinant avec son propre état, action et récompense pour prendre les décisions adéquates.

Les résultats expérimentaux sous différents scénarios montrent que les approches proposées surpassent de nombreuses approches de pointe en termes de trois paramètres : temps d'attente moyen (en anglais *AWT : Average Waiting Time*), longueur moyenne de la file d'attente (en anglais *AQL* : *Average Queue Length*) et émission moyenne de CO_2 (en anglais *AEC : Average Emission* CO_2)

Mots-clés : la congestion, contrôle des feux de circulation, apprentissage par renforcement, plusieurs intersections.

Abstract

Nowadays, automobiles have become very useful for the daily transportation of both people and goods. However, the increase in their numbers has generated a significant rise in demands for the use of road networks, which can lead to delays, traffic congestion, and poor traffic flow, especially in large cities and global metropolises. This type of problem often referred to as road congestion, can be caused by other factors such as road works, accidents, insufficient traffic management, etc. Many strategies to reduce traffic congestion have been adopted by governments and transport agencies. These strategies include building new roads, planning additional public transport routes, real- time traffic management, etc.

On the other hand, traffic signal control (*TSC*) represents one of the modern trends that can play a crucial role in traffic management. It also refers to the management and coordination of traffic lights on road networks in order to effectively control traffic and improve traffic flow. This can involve adjusting the traffic light cycle time based on real traffic to optimize certain performance parameters. Modern *TSC* systems can adopt technologies such as traffic sensors, surveillance cameras and data processing algorithms to optimize traffic management.

Reinforcement learning (*RL*) is an intelligent approach widely adopted by adaptive *TSC* systems to optimize traffic signal management, enhancing traffic flow. Indeed, it is possible to train a system to learn how to adjust traffic light cycle times based on the state of real traffic in order to reduce road congestion.

In this thesis, we propose several cooperative approaches based on Deep Reinforcement Learning (*DRL*) to intelligently optimize the management of traffic lights in a road network with multiple intersections. We have thus modeled our problem as a multi-agent reinforcement learning system (*MARL*). This involves the use of multiple agents each of whom can learn to make decisions in terms of adjusting light cycle times to the local traffic situation, and these decisions can be synchronized with the decisions made by other agents to ensure optimal functioning of the entire system. In such approaches, each agent can receive from its neighbors their states, actions and rewards, combining them with its own state, action and reward to make the appropriate decisions.

Experimental results under different scenarios show that the proposed approaches outperform many state-of-the-art approaches in terms of three parameters: Average Waiting Time (*AWT*), Average Queue Length (*AQL*) and average CO_2 emission (*AEC*).

Keywords: congestion, traffic signal control, reinforcement learning, multiple intersections.

ملخص

في الوقت الحاضر، أصبحت السيارات مفيدة جدًا للحركة اليومية لكل من الأشخاص والبضائع. ومع ذلك ، أدت الزيادة في عددها إلى زيادة كبيرة في الطلب على استخدام شبكات الطرق ، مما قد يؤدي إلى التأخير والاختناقات المرورية وتدفق حركة المرور الضعيفة ، لا سيما في المدن الكبرى والعواصم العالمية. هذا النوع من المشاكل ، الذي يشار إليه غالبًا بمفهوم ازدحام الطرق ، يمكن أن يكون ناتجًا عن عوامل أخرى ، مثل أعمال الطرق ، والحوادث ، وإدارة المرور غير الكافية ، وما إلى ذلك. تم تبني العديد من الاستراتيجيات لتقليل الازدحام المروري من قبل الحكومات ووكالات النقل . تتضمن هذه الاستراتيجيات بناء طرق جديدة ، وتخطيط طرق نقل عام إضافية ، وإدارة حركة المرور في الوقت الفعلي ، وما إلى ذلك.

من ناحية أخرى، يمثل التحكم في إشارات المرور أحد الاتجاهات الحديثة التي يمكن أن تلعب دورًا مهمًا في إدارة حركة المرور. كما يشير إلى إدارة وتنسيق إشارات المرور على شبكات الطرق من أجل التحكم الفعال في حركة المرور وتحسين تدفق حركة المرور. قد يتضمن ذلك ضبط وقت دورة إشارة المرور استنادًا إلى حركة المرور الفعلية لتحسين معايير أداء معينة. يمكن لأنظمة التحكم في إشارات المرور الحديثة اعتماد تقنيات مثل أجهزة استشعار حركة المرور وكاميرات المراقبة وخوارزميات معالجة البيانات لتحسين إدارة حركة المرور.

التعلم المعزز هو نهج ذكي من بين أكثر الأنظمة التكيفية التي تتبناها أنظمة التحكم في إشارات المرور لتحسين إدارة إشارات المرور ، وتحسين تدفق حركة المرور. في الواقع ، من الممكن تدريب نظام لمعرفة كيفية ضبط أوقات دورة إشارات المرور بناءً على ظروف المرور الفعلية لتقليل الازدحام المروري.

في هذه الأطروحة، نقترح العديد من الأساليب التعاونية القائمة على التعلم المعزز العميق لتحسين إدارة إشارات المرور بذكاء في شبكة طرق ذات تقاطعات متعددة. لذلك قمنا بصياغة مشكلتنا كنظام تعلم معزز متعدد العوامل. يتضمن ذلك استخدام عدة وكلاء ، يمكن لكل منهم تعلم اتخاذ القرارات من حيث ضبط أوقات دورة إشارات المرور بناءً على حالة المرور المحلية ، ويمكن مزامنة هذه القرارات مع القرارات التي يتخذها وكلاء آخرون لضمان الأداء الأمثل للنظام بأكمله. في مثل هذه الأساليب، يمكن لكل وكيل أن يتلقى من جيرانه حالاتهم وإجراءاتهم ومكافآتهم، ويجمعها مع حالته وإجراءاته ومكافأته الخاصة لاتخاذ القرارات المناسبة.

تظهر النتائج التجريبية في ظل سيناريوهات مختلفة أن الأساليب المقترحة تتفوق على العديد من الأساليب الحديثة من حيث ثلاث معايير: متوسط وقت الانتظار ومتوسط طول قائمة الانتظار ومتوسط انبعاثات ثاني أكسيد الكربون.

كلمات البحث : الازدحام ، التحكم في إشارات المرور ، التعلم المعزز ، التقاطعات المتعددة.

Table des Matières

Liste des figures

Liste des tables

Liste des Algorithmes

Liste des abréviations et des sigles

Adam	ADAptive Moment Estimation
AI	Artificial Intelligence
ATSC	Adaptive Traffic Signal Control
DDQN	Double Deep Q-Network
DP	Deep Learning
DQN	Deep Q-Network
DRL	Deep Reinforcement Learning
DNN	Deep Neural Network
FTSC	Fixed Traffic Signal Control
GAT	Graph Attention Network
HDP	Hierarchical Decision Process
IoT	Internet of Things
IT	Information Technologies
ITS	Intelligent Transportation Systems
ITSCP	Intersection Traffic Signal Control Problem
MARL	Multi-Agent Reinforcement Learning
MAS	Multi-Agent System
MDP	Markov Decision Process
ML	Machine Learning
MMDP	Multi-agent Markov Decision Process
MSE	Mean Squared Error
ReLU	Rectified Linear Unit
RL	Reinforcement Learning
SGD	Stochastic Gradient Descent
SUMO	Simulation of Urban MObility
TL	Traffic Light
TSC	Traffic Signal Control
UCB	Upper Confidence Bound
VII	Vehicle Infrastructure Integration
V2V	Vehicle to Vehicle

Introduction générale

1. Introduction

Depuis plusieurs décennies, le phénomène de la congestion du trafic routier devient l'un des problèmes les plus importants caractérisant la majorité des réseaux routiers des grandes villes et métropoles mondiales. Cela conduit non seulement à augmenter le temps des déplacements et à réduire les conditions de sécurité, mais aussi à exacerber le bruit et la pollution environnementale. D'autre part, la vie économique est sensiblement influencée par ce phénomène en générant des surcoûts financiers liés à la surconsommation du carburant et à la dégradation des véhicules et surtout à la perte du temps de travail. En effet, pour l'une des régions canadiennes les plus peuplées, la région du grand Toronto (*GTA : Greater Toronto Area*), le coût de la congestion a été estimé à environ 5,5 milliards de dollars pour l'année 2006 (Mekky, 2007). En 2014, les embouteillages ont coûté aux Américains plus de 160 milliards de dollars en perte de productivité en gaspillant plus de 3,1 milliards de gallons de carburant («The Economist», 2022). En 2018, les États-Unis déclarent un coût global de 87 milliards de dollars causé de la congestion routière selon (CNBC : Consumer News and Business Channel) («CNBC», 2022).

Par ailleurs, les zones urbaines se sont élargies, de manière continue, au fil des siècles ainsi que la majorité d'entre elles n'ont pas été développées en prévision de l'invention des moyens de transport. De plus, le renforcement du réseau routier par de nouvelles routes n'est pas toujours faisable et l'augmentation du parc automobile s'amplifier de façon plus rapide que celle de la population. Par exemple, le parc automobile de

l'Algérie comptait plus de 6,5 millions de véhicules à la fin de l'année 2019, contre plus de 6,4 millions en 2018, soit une hausse de 2,47% (selon *ONS*[1]). Ceci explique l'augmentation de la demande de trafic routier, environ 85% des mouvements de voyageurs se font par la route et 90% du volume des échanges (transport intérieur de marchandises hors transit) est réalisé par le transport routier (Bouriachi et al., 2021). C'est ainsi que la congestion routière, qui ne cesse de s'accroitre, reste un sérieux problème attirant l'attention d'aussi bien les chercheurs que les autorités afin d'imaginer des stratégies adéquates permettant d'atténuer ses conséquences. En effet, de nombreuses stratégies peuvent être considérées dans ce contexte à savoir : La mise en place de nouvelles infrastructures qui est généralement très difficile et trop coûteuse à concrétiser, ainsi qu'elle ne peut pas offrir une solution radicale au problème pour des raisons économiques et environnementales ou plus clairement du fait de limitations de l'espace disponible. Pareillement, minimiser la sollicitation du réseau, en encourageant les usagers à exploiter les transports en communs, n'est pas chose facile et peut durer longtemps et requérir des coûts supplémentaires importants. Toutefois, le management ou l'exploitation bien ordonnée de la sollicitation du trafic reste actuellement la solution la plus efficace et la plus facile à mettre en œuvre.

2. Contexte de travail

Le travail de recherche, décrit dans cette thèse, s'inscrit dans le contexte de la gestion de la congestion routière, plus particulièrement congestion urbaine. La congestion urbaine se manifeste généralement aux intersections. L'intersection, constitue, sans aucun doute, le composant le plus important et complexe du réseau routier urbain. Elle se caractérise par des agglomérations de population adjacentes relativement plus élevées et un flux de trafic complexe, entraînant des points de conflit de trafic plus élevés, des retards et une pollution accrue. Le coût des retards aux intersections est considéré comme l'élément clé de l'évaluation économique de la congestion du trafic.

En fait, notre défi majeur, dans ce contexte, est de proposer une solution intelligente permettant de gérer de manière dynamique les feux de signalisations associées à ensemble de d'intersections afin de faire face au problème de la congestion routière urbaine.

[1] **ONS : Office National des Statistiques**

3. Problématique de recherche

Le problème de contrôle des feux de circulation aux intersections baptisé en anglais *Intersection Traffic Signal Control Problem* (*ITSCP*) est devenu un sérieux problème nécessitant de sérieuses solutions, en particulier pour les zones métropolitaines (Haddad et al., 2022a). Dans cette perspective, une méthode efficace de contrôle de tels feux est requise de toute urgence lorsque plusieurs véhicules tentent d'utiliser une infrastructure de transport commune avec une capacité limitée. L'*ITSCP* a été résolu au début en utilisant des contrôleurs à base de valeurs fixes pour les différentes durées de feux de signalisation. Dans ce type de contrôleur nommé en anglais *Fixed Traffic Signal Control* (*FTSC*), les feux de signalisation changent entre le vert, le jaune et le rouge selon un schéma de synchronisation fixe. Néanmoins, Ce type d'approches ne peut pas répondre parfaitement aux changements des besoins du trafic. De plus, l'avancement perpétuel des technologies de l'information (*IT : Information Technology*) et de l'intelligence artificielle (*AI : Artificial Intelligence*), qu'a connu le monde actuel a également favorisé le développement des systèmes de transport intelligent (*ITS : Intelligent Transport System*). Par conséquent, la recherche sur l'*ITS* devienne l'une des thématiques les plus importantes et omniprésentes dans la majorité des manifestations scientifiques. Dans ce contexte, les travaux de recherche proposés ont joués un rôle primordial afin de proposer des approches intelligentes ayant la capacité de percevoir, de raisonner et d'agir en vue de fournir des services améliorant la qualité de vie des personnes particulièrement lorsqu'il s'agit du trafic routier. C'est-à-dire que l'approche intelligente implémentée par un contrôleur de signaux s'adapte à l'environnement de trafic, en utilisant les données de trafic reçues. De telles approches, autrement dites les méthodes adaptives (*ATSC : Adaptive Traffic Signal Control*), sont plus adoptées que les méthodes fixes du fait qu'elles ne demandent pas beaucoup d'investissements à l'exception de certains simples moyens technologiques permettant d'agir de manière dynamique sur les éléments du trafic. Dans cette dialectique, l'utilisation du réseau routier existant peut devenir meilleure et le système de contrôle peut préserver un certain niveau de performance. En collectant, de manière continue, des données du trafic à partir de plusieurs capteurs déployés à différents endroits du réseau routier, un système basé *ATSC* vise à réduire à la fois la congestion potentielle et les émissions de dioxyde de carbone, tout en ajustant dynamiquement la synchronisation des signaux (Aslani et al., 2019). De nombreux systèmes adaptifs ont été proposés dans la littérature à savoir : *Split Cycle Offset Optimisation Technique* (*SCOOT*) (Robertson and Bretherton, 1991), *Sydney Coordinated Adaptive Traffic System* (*SCATS*) (Sims and Dobinson, 1980), *Real Time Hierarchical Optimized Distributed Effective System (RHODES)* (Head et al., 1992), *Optimized Policies for Adaptive Control (OPAC)* (Gartner, 1983), etc. Ces systèmes

implémentent généralement des algorithmes intelligents se basant sur les algorithmes évolutionnaires (Montana and Czerwinski, 1996), les algorithmes génétiques (Guo et al., 2019), la programmation dynamique (Yousef et al., 2010), les systèmes multi-agents (Eom and Kim, 2020; Zhang and Zhang, 2020), la théorie des jeux (Nam Bui and Jung, 2018), l'apprentissage automatique (Eom and Kim, 2020), etc.

Parmi les techniques d'*IA* les plus populaires et que beaucoup de recherches adoptent pour mettre en œuvre des solutions adaptives au *ITSCP,* nous distinguons l'apprentissage par renforcement (en anglais *RL : Reinforcement Learning*). Il s'agit d'une méthode qui fait référence à une classe de méthodes d'apprentissage automatique (en anglais *ML* : *Machine Learning*), dont le but de rendre le système *plus* capable d'apprendre, à partir de ses expériences antérieures via une méthode de récompense ou de pénalité. Le *RL* a largement prouvé ses performances exceptionnelles en résolvant des problèmes complexes (Shamsi et al., 2022), lui ont permis d'occuper une place importante dans le domaine de *ML* (Wang et al., 2021b). D'autre part, la naissance des méthodes d'apprentissage par renforcement profond (en anglais *DRL : Deep Reinforcement Learning*), qui ont été bien appliquées dans le contexte *ATSC*, a particulièrement jouée un rôle important en améliorant les performances des *STI* (Genders and Razavi, 2018). La méthode *DRL* peut donc produire des résultats suffisamment efficaces sans avoir besoin d'une compréhension analytique explicite ou d'une modélisation de la dynamique du trafic.

Par ailleurs, dans la littérature du contexte *ATSC*, de nombreux efforts croissants ont été étudiés pour proposer des solutions distribuées, que la majorité d'entre eux considèrent le système multi-agents (*MAS*) comme une approche de solution puissante (Eom and Kim, 2020; Zhang et al., 2022). Par ailleurs, modéliser le problème *ATSC* sous forme de *MAS* intégrant *DRL* (appelé *MARL : Multi-Agent Reinforcement Learning*) semble également très important dans la mesure où il permet aux agents d'apprendre de l'environnement de manière indépendante (chaque agent exécute un algorithme *DRL* sans échanger d'informations avec d'autres agents) ou en coopération (chaque agent prend sa propre décision en fonction des informations de lui-même et des autres agents) (Wang et al., 2021a).

4. Objectif de recherche et contributions

La présente thèse de Doctorat cible également le problème *ITSCP* pour plusieurs intersections en proposant une approche coopérative adaptive basée sur *MARL.* Pour atteindre un tel objectif, plusieurs solutions sont proposées de manière graduelle, en commençant par la résolution du problème pour une intersection isolée en arrivant à la

résolution du problème pour un réseau d'intersections adjacentes. Les différentes propositions considèrent toute intersection comme un agent *DRL*. En assurant la coopération entre les agents, les approches proposées permettent aux agents de partager leurs décisions et leurs observations les uns avec les autres et les différents agents se comportent comme un groupe synergique, plutôt qu'un ensemble d'individus.

Bien que l'algorithme d'apprentissage *Q-learning* a été considéré comme l'un des algorithmes *RL* les plus populaires et les plus utilisés (Wang et al., 2020), il a été largement appliqué pour résoudre le problème *ITSCP* dans de nombreuses recherches (Eom and Kim, 2020). De plus, la majorité des méthodes basées sur *MARL,* proposées dans littérature, se sont principalement concentrées sur le *Q-learning* pour distribuer la fonction *Q* aux agents. Par conséquent, notre contribution applique *Q-learning* pour permettre à un agent de sélectionner des actions adéquates en recevant de manière récurrente les récompenses associées à l'environnement. L'aspect coopération se manifeste dans cette contribution par l'échange de valeurs d'états, d'actions et de récompenses entre les agents adjacents.

Pour mesurer les performances de nos propositions, des expérimentations ont été menées pour plusieurs circonstances de flux de trafic en utilisant le simulateur de trafic microscopique *SUMO* (*Simulation of Urban MObility*) (Krizhevsky et al., 2017). Les résultats sont comparés à des algorithmes de la littérature à savoir : *QT-CDQN* (Ge et al., 2019), *MADRL* (Liu et al., 2017) et *CODRL* (Hussain et al., 2020).

Ainsi, les contributions apportées dans cette thèse peuvent être résumées en ce qui suit :

- Tout d'abord, nous proposons une approche dynamique de contrôle des feux de signalisation pour une intersection isolée. Cette approche a pour objectif l'amélioration de la qualité du trafic routier en décidant dynamiquement la séquence des phases et les durées de feu vert. Toute décision est prise en fonction du temps d'attente moyen (*AWT*) et du nombre de véhicules dans les files d'attente des différentes phases. Ainsi, la priorité est donnée à la phase avec les valeurs les plus élevées de ces deux paramètres. Une évaluation des performances de cette nouvelle approche est également proposée en développant une simulation de trafic via la plateforme *NetLogo*. Des comparaisons sont, en outre, faites avec d'autres algorithmes adaptifs proposés dans (Rida and Hasbi, 2019; Yousef et al., 2010) ainsi qu'avec l'algorithme classique qui fixe les durées des feux tout au long de l'utilisation du réseau routier.
- Ensuite, nous proposons une nouvelle approche basée sur le *DRL* afin de contrôler une intersection isolée. Le contrôleur des feux est modélisé par un agent

intelligent qui perçoit l'état du trafic. Cette contribution adopte l'algorithme *Double Deep Q-Network (DDQN)* pour apprendre et optimiser le comportement du contrôleur des feux de signalisation en fonction des conditions de trafic actuelles. Cela fait que l'idée d'avoir une formule simplifiée d'état et de récompense facilite la formation de l'agent en simplifiant la convergence de ces derniers. Il sélectionne dynamiquement la séquence des phases pour objectif améliorant la qualité du trafic.

- Enfin, l'approche précédente est également améliorée afin qu'elle soit capable de gérer un réseau d'intersections. Ceci nous a permet de proposer une nouvelle solution coopérative. Elle est basée sur l'apprentissage par renforcement multi-agents (*MARL*) qui considère chaque intersection comme un agent et le groupe d'agents coopèrent afin d'ajuster les durées des différents feux de signalisation tout en optimisant certaines fonctions objectifs. L'aspect coopération se manifeste dans cette approche par l'échange de valeurs d'états, d'actions et de récompenses entre agents adjacents.
- La simulation de toutes ces propositions en utilisant *NetLogo* et *SUMO* montre nos contributions avec des résultats prometteurs.

5. Organisation de la thèse

Cette thèse est structurée en cinq chapitres. Les deux chapitres, qui suivent la présente introduction générale, décrivent l'état de l'art sur aussi bien les systèmes de transport intelligents que les méthodes d'apprentissage par renforcement appliquées au contrôle des feux de signalisation, ensuite les derniers chapitres présentent respectivement les approches proposées, les résultats obtenus et discussions.

Le chapitre 1, intitulé « *Généralités sur la théorie du trafic routier*», introduit, en premier lieu, les différents concepts fondamentaux liés à notre domaine de recherche qui la gestion et contrôle du trafic routier. Il présente, en deuxième lieu, un état de l'art sur les systèmes, les modèles et les méthodes de contrôle des feux de circulation.

Le chapitre 2, intitulé « *Apprentissage par renforcement pour le contrôle du trafic routier*», présente les bases de l'apprentissage par renforcement sur laquelle s'appuient nos contributions décrites dans cette thèse. Il décrit, ensuite, un état de l'art sur l'apprentissage par renforcement appliqué à l'optimisation du trafic routier.

Le chapitre 3, intitulé «*Approche réactive pour le contrôle adaptatif des feux de signalisation dans les intersections isolées*», nous proposons et étudions un algorithme dynamique qui permet de contrôler les feux des signalisations pour une seule intersection. Une

évaluation des performances de cette nouvelle approche est également présentée pour démontrer l'efficacité de la méthode proposée.

Le chapitre 4, intitulé « *Approche intelligente basée DRL pour le contrôle adaptatif des feux de signalisation dans les intersections isolées* », décrit l'approche proposé basée *DRL* pour le contrôle de feux de signalisation à une intersection isolée et présente des résultats expérimentaux pour démontrer l'efficacité de la méthode proposée.

Le chapitre 5, intitulé « *Approche intelligente basée MARL pour le contrôle adaptatif des feux de signalisation dans les réseaux à intersections multiples* », décrit l'approche coopératif de contrôle des feux de signalisation proposé qui basé sur *DRL* pour plusieurs intersections et présente des résultats expérimentaux pour démontrer l'efficacité dans un réseau routier.

Enfin, la dernière partie fournit la conclusion et les perspectives, y compris les apports et les limites du travail.

Chapitre 1 : Généralités sur la théorie du trafic routier

1.1 Introduction

Le trafic routier se définit comme l'ensemble des phénomènes complexes qui résultent du déplacement d'usagers sur un réseau routier à capacité limitée. Ces phénomènes sont étudiés, depuis plusieurs années, pour permettre principalement le contrôle de l'évolution du trafic au cours du temps (Ardekani and Herman, 1987; Hall, 1997; Rothrock and Keefer, 1957), tout en mesurant les performances de la circulation, notamment au niveau des intersections (Greenshields et al., 1935). C'est ainsi le recours à des théories semble assez important afin de tirer les bonnes explications de ces phénomènes.

L'origine du terme **théorie** vient du mot grec « *theoria* » qui veut dire « *contempler, observer, examiner* ». Une théorie représente alors un ensemble cohérent d'explications, de notions ou d'idées sur une thématique précise, pouvant inclure des lois et des hypothèses, induites par l'accumulation de faits provenant de l'observation, l'expérimentation ou, dans le cas des mathématiques, déduites d'une base axiomatique donnée («Théorie», 2023; Daston, 2020). C'est sur la base de la théorie que les gens peuvent faire des prédictions sur ce qu'ils comptent observer et étudier. En outre, une

théorie scientifique doit être en mesure de respecter pas mal de critères à savoir la correspondance entre les principes théoriques et les phénomènes étudiés.

Dans cette thèse, les théories sont considérées comme le fondement de la science du trafic routier. Elles regroupent une variété de lois mathématiques, physiques et des systèmes informatiques permettant de donner une compréhension profonde des phénomènes en relation avec la circulation des véhicules sur le réseau routier ainsi que leurs interactions avec l'environnement. La robustesse de telles théories est également conditionnée par leurs capacités à fournir des explications objectives aux évolutions, au cours du temps, des conditions de circulations dans des zones de route (Papageorgiou et al., 2009; Greenshields et al., 1935; Adams, 1936)

Dans ce chapitre, nous présentons de manière générale le domaine de recherche pluridisciplinaire qui étudie les mouvements des véhicules sur les réseaux routiers. Nous exposons en premier lieu un bref historique sur le domaine de la théorie du trafic. Ensuite, nous discutons les modèles de la circulation routière. Après, nous expliquons les différentes variables élémentaires du trafic. Ces dernières sont des mesures quantitatives représentant les différents aspects de la circulation sur les réseaux routiers. Enfin, nous discernons le principe de la gestion des feux de signalisation aussi bien dans les intersections isolées que multiples.

1.2 Historique

La naissance de la théorie du trafic routier date des années 30 suite aux recherches menées par Greenshields et al., (1935), qui a proposé un modèle mathématique modélisant le trafic sur des autoroutes Américaines. Ce modèle, qui est basé sur des observations de l'état du trafic, est capable de mesurer le débit, la vitesse et la densité du trafic en utilisant des capteurs photographiques.

Ensuite, l'intérêt à cette thématique est en forte croissance, notamment après la deuxième guerre mondiale qui a connu une augmentation du besoin d'utilisation de l'automobile. En effet, plusieurs rencontres scientifiques internationales ont été organisées à savoir :

- *Symposium sur la théorie du flux de trafic, Détroit, Michigan, 7-8 Décembre 1959,*
- *Deuxième colloque international sur la Théorie du flux de Trafic routier, Londres, Angleterre, 25-27 juin 1963.*
- *Troisième colloque international sur la théorie du flux de trafic routier, New York, Juin 1965,*

- *Quatrième colloque international sur la théorie du flux de trafic routier, Karlsruhe, Allemagne, Juin 18-20, 1968.*
- *Cinquième symposium international sur la théorie de trafic et des transports, Berkeley, Californie, 16-18 Juin, 1971.*
- *Sixième symposium international sur le transport et la théorie du trafic, Sydney, Australie, 26-28 Août 1974.*

En outre, plusieurs journaux scientifiques, abordant la thématique de recherche du transport et gestion du trafic routier, ont été proposés entre les années 58 et 72. Ceci explique l'intérêt important que la communauté scientifique a attribué à cette problématique. Parmi les principaux journaux créés dans cette période, nous citons :

- *Traffic Engineering and Control, Editeur: Printerhall Ltd., 29, rue Newman, Londres W1P3PE, Angleterre. (créé en 1958).*
- *Transportation Science, Section de la science des transports, Editeur: Société de recherche opérationnelle d'Amérique, 428 East Preston Street, Baltimore, Maryland 21202. (créé en 1967).*
- *Transportation Research Elsevier, Editeur: Pergamon Press Inc., Maxwell House, Fairview Park, Elmsford, New York 10523. (créé en 1967).*
- *Transportation Planning and Technology. Editeur: Taylor francis, New York, New. (créé en 972).*
- *Transportation, Editeur : Springer US. (créé en 1972).*

Au milieu des années 1950, les modèles de trafic avaient attiré l'attention de nombreux scientifiques renommés. Dans cette époque, aurait été le commencement de la théorie des flux de trafic (Maerivoet and De Moor, 2005; Pipes, 1953). Par conséquent, des chercheurs scientifiques de nombreux horizons ont tenté de modéliser le mouvement de la circulation, dans le but extrême de trouver des améliorations aux problèmes de circulation. Certaines des premières contributions à la modélisation du trafic ont été celles de Reuschel (1950) et Pipes (1953), d'une part, et de Lighthill and Whitham (1955), d'autre part. *Reuschel* et *Pipes* ont proposé un modèle (*microscopique*) de trafic décrivant le déplacement détaillé des voitures circulant à proximité les unes des autres sur une seule voie. *Lighthill*, un théoricien de la mécanique des fluides de renommée mondiale, avec *Whitham*, ont proposé un modèle (*macroscopique*) de trafic, modélisant le trafic comme un continuum assimilable à un fluide.

Les travaux sur la modélisation (*microscopique* et *macroscopique*) ont ouvert de nouvelles directions de recherches dans le domaine de la théorie du trafic. Ces travaux portent sur la modélisation des interactions des mouvements du trafic dans le but de les

améliorer par le biais d'un contrôle approprié. Ce contrôle s'effectue au moyen de dispositifs de dispositifs de régulation tels que les feux de signalisation, ainsi que d'autres moyens de régulation du débit du trafic (panneaux messages variables, dispositifs de coordination, stations de comptage, capteurs routiers, ...).

1.3 Concepts fondamentaux

1.3.1 Variables élémentaires du trafic routier

Les variables élémentaires du trafic routier sont des mesures utilisées pour décrire le comportement et la performance des véhicules et des usagers de la route. Elles sont importantes pour la planification, la conception et la gestion des routes et des intersections, ainsi que pour l'évaluation de la performance des systèmes de transport. Les ingénieurs et les planificateurs de transport travaillent sur ces variables pour proposer de meilleures conceptions des routes et des intersections, gérer la circulation, améliorer la sécurité et la qualité de service pour les usagers de la route. Les variables élémentaires comprennent entre autres :

1.3.1.1 Débit de trafic

Le débit de trafic est une mesure importante dans la gestion du trafic routier. Il représente le nombre de véhicules qui passent par un point donné sur le réseau routier en une heure (*une journée, un mois, une année, etc.*), généralement exprimé en véhicules par heure.

L'équation (1.1), définie le débit moyen q (t_1, t_2, x) au point d'abscisse x entre les deux instants t_1 et t_2.

$$q\ (t_1, t_2, x) = \frac{n(x, \Delta t, t)}{\Delta t} \tag{1.1}$$

Où

$N(x, \Delta t, t)$, désigne le nombre de véhicules passés par le point x entre les deux instants t et $t+\Delta t$ (Figure 1.1). Dans certaines théories, le flot des véhicules est parfois considéré comme continu ainsi qu'il est évalué par l'équation (1.2) suivante :

$$q(x,t) = \lim_{\Delta t \to 0} q(t - \frac{\Delta t}{2}, t + \frac{\Delta t}{2}, x) \tag{1.2}$$

Une telle définition n'est pas applicable en ces termes à une théorie des flux de circulation discrets puisque cette limite fluctuerait entre l'infini et zéro selon qu'un véhicule était présent ou non à ce moment *t*. Il faut noter qu'il préférable de

considérer une petite valeur de Δt (de l'ordre de quelques secondes par exemple). Cette valeur de Δt désigne qu'il y une identité entre $q(x,t)$ et $q(t-\frac{\Delta t}{2}, t+\frac{\Delta t}{2}, x)$.

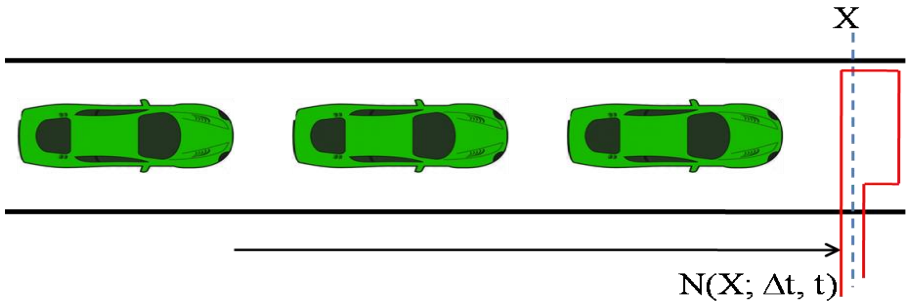


Figure 1.1 Débit de trafic passant par le point X.

Vu que le débit de trafic est une variable continue qui requiert un intervalle de temps précis pour être bien estimée ainsi qu'il y a toujours des écarts importants entre la valeur estimée et la valeur réelle. En effet, il est souvent pénible d'exploiter le potentiel des infrastructures des réseaux routiers à un degré maximal.

Néanmoins, l'avancement considérable qu'a connu le monde actuel, particulièrement en termes de technologies de l'information (en anglais *IT* : *Information Technologies*), nous a offert plusieurs opportunités, notamment pour faire face au problème d'estimation de débit de trafic. Les technologies à base d'*IoT* ont permis aux systèmes de gestion du trafic de mesurer, de manière précise, et à n'importe quel moment la valeur du débit de trafic.

Par ailleurs, les technologies de communications sans fil (comme : *WiFi, WiMax et 3G*) ont simplifiées les échanges inter-véhicules (*V2V : Vehicle to Vehicle*) et les échanges entre véhicules et infrastructures (*VII : Vehicle Infrastructure Integration*), ce qui rend la gestion du trafic plus rationnelle en agissant sur des données réelles du trafic.

1.3.1.2 Densité de trafic

La densité, appelée aussi concentration, détermine le nombre de véhicules qui se trouvent sur une section de la route à un instant donné. La densité moyenne $k(x_1, x_2, t)$ à l'instant t sur une section de route limitée par les deux points d'abscisses x_1 et $x_2 = x1 + \Delta x$ se définit comme suit :

$$k(x_1, x_2, t) = \frac{n(x, \Delta x, t)}{\Delta x} \tag{1.3}$$

Avec $n(x, \Delta x, t)$ présente le nombre des véhicules présents sur la section à l'instant t (Figure 1.2). La densité est exprimée en nombre de véhicules par unité de longueur comme par exemple véhicules/mètre (*veh/m*).

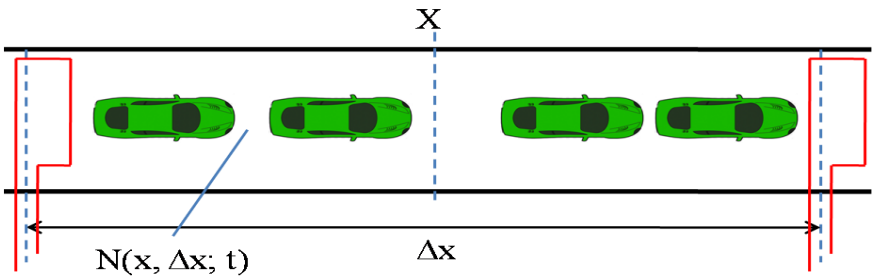


Figure 1.2 Densité de trafic.

1.3.1.3 Vitesse de circulation

La vitesse de circulation dans le trafic routier peut varier considérablement en fonction de plusieurs facteurs tels que la densité du trafic, les conditions météorologiques, les limitations de vitesse, les travaux routiers, les accidents de la route, etc. Généralement, la vitesse de circulation dans les réseaux urbains denses est inférieure à la vitesse limite, vu que les véhicules sont souvent ralentis par la circulation, les feux de signalisation, les arrêts et les embouteillages. D'autre part, sur les autoroutes et les routes à grande circulation, la vitesse de circulation peut être plus élevée, mais elle peut également varier en fonction de la densité du trafic et des conditions météorologiques.

La vitesse de circulation peut être vue comme un élément d'équilibre entre les contraintes de la circulation et les limites de sécurité imposées pour protéger les usagers de la route.

Par ailleurs, La vitesse moyenne du flux routier est la vitesse à laquelle les véhicules se déplacent sur une certaine distance d'une route. Nous distinguons deux types de vitesse moyenne à savoir : la vitesse moyenne temporelle et la vitesse moyenne spatiale.

La ***vitesse moyenne temporelle*** (vitesse moyenne du flot), notée $\boldsymbol{v}_{\text{moy\_t}}(\boldsymbol{t})$, représente la moyenne de toutes les vitesses de véhicules qui traversent en point x sur un intervalle de temps donné. Elle est mesurée par la formulation suivante :

$$v_{moy\_t}(t) = \frac{1}{N}\sum_{i=1}^{N} v_i \qquad (1.4)$$

Où N est le nombre de véhicules passant par un point. v_i est la vitesse du i^{ème} véhicule. L'unité utilisée pour quantifier $v_{moy\_t}(t)$ est généralement le *mètres/seconde* [*m/s*].

Quant à la ***vitesse moyenne spatiale*** **(*notée*** v_{moy_s}**)**, est une mesure de la vitesse à laquelle les véhicules se déplacent dans l'espace. Elle consiste en la vitesse moyenne

harmonique des véhicules passant par à un point pendant un intervalle de temps (Wardrop, 1952).

Pour calculer la v_{moy_s} sur une route, on peut mesurer la distance totale parcourue par tous les véhicules sur cette route pendant une certaine période de temps, puis diviser cette distance par le temps écoulé.

$$v_{moy\_s} = \frac{N}{\sum_{i=1}^{N} \frac{1}{v_i}} \tag{1.5}$$

La différence entre la vitesse moyenne temporelle et la vitesse moyenne spatiale peut être illustrée graphiquement par la courbe décrite sur la Figure 1.3. La v_{moy_t} (décrite en couleur rouge) est calculée sur la base des vitesses individuelles des différents véhicules traversant un point donné pendant un intervalle de temps Δt. La v_{moy_s} (décrite en couleur bleu) est calculée à partir des vitesses individuelles de tous les véhicules présents sur un segment de route de longueur Δx.

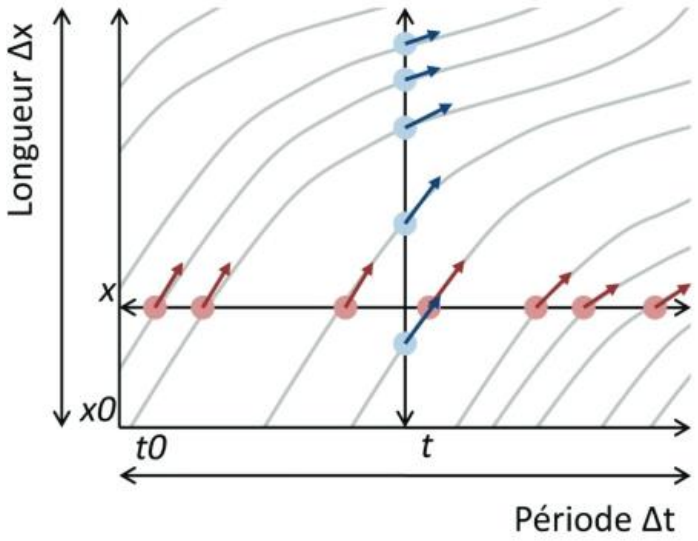


Figure 1.3 Différence entre vitesse moyenne temporelle et vitesse moyenne spatiale (Buisson and Lesort, 2010).

1.3.1.4 Temps de trajet

L'estimation du temps de trajet est un élément clé de la gestion du trafic routier. Les autorités de la circulation utilisent des données en temps réel pour estimer le temps de trajet sur les principales routes et autoroutes, ce qui leur permet de prendre des décisions éclairées sur la gestion du trafic.

L'estimation du temps de trajet peut être utilisée pour prédire les temps de voyage futurs et aider les conducteurs à éviter les embouteillages. Par exemple, les panneaux d'affichage électroniques sur les routes peuvent fournir des estimations de temps de trajet pour les différentes destinations, permettant aux conducteurs de choisir l'itinéraire le plus rapide en fonction des conditions actuelles de la circulation.

En outre, l'estimation du temps de trajet peut être utilisée pour ajuster les feux de circulation en temps réel et pour coordonner les travaux de réparation sur les routes afin de minimiser les perturbations du trafic.

En somme, l'estimation du temps de trajet est une composante essentielle de la gestion du trafic routier et permet de réduire les embouteillages, les temps de voyage et d'améliorer la sécurité routière.

Prenant en considération la longueur unitaire d'un segment de route donné, et $\boldsymbol{v_i}$ la vitesse du $i^{ième}$ véhicule. Soit t_i le temps que prend le véhicule pour traverser la distance unitaire alors t_i=1/v_i. S'il y a N véhicules de ce type sur la voie, le temps de trajet moyen t_{moyen} est donné par la formule suivante :

$$t_{moyen} = \frac{\sum t_i}{N} \tag{1.6}$$

1.3.1.5 Capacité de la route

La capacité de la route est la quantité maximale de véhicules qu'une route peut transporter dans des conditions régulières de circulation. Elle est généralement exprimée en termes de nombre de véhicules par heure et par voie. La capacité de la route dépend de nombreux facteurs tels que la largeur de la chaussée, le nombre de voies de circulation, les limites de vitesse, la gestion des feux de signalisation aux intersections, les conditions météorologiques, la densité du trafic, etc.

Ce type de mesure est généralement utilisé par les autorités afin de bien évaluer les niveaux de service de la route et pour planifier et concevoir des améliorations de la qualité du trafic. Lorsque la capacité de la route est dépassée, cela peut entraîner des congestions routières, des retards et des temps de trajet plus grands pour les conducteurs de véhicules.

La détermination de la capacité d'une route est un processus complexe qui prend en compte de nombreux facteurs. Plusieurs étapes générales sont souvent à considérer pour estimer la capacité d'une route entre autres :

- *Mesurer la largeur de la chaussée et le nombre de voies de circulation disponibles. Ceci permettra de déterminer la largeur disponible pour chaque véhicule et le nombre de voies pouvant être utilisées simultanément.*
- *Évaluer les limitations de la route telles que les intersections, les feux de signalisation, les ponts, les tunnels et autres points de congestion potentiels. Cela aidera à estimer la capacité réelle de la route en tenant compte des points où les véhicules doivent ralentir ou s'arrêter.*
- *Prendre en compte les caractéristiques de la circulation telles que la densité de trafic, les vitesses moyennes et les taux d'occupation pour chaque voie.*
- *Utiliser des modèles de capacité de la route pour estimer la capacité maximale de la route. Ces modèles prennent en compte les facteurs ci-dessus pour estimer la capacité maximale de la route en nombre de véhicules par heure et par voie.*

Il est important de noter que la capacité de la route peut varier en fonction de nombreux facteurs tels que l'heure de la journée, le jour de la semaine, les conditions météorologiques et d'autres événements spéciaux qui peuvent affecter la circulation. Par conséquent, les estimations de la capacité de la route doivent être mises à jour régulièrement pour refléter les conditions de circulation actuelles.

1.3.1.6 Taux d'occupation

Le taux d'occupation du trafic routier est une mesure qui indique le pourcentage de la capacité maximale d'une route ou d'une voie de circulation qui est utilisé par les véhicules pendant une période de temps donnée. Il est généralement exprimé en pourcentage de la capacité maximale de la route ou de la voie de circulation.

Cette mesure est utilisée pour estimer l'efficacité de la capacité de la route à répondre aux besoins du trafic routier. Un taux d'occupation élevé peut montrer une vigoureuse demande pour l'utilisation de la route, cependant ceci peut également entraîner des problèmes de congestion routière.

Pour mesurer le taux d'occupation, plusieurs méthodes peuvent être utilisées comme par exemple en menant des enquêtes sur le terrain, en utilisant des modèles de simulation du trafic, ou en utilisant des capteurs ponctuels appelés *boucles électromagnétiques.* Ces derniers sont souvent installés à des endroits bien déterminés sur la route pour capter ce qui se passe à ces endroits. C'est à partir de données collectées par ces capteurs que le taux d'occupation (noté TO) peut être calculé. Soit θ_i, le temps pendant lequel la boucle électromagnétique détecte un signal. Ce temps est étroitement lié à la vitesse v_i du $i^{ième}$ véhicule, à sa longueur L_i.

TO est exprimé en pourcentage par la formule suivante :

$$TO = \frac{100}{\Delta t} \times \sum_{i=1}^{N} \theta_i \tag{1.7}$$

D'autre part, le temps de séjour θ_i sur la boucle de l'i^{ième} véhicule est relié directement à la longueur L_i et à la vitesse de ce véhicule ainsi qu'à la longueur de la boucle (notée λ).

$$TO = \frac{\lambda + L_i}{v_i} \tag{1.8}$$

Il est utile de préciser que le taux d'occupation du trafic routier est une mesure dynamique qui peut varier en fonction de différents facteurs tels que la densité du trafic, les conditions météorologiques et la présence d'accidents ou de travaux routiers.

On peut utiliser le TO pour mesurer la concentration (notée K), qui désigne le nombre de véhicules présents à un instant donné sur une longueur de route bien déterminée. En effet, nous pouvons utiliser la formulation suivante :

$$TO = 100x\,(\lambda + L)K \tag{1.9}$$

Cette relation n'est exactement vraie que lorsque les longueurs des véhicules sont identiques (notée L).

1.3.2 Infrastructures de mesure

La collecte de données sur le trafic routier est une tâche extrêmement importante pour plusieurs raisons. Tout d'abord, pour la planification du transport dans la mesure où les données collectées sont utilisées pour programmer la construction de nouvelles routes, l'agrandissement des routes existantes et la création de nouvelles lignes de transport en commun. D'autre part, les gestionnaires du trafic exploitent de telles données pour gérer la circulation en temps réel. En outre, les autorités responsables de la gestion des réseaux routiers peuvent travailler sur les données collectées pour proposer des ajustements des feux de signalisation en fonction des heures d'encombrements, des modifications des itinéraires des transports publiques. L'amélioration des conditions de sécurité routière est une tâche très utile qui devrait étudier les données mesurées sur le trafic, notamment pour identifier les zones à haut risque et à développer des programmes pour réduire par exemple le nombre d'accidents.

Les infrastructures de mesure de variables élémentaires du trafic routier sont des dispositifs utilisés afin de collecter, de façon directe ou indirecte, des données sur le trafic routier (Klein et al., 2006; Mimbela and Klein, 2007). En effet, de nombreux

capteurs de trafic peuvent être exploités, chacun ayant des avantages et des inconvénients en termes de précision, de coût et d'installation. Parmi les capteurs les plus couramment utilisées dans le contexte du trafic routier sont :

- *Capteurs magnétiques* qui utilisent des champs magnétiques pour détecter la présence de véhicules en mouvement. Ils sont généralement installés sous la chaussée et sont relativement peu coûteux, mais peuvent être moins précis que d'autres types de capteurs.
- *Capteurs à induction* qui sont similaires aux capteurs magnétiques, ils utilisent des champs électromagnétiques pour détecter les véhicules. Ils sont également installés sous la chaussée, mais peuvent offrir une meilleure précision que les capteurs magnétiques.
- *Capteurs de pression* qui utilisent des cellules de charge pour mesurer la pression exercée sur la chaussée par les véhicules en mouvement. Ils peuvent être installés à la surface de la chaussée ou encastrés, et sont généralement plus précis que les capteurs magnétiques ou à induction.
- *Capteurs optiques* qui utilisent des caméras pour détecter la présence et le mouvement des véhicules. Ils peuvent être installés en hauteur ou sur le côté de la route, et peuvent offrir une précision élevée. Cependant, ils peuvent être plus coûteux que d'autres types de capteurs et nécessitent une maintenance régulière.
- *Capteurs acoustiques* qui utilisent des microphones pour détecter le bruit des véhicules en mouvement. Ils peuvent être installés à la surface de la chaussée ou encastrés, et peuvent offrir une précision raisonnable. Cependant, ils peuvent être sensibles aux bruits ambiants et peuvent nécessiter une calibration régulière.

Pour collecter des données plus complètes sur le trafic routier, les infrastructures suscitées peuvent être installées individuellement ou combinées ainsi que les données collectées peuvent être analysées pour assister les planificateurs de transport à prendre des décisions pertinentes sur la construction de routes, la gestion du trafic, l'amélioration de la sécurité routière.

1.4 Modélisation de la circulation routière

Une approche méthodologique qui s'est avérée être couronnée de succès pour l'étude des phénomènes complexes est l'approche systémique qui considère le système dans son ensemble, composé de composants interconnectés, complexes et fonctionnellement liés, pouvant être étudiés scientifiquement en utilisant une représentation formelle ou un modèle du système (Barceló, 2010). Au niveau le plus simple, un modèle est une représentation de quelque chose.

Dans le domaine de la théorie du trafic, la modélisation de la circulation routière semble assez importante. Elle vise principalement à comprendre et à prédire le

comportement des automobiles sur les réseaux routiers en utilisant des méthodes mathématiques et informatiques. Les modèles de circulation routière peuvent aider à estimer les impacts des projets d'infrastructure de transport sur la circulation du trafic, à planifier les moyens de transport publique, à optimiser les itinéraires de livraison de la marchandise, à améliorer la sécurité routière, etc.

Les modèles de flux de circulation peuvent être appuyés sur des données de recensement du trafic réel ou sur des approximations basées sur des algorithmes de simulation. Les algorithmes de simulation peuvent prendre en compte plusieurs facteurs en relation avec l'environnement, les politiques de transport, les vitesses de circulation des automobiles, les distances entre véhicules, les habitudes de conduite, etc., pour estimer le flux de circulation sur un réseau routier ou sur une partie de ce dernier.

Par ailleurs, la simulation du trafic est un outil crucial pour la prise de décision en matière de transport et le développement de politiques (Wang et al., 2023b). Elle a attiré l'attention de nombreux chercheurs dans le domaine du transport, ainsi que divers modèles de simulation du trafic tels que *SUMO*, *MATSim*, *AimSun*, *VISSIM*, et d'autres ont été développés pour simuler des systèmes de trafic à différentes échelles (Alghamdi et al., 2022). De plus, diverses problématiques, telles que la congestion et la gestion des priorités à une intersection, échappent fréquemment aux résolutions par les méthodes conventionnelles d'analyse en raison de leur complexité. C'est ainsi que les simulateurs de trafic peuvent jouer un rôle fondamental en aidant à prévoir les degrés de congestion, à estimer les effets de nouvelles extensions des réseaux routiers, à évaluer les impacts des changements de circulation, et à planifier les itinéraires pour éviter les embouteillages.

Dans la littérature, il existe plusieurs types de modèles de circulation routière parmi lesquels nous distinguons : les modèles *microscopiques*, *macroscopiques* et *mésoscopiques* (Savrasovs, 2011).

1.4.1 Modèles microscopiques

Le modèle microscopique du trafic routier considère chaque véhicule individuel et sa dynamique en interaction avec les autres véhicules. Cela peut inclure des éléments tels que la vitesse, l'accélération, la distance entre les véhicules et les réactions des conducteurs. Ces modèles sont souvent utilisés pour des études de simulation à petite échelle ou pour des études approfondies de problèmes spécifiques. Le point fort de tels modèles réside dans leur facilité et leur précision du point de vue modélisation. Par conséquent, à travers ce type de modèles, nous pouvons décrire, de façon plus précise, le comportement individuel de n'importe quel véhicule ou conducteur, et ce en fonction

de toute situation comme par exemple les feux de signalisation, le ralentissement ou le changement de la vitesse maximale, les conditions environnementales, les changements climatiques, etc. De nombreux types de modèles, peuvent être considérés tels que :

- *Modèles de poursuite de véhicules* (en anglais *car-followingmodels*) : Ces modèles permettent de modéliser les phénomènes de congestion, d'accélération et de décélération du trafic (Chandler et al., 1958). Ils simulent le comportement d'un véhicule en mouvement en prenant en compte sa vitesse *v(t),* son accélération *a(t)*, sa position *x(t)* et sa distance *d(t)* par rapport au véhicule qui le précède (appelé *lead*) sur la route (Figure 1.4). Par conséquent, la loi de comportement d'un véhicule *i* à l'instant *t* est également définie par une équation différentielle reliant $a_i(t)$ à $v_i(t)$, $v_i(t)$ à $d_i(t)$ comme suit :

$$d_i(t) = x_{i-1}(t) - x_i(t) - l_{i-1}./\ l_{i-1}longueur\ du\ véhicle\ qui\ le\ précède \qquad (1.10)$$

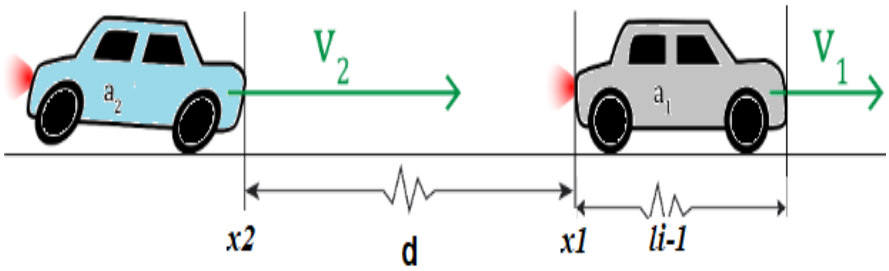


Figure 1.4 Principe du modèle de poursuite de véhicules.

- *Modèles de changement de voie* (en anglais *lane-changingmodels*) : Ces modèles simulent le comportement d'un véhicule qui change de voie sur la route. Ils prennent en compte plusieurs facteurs tels que la vitesse des véhicules sur les autres voies, la densité du trafic et les préférences des conducteurs. Divers modèles de changement de voie ont été proposés dans la littérature à savoir : Gipps (1986), Kita (1999), Hidas (2005), Kesting (2008), etc.
  Les notations de base pour décrire les interactions des véhicules dans un processus de changement de voie sont illustrées par Figure 1.5. Le véhicule S1 est le véhicule sujet, qui a l'intention de rejoindre la voie cible à partir de la voie actuelle. S0 est le véhicule devant S1 dans la voie actuelle. T1, T2 et T3 sont des véhicules dans la voie cible, qui peuvent affecter et être affectés par la tentative de changement de voie de S1. T1 et T2 sont respectivement les véhicules *lead* et *lag*. T3 est le véhicule *lag* suivant T2, et peut devenir le véhicule *lag* si S1 n'est pas capable de fusionner devant T2. Les conditions initiales (gap1 : écart spatial existant entre T1 et T2 ; et H_{lag} : avance spatiale entre T2 et S1), comme illustré à la Figure 1.5, sont utilisées pour décider si un changement de voie compétitif/coopératif est nécessaire ou non.

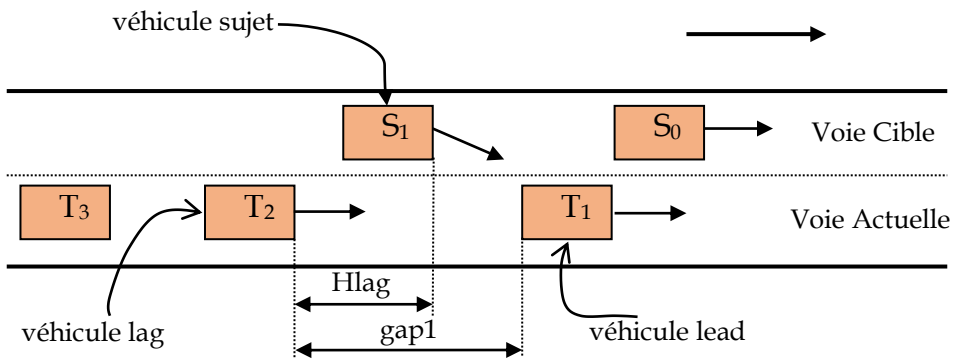


Figure 1.5 Principe du modèle de changement de voie (Sun and Kondyli, 2010).

- *Modèles de comportement des conducteurs* (en anglais *driver behavior models*) : Ces modèles simulent le comportement des conducteurs sur la route, en prenant en compte leur perception de la situation, leur prise de décision et leur action. Ils permettent de modéliser la variabilité du comportement des conducteurs, ce qui peut affecter la sécurité et l'efficacité du trafic. Un état de l'art sur ce type de modèles et ses applications est bien présenté dans le travail d'AbuAli and Abou-zeid (2016).
- *Modèles de mouvement des piétons* (en anglais *pedestrian movement models*) : Ces modèles simulent le mouvement des piétons dans les zones urbaines, en prenant en compte les interactions entre les piétons et les véhicules, ainsi que les obstacles et les contraintes de l'environnement urbain (Teknomo et al., 2016).
- Des *modèles hybrides*, combinant les modèles précédents, peuvent être utilisés afin de créer des modèles de trafic plus complexes et plus réalistes.

Toutefois, ces modèles présentent un inconvénient majeur, qui réside dans leur limite en termes de modélisation du trafic routier à grande échelle. En effet, il est pénible, lors d'une simulation, de tenir compte de toutes les véhicules, toutes les interactions inter-véhicules, toutes les zones du réseau routier, etc., qui génèrent une grande masse de données qui est très difficile à traiter, voire impossible. Ceci rend plus judicieux d'adopter la modélisation macroscopique qui modélise le trafic grâce à entités macroscopiques.

1.4.2 Modèles macroscopiques

Contrairement à la considération individuelle des modèles microscopiques, les modèle macroscopiques examinent le trafic dans son ensemble et utilise des variables agrégées telles que la densité de véhicules, le débit de trafic et la vitesse moyenne, etc. Ils sont considérés parmi les modèles les plus anciens définissant le trafic comme étant un phénomène continu (Whitham, 2011; Faye, 2014; Lighthill and Whitham, 1955). Ils sont souvent provenus de la mécanique des fluides.

Ces modèles sont souvent utilisés pour des prévisions à grande échelle du trafic routier. Ils sont moins couteux en termes de calcul par rapport aux modèles microscopiques, et en effet, ils sont recommandés pour mener des simulations, des estimations, des prévisions et des contrôles du trafic. A travers de tels modèles, nous pouvons être capable de décrire le changement des situations de trafic, telle que la congestion routière, et de modéliser le flux de circulation indépendamment des paramètres et des comportements individuels des véhicules à savoir le changement de voie et le type de véhicule (Ouessai, 2020).

Les premiers modèles macroscopiques du trafic ont été introduit initialement dans les travaux de (Lighthill and Whitham, 1955; Richards, 1956). Ce type de modèle nommé modèle du premier ordre considère que les états du trafic sont des états d'équilibre et que ce modèle évolue d'un état d'équilibre à un autre. Il se base principalement sur la ressemblance avec la mécanique des fluides. Cette dernière permet de synthétiser une première relation de base entre les trois variables à savoir la *densité*, la *capacité* et la *vitesse*. Trois équations importantes sont définies dans ce modèle à savoir :

- L'équation reliant la vitesse, le débit et la concentration :

$$q(x,t) = \rho(x,t)\,.\,v(x,t) \tag{1.11}$$

Tel que :

$q(x,t)$ *est le débit du trafic en un point x et à l'instant t.*
$\rho(x,t)$ *est la densité du trafic en un point x et à l'instant t.*
$v(x,t)$ *est la vitesse moyenne du trafic à la position x et à l'instant t.*

- L'équation de conservation, provenant de la conservation du nombre de véhicules sur une section de longueur infinitésimale et pendant un intervalle de temps. Il s'agit d'une dérivée de la loi de conservation du nombre de véhicules :

$$\frac{\partial \rho(x,t)}{\partial \rho\, t} + \frac{\partial q(x,t)}{\partial \rho\, t} = 0 \tag{1.12}$$

- Le diagramme fondamental du trafic routier, qui est un graphique représentant la relation entre le débit du trafic (en véhicules par heure) et la densité du trafic (en véhicules par unité de distance) pour une route donnée. Cette relation est souvent modélisée par une courbe appelée "*Diagramme Fondamental du Trafic*" ou équation hydrodynamique, par analogie à la mécanique des fluides. la section 1.3.4 offre une présentation plus approfondie de la notion de diagramme fondamental.

Quant aux modèles de second ordre, qui sont des modèles mathématiques utilisés pour décrire le trafic via des équations aux dérivées partielles[2], permettent de prendre en considération les états de non équilibre ainsi que les conditions de convergence vers un état d'équilibre. En effet, l'équation d'équilibre est également dynamique exprimant l'accélération du flux. Ce type de modèles macroscopiques peut être formulé en utilisant des équations telles que l'équation de conservation du trafic, qui décrit la variation de la densité du trafic sur une route en fonction de la vitesse du trafic et des taux d'entrée et de sortie des véhicules. D'autres équations peuvent également être utilisées pour modéliser les interactions entre les véhicules, les conditions de la route et les facteurs environnementaux tels que le temps et l'éclairage.

Il existe plusieurs types de modèles macroscopiques du trafic routier. Voici quelques exemples de ces modèles :

- *Modèles de flux* : Ces modèles décrivent le flux de véhicules sur un réseau routier en utilisant des équations mathématiques qui prennent en compte des paramètres tels que la vitesse et la densité du trafic. Les modèles de flux peuvent être utilisés pour prédire les niveaux de congestion sur un réseau routier donné.
- *Modèles de capacité* : Ces modèles évaluent la capacité d'un réseau routier en termes du nombre maximal de véhicules qui peuvent circuler sur ce réseau en une période de temps donnée. Les modèles de capacité sont utiles pour la planification des infrastructures routières et pour estimer les coûts de construction de nouvelles routes.
- *Modèles de file d'attente* : Ces modèles se réfèrent à des méthodes mathématiques utilisées pour modéliser et analyser les files d'attente de véhicules à divers points du réseau routier. Ils permettent de comprendre les phénomènes liés à la congestion, aux retards et aux temps d'attente des véhicules. Ils sont souvent utilisés pour évaluer l'efficacité des systèmes de signalisation, de gestion du trafic et d'infrastructures routières afin d'optimiser la fluidité et de réduire les temps d'attente. En utilisant des concepts tels que la théorie des files d'attente, ces modèles contribuent à une meilleure planification et gestion du trafic routier.
- *Modèles de diffusion* : Ces modèles décrivent la propagation des congestions et des perturbations du trafic à travers un réseau routier. Les modèles de diffusion sont utiles pour prédire les impacts des accidents, des travaux routiers ou d'autres événements qui pourraient perturber la circulation.
- *Modèles d'équilibre de réseau* : Ces modèles font référence à des modèles mathématiques qui tentent de représenter l'équilibre entre l'offre et la demande sur

[2]Une équation aux dérivées partielles (EDP) est une équation qui lie une fonction de plusieurs variables à ses dérivées partielles.

un réseau routier. Ils décrivent les interactions entre les différents flux de trafic sur un réseau routier et cherchent à établir un équilibre entre ces flux. Les modèles d'équilibre de réseau peuvent aider à identifier les zones de congestion et à proposer des solutions pour les réduire.

1.4.3 Modèles mésoscopiques

Entre les deux types de modèles précédents, des modèles baptisés « *mésoscopiques* » peuvent être utilisés afin de modéliser le trafic routier dans un niveau d'abstraction intermédiaire. Il s'agit d'une hybridation des approches microscopiques et macroscopiques. À travers de tels modèles, les comportements des véhicules ne sont pas modélisées directement de façon individuelle mais sur la base des distributions de probabilités (Treiber and Kesting, 2013).

Les modèles *mésoscopiques* s'appuient généralement sur trois approches décernant ainsi trois types de modèles à savoir : les modèles de cluster, les modèles à distribution du temps inter-véhiculaires et les modèles à gaz cinétique. Dans les modèles de cluster, les véhicules voisins sont regroupés en clusters ou en paquets en fonction de certaines caractéristiques de mobilité telles que la vitesse moyenne, la destination, etc. Ces regroupements sont établis sur la base de caractéristiques similaires, permettant ainsi de représenter des comportements collectifs. Dans les modèles à distribution du temps inter-véhiculaires, la répartition des probabilités du temps séparant deux véhicules consécutifs constitue l'élément de base de la modélisation *mésoscopique*. Ils examinent la séquence de temps entre le passage de chaque paire consécutive de véhicules. Ces distributions temporelles permettent de déduire des informations sur la fluidité du trafic, les périodes de congestion, et les phénomènes d'interactions entre les véhicules. Enfin, les modèles à gaz cinétique sont utilisés en s'inspirant des modèles décrivant le déplacement continu des particules dans un gaz. Ils constituent un outil essentiel pour la modélisation des réseaux routiers en augmentant les détails par rapport aux modèles macroscopiques. Par conséquent, l'utilisation de *modèles à gaz cinétique* permet d'obtenir une compréhension plus approfondie des interactions entre les véhicules et de simuler des scénarios de trafic complexes.

1.4.4 Diagrammes fondamentaux

L'une des alternatives des modèles discutés précédemment, les diagrammes fondamentaux, autrement dits « *diagrammes de la capacité d'une route* », sont utilisés afin de donner des informations sur la dynamique du trafic, et plus particulièrement sur les vitesses de propagation dans les réseaux routiers. Le diagramme fondamental est également un outil utilisé pour décrire la relation entre la *densité* de trafic, la *vitesse* de circulation et le *débit* de véhicules sur une route. Il est basé sur la loi de conservation du

trafic, qui stipule que le nombre de véhicules sur une route à un instant donné est le résultat de l'accumulation ou de la dispersion des véhicules qui y circulent. Cette loi est fondée sur l'hypothèse que le nombre de véhicules sur une route donnée est une fonction multipliant la *densité* de trafic fois la *vitesse* de circulation. En effet, à mesure que la *densité* accroît, la *vitesse* diminue, ce qui peut apparaitre une congestion routière. Cette relation est généralement schématisée par le diagramme fondamental, qui montre la relation entre la *densité*, la *vitesse* et le *débit* du trafic. Il semble très utile pour aussi bien les ingénieurs de la circulation que les planificateurs de transport, notamment lors de la conception de nouvelles infrastructures routières ou intersections, ainsi pour dégager les bonnes décisions à propos de la gestion et le contrôle du trafic et l'optimisation de la congestion routière.

Le diagramme fondamental est sensible à plusieurs facteurs, notamment la géométrie de l'itinéraire, la composition du trafic, la météo et les mesures opérationnelles. A travers ce type de diagrammes, nous pouvons distinguer deux états différents du trafic :

- *Etat fluide*, où les véhicules peuvent se déplacer sans contraintes à une vitesse égale à la vitesse libre. Il correspond également à un régime normal du trafic.
- *Etat congestionné*, où les véhicules se déplacent avec une vitesse inférieure à la vitesse libre. Il correspond ainsi à un régime congestionné du trafic.

La Figure 1.6 présente la forme générale d'un diagramme fondamental. En examinant l'évolution de la courbe, la *densité* dans une zone de trafic fluide se situe quelque part entre *zéro* et la *densité critique*. Le cas où il n'y a absolument pas de véhicules sur la route correspond à la valeur de *densité* de *zéro*. La valeur critique de la *densité* indique qu'il y a des véhicules qui circulent avec une vitesse libre. Au-delà de cette valeur, la vitesse de circulation des véhicules commence à diminuer, ce qui rend ces véhicules incapables de circuler à vitesse libre. D'autre part, la zone de trafic congestionné se situe entre les deux valeurs de la densité : *critique* et *maximale.* La valeur *maximale* de la densité stipule que la route est totalement pleine ainsi qu'aucun véhicule ne peut se déplacer (vitesse est égale à zéro). La densité critique est représentée sur le diagramme par la valeur de débit maximal.

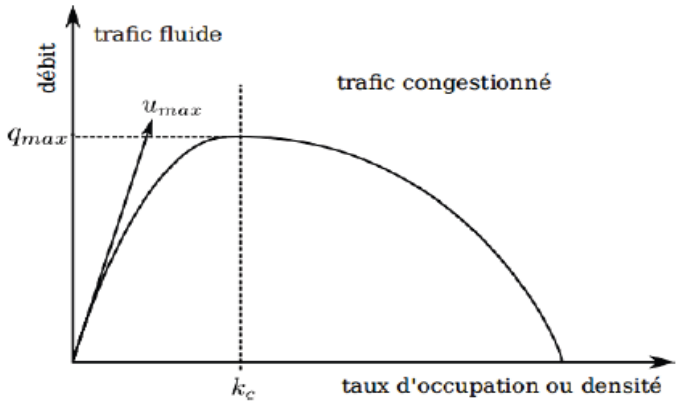


Figure 1.6 Diagramme fondamental.

1.5 Intersections à feux de signalisation

Une intersection est une zone du réseau routier dont deux ou plusieurs voies se croisent. Le trafic dans de telle zone est souvent contrôlé par des feux de signalisation.

Les intersections à feux de signalisation sont souvent utilisées dans les zones urbaines où le trafic est dense ainsi qu'il est important de réguler la circulation pour éviter les conflits d'accès aux différentes voies. Les feux de signalisation devraient être synchronisés pour assurer une circulation fluide des véhicules et des piétons. Ils peuvent être programmés pour fonctionner à des heures spécifiques de la journée en fonction des niveaux du trafic, par exemple, en augmentant le temps alloué pour les voies les plus fréquentées pendant les heures de pointe.

Les feux de signalisation sont souvent installés sur des poteaux ou des supports en hauteur pour que les conducteurs et les piétons puissent facilement les voir. Trois couleurs sont à distinguer dans ces feux : *rouge*, *jaune* et *vert*, chacune ayant une signification spécifique. Le feu *rouge* signifie "*arrêt*" et les conducteurs doivent s'arrêter complètement à la ligne d'arrêt ou avant le passage piéton. Le feu *jaune* signifie "*attention*" et les conducteurs doivent ralentir et être prêts à s'arrêter avant que le feu ne passe au *rouge*. Enfin, le feu *vert* signifie "*traverser*" et les conducteurs peuvent continuer à marcher en toute sécurité.

Par ailleurs, les technologies à base d'*IoT* peuvent également être utilisées dans les réseaux routiers modernes, notamment pour détecter l'état du trafic dans le but d'ajuster, de manière adaptive, les cycles de signalisation en conséquence.

1.5.1 Zones fonctionnelles d'une intersection

Dans une intersection routière, on peut généralement identifier plusieurs zones dont les plus importantes sont (Figure 1.7) :

- La *zone de conflit* : c'est une zone où les véhicules peuvent se croiser ou se rencontrer de manière dangereuse. Elle est baptisée aussi la *zone critique*. Elle est partagée entre tous les véhicules qui traversent l'intersection. Le gestionnaire des feux de signalisation a un rôle important pour contrôler l'accès à cette zone.
- La *zone de stockage* : c'est une zone située avant l'intersection où les véhicules peuvent s'arrêter ou ralentir en attendant que le feu passe au *vert* ou que le trafic se fluidifie. Cette zone est généralement située en amont des feux de signalisation.
- La *zone de sortie* : c'est la partie de l'intersection où les véhicules quittent la voie de circulation et se dirigent vers leur destination. Cette zone est généralement située après la zone de conflit et peut être signalée par des panneaux de signalisation et des marquages au sol pour indiquer la direction à suivre. Les piétons qui traversent l'intersection doivent également faire attention à cette zone pour éviter d'être heurtés par des véhicules qui sortent de l'intersection.

Les véhicules franchissant une intersection traversent ces trois zones dans l'ordre : *zone de stockage*, *zone de conflit* puis *zone de sortie*, ainsi que les véhicules dans chaque voie doivent traverser l'intersection en mode *premier arrivé premier sorti* (*FIFO* : *First in First out*). La direction prise par un flux de véhicules pour traverser l'intersection est baptisée une *trajectoire*. Cette dernière relie la voie d'entrée de l'intersection à sa voie de sortie. D'autre part, les véhicules appartenant à un flux donné peuvent avoir des trajectoires différentes.

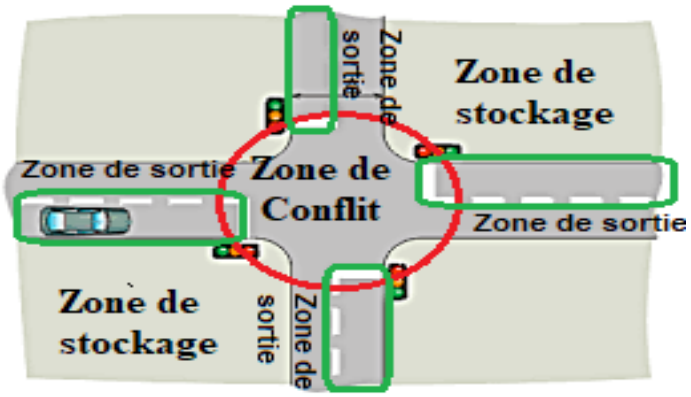


Figure 1.7 Zones fonctionnelles d'une intersection à feux simple de deux routes à sens unique (Wu, 2011).

La gestion des flux dans une intersection s'appuient également sur la détermination des flux *compatibles* et des flux *incompatibles.* Les flux *compatibles* sont les flux dont leurs trajectoires ne se croisent absolument pas, ainsi qu'ils peuvent avoir l'autorisation d'accès à la zone de conflit de l'intersection simultanément. Comme il est illustré sur la Figure 1.8, les flux 1 et 7 sont *compatibles*. Contrairement, les flux 1 et 8 sont deux flux *incompatibles* et l'accès simultané, par ces dernières, à la zone de conflit ne devrait en effet pas être permis. Dans le cas où une intersection dispose de plus deux flux

compatibles, ces derniers peuvent être regroupés pour constituer ce qu'on appelle *Groupe de Flux Compatibles* (*GFC*). Par conséquent, les douze flux décrits sur la Figure 1.8 peuvent être organisés en quatre *GFC* à savoir : *GFC 1= {flux 1 et 7}*, *GFC 2 = {flux 2, 3, 8 et 9}, GFC 3 = {flux 4 et 10} et GFC 4 = {flux 5, 6, 11 et 12}.* En outre, un véhicule ne peut appartenir qu'à un seul *GFC* à un moment donné.

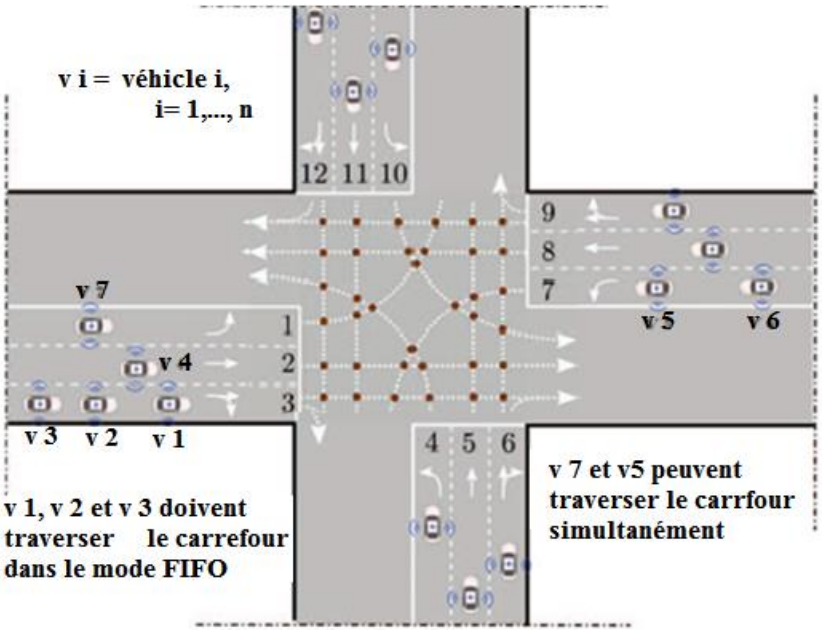


Figure 1.8 Flux compatibles et flux incompatibles (Yan, 2012).

1.5.2 Fonctionnement classique d'une intersection

La Figure 1.9 décrit le principe de fonctionnement du modèle d'intersection traditionnel, qui est le plus souvent référencé dans la littérature du domaine. Un tel modèle est formulé par un ensemble de quatre directions (noté *D*). Chaque direction pourra être une direction d'entrée, direction de sortie ou bien les deux (c'-à-d. Une direction d'entré et de sortie en même temps). La direction d'entrée permet aux différents véhicules en provenance de diverses voies de franchir l'intersection vers la direction de sortie. La direction de sortie établie généralement un lien avec une intersection voisine. Chaque direction d'entrée dispose de deux voies, les véhicules tournant à gauche utilisent la voie la plus à gauche, alors que la voie la plus à droite est généralement utilisée par les véhicules allant tout droit ou tournant à droite.

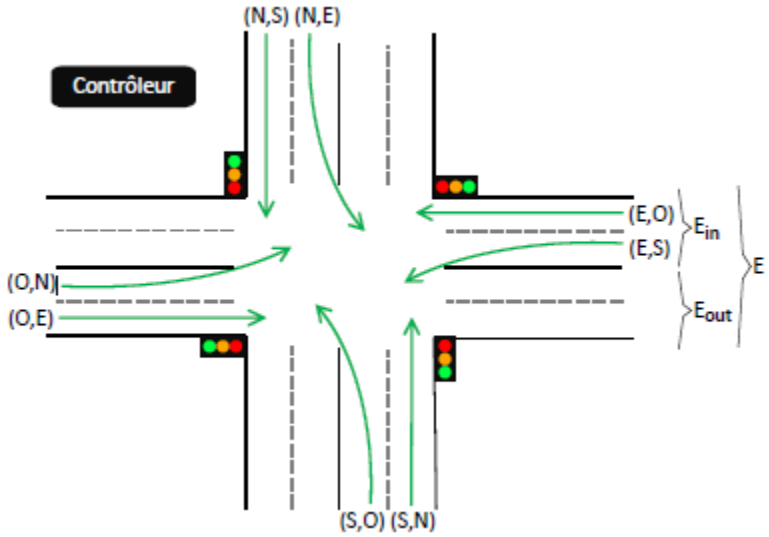


Figure 1.9 Modèle classique d'intersection à feux de signalisation (Faye, 2014).

Généralement, les intersections sont équipées d'un système de contrôle de feux de signalisation (feux tricolores) pour gérer l'accès simultané des véhicules aux zones de conflit. C.à.d. les feux de signalisation servent à attribuer aux usagers de l'intersection (piétons, cyclistes ou véhicules) le droit de s'engager dans la zone de conflit ou sur un passage piéton ou cycliste. Ils fonctionnent selon un système de phases, de cycles et de plans:

- *Phase* : une phase est une période de temps pendant laquelle les feux de signalisation indiquent un comportement spécifique pour les véhicules, les cyclistes et les piétons. Chaque *phase* correspond à un ensemble de mouvements autorisés pour les véhicules et les piétons dans l'intersection. Par exemple, une phase peut autoriser les véhicules venant d'une direction à passer tandis que les feux interdisent le passage aux autres directions. De même, une phase peut autoriser les piétons à traverser l'intersection tandis que les véhicules sont stoppés.
- *Cycle* : un cycle est la durée totale de toutes les phases d'un feu de signalisation. Un cycle typique dure généralement de 60 à 120 secondes, mais cela peut varier en fonction des besoins locaux. Pendant un cycle, toutes les directions de circulation passent par une phase de feu vert.
- *Plan* : un plan de feux de signalisation est une séquence prédéfinie de phases et de durées qui est programmée dans le contrôleur de feux de signalisation. Le plan est conçu pour répondre aux besoins de circulation locaux et peut être ajusté en fonction des fluctuations de la circulation, des événements spéciaux, etc. Le plan peut également inclure des phases spéciales pour les piétons, les cyclistes ou les transports en commun.

La Figure 1.10 illustre une intersection classique avec quatre branches ainsi que tous les mouvements possibles autorisés. Il s'agit d'un cycle composé de quatre phases : *phase 1 (flux 1, 7), phase 2 (flux 2, 8), phase 3 (flux 11, 5), phase 4 (flux 10, 4)*.

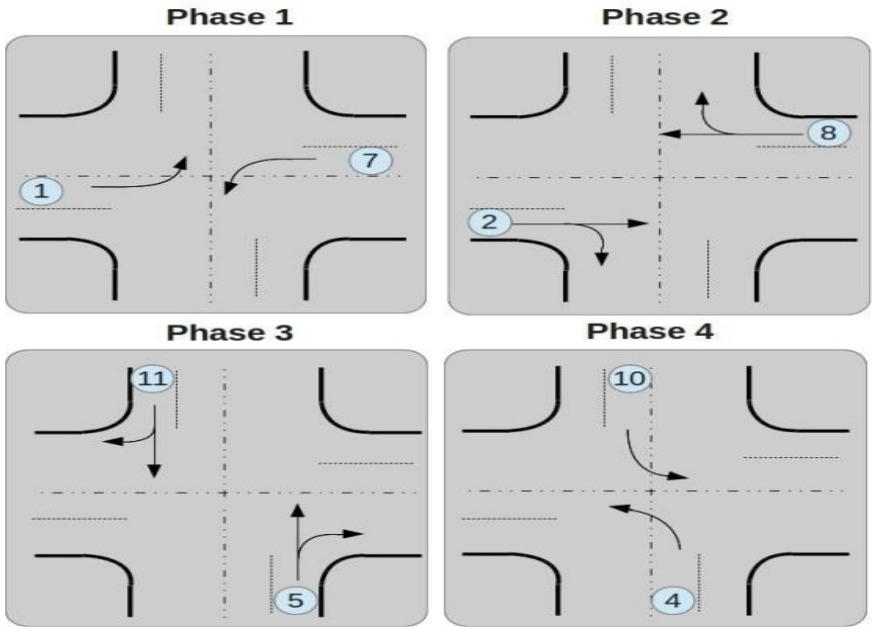


Figure 1.10 Exemple de découpage en phase d'une intersection à quatre directions (Sammoud, 2015).

La Figure 1.11 décrit le découpage temporel des phases de l'exemple de la Figure 1.10. Chaque phase se compose d'un temps *vert effectif* et d'un temps *rouge intégral*. Le temps *vert effectif* est la somme du temps de *vert réel* et du temps de *jaune*. C'est le temps effectivement attribué aux véhicules pour traverser l'intersection. Il existe également un retard du mouvement qui est appelé temps de retard de démarrage situé au début de chaque temps vert effectif. Il utile de noter qu'il est nécessaire d'insérer un temps de *rouge intégral* au niveau de la transition entre deux phases, de manière à garantir un certain niveau de sécurité. Ce temps permet de s'assurer que la zone de conflit est bien vide avant d'autoriser un courant conflictuel. Par conséquent, nous distinguons également les notions de *vert utile* et de *rouge utile*. Le temps de *vert utile* est la différence entre le temps de *vert effectif* et celui du *temps perdu*. Le *rouge utile* est obtenu en retranchant, de la durée du cycle, le temps du *vert utile*.

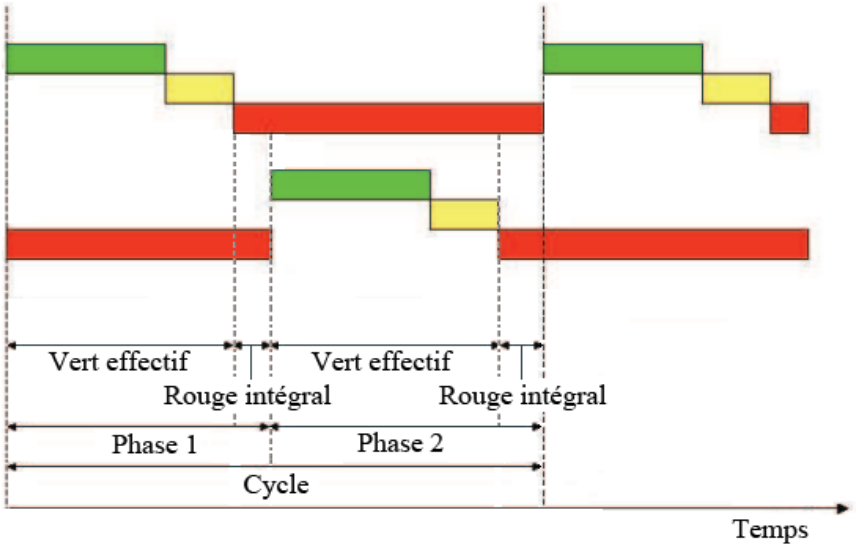


Figure 1.11 Découpage d'un cycle en phases (Perronnet, 2015).

La fluidité du trafic est étroitement liée à la *durée de cycle*, au *nombre de phases* et à la *durée des phases* d'un feu de signalisation.

- *Durée de cycle* : Une durée de cycle plus longue permet de donner plus de temps aux différents flux de circulation pour passer. Cependant, si la durée de cycle est trop longue, les temps d'attente pour les véhicules peuvent être excessifs, ce qui peut causer des congestions. Ainsi, la durée de cycle doit être optimisée en fonction des besoins locaux de circulation pour assurer une circulation fluide et minimiser les temps d'attente.
- *Nombre de phases* : Plus il y a de phases, plus il est possible de donner la priorité à différents flux de circulation et de gérer la circulation dans des situations de trafic complexes. Néanmoins, si le nombre de phases est trop élevé, cela peut augmenter les temps d'attente pour les véhicules et réduire la fluidité du trafic. Encore une fois, le nombre de phases doit être optimisé pour répondre aux besoins locaux de circulation.
- *Durée des phases* : La durée de chaque phase doit être suffisamment longue pour permettre à tous les véhicules de passer, mais pas trop longue pour éviter les temps d'attente excessifs. Une durée de phase trop courte peut également perturber la fluidité du trafic en permettant à trop peu de véhicules de passer pendant chaque phase. Par conséquent, la durée des phases doit également être optimisée pour répondre aux besoins locaux de circulation. En effet, pour équilibrer la charge sur les zones de stockage de l'intersection, cette durée doit être définie de manière à permettre aux véhicules accumulés lors des phases précédentes de quitter la zone de stockage. D'autre part, la mesure de la durée des phases doit tenir compte la taille de la zone de stockage en réduisant le nombre de véhicules en attente à cette dernière. En outre, une saturation peut avoir lieu à l'intersection lorsqu'un véhicule rencontre plus d'un feu rouge dans sa phase individuelle avant d'atteindre la zone de conflit.

De nos jours, notamment dans les villes modernes, les *durées des phases, les durées des cycles* et *le nombre de phases* sont déterminés par des systèmes de contrôle intelligents qui utilisent des capteurs pour détecter la présence de véhicules et de piétons. Le système de contrôle des feux de signalisation peut ajuster ces durées en fonction des conditions de circulation en temps réel.

1.5.3 Régulation des feux de signalisation

Une intersection est un lieu où se croisent deux routes ou plus. Une *route* est caractérisée par sa longueur, son nombre de voies, ainsi que par les sens de circulation. Un *mouvement* est défini par son origine et sa destination. On nomme *courant* l'ensemble des *mouvements* qui appartiennent à la même voie d'origine. Ils sont admis simultanément sans conflit. On appelle *point conflictuel* tout croisement entre deux *mouvements*. En effet, le croisement des véhicules, de plus en plus nombreux et de plus en plus rapides, comporte des risques mettant en péril des vies humaines. Pour des raisons de sécurité, certaines paires de *mouvements* sont interdites. Bien que des règles générales existent concernant la détermination de ces interdictions, leur élaboration est à la discrétion de l'exploitant d'intersection.

1.5.3.1 Modélisation mathématique pour la régulation d'intersection

Un modèle mathématique pour un régulateur d'intersection à feux peut être basé sur des équations d'état et des équations de contrôle. Les équations d'état décrivent l'évolution de l'état du système au fil du temps, tandis que les équations de contrôle décrivent les actions de contrôle à prendre en fonction de l'état actuel du système. Les variables d'état peuvent inclure le nombre de véhicules dans chaque file d'attente, le temps écoulé depuis le dernier changement de feu, le temps restant avant le prochain changement de feu, etc. Les actions de contrôle peuvent inclure le changement de la durée de chaque cycle de feu, la priorisation d'une file d'attente par rapport à une autre, la synchronisation des feux avec d'autres intersections, etc.

Dans cette optique, nous pouvons aborder le problème de régulation de l'intersection comme étant un problème d'ordonnancement des véhicules pour traverser la zone de conflit d'une intersection. Dans le cas des intersections isolées, il est possible de modéliser le problème comme une ressource critique à plusieurs points d'entrées. Une telle ressource autorise un certain nombre de véhicules à traverser la zone de conflit simultanément. Chaque véhicule est considérée comme une tâche qui a une date d'arrivé et qui souhaite d'être traitée. Les véhicules sont organisés en plusieurs *GFCs*. Les véhicules dans un même flux sont traités selon la politique *FIFO*. Les véhicules dans des flux distincts mais appartenant au même *GFC* peuvent traverser la zone de conflit de l'intersection simultanément.

Un régulateur d'intersection à feux peut être modélisé mathématiquement de différentes manières, selon le degré de détail souhaité dans la modélisation et l'objectif de l'étude en question. Toutefois, nous pouvons donner une description générale de la façon dont on peut modéliser un système de contrôle des feux d'intersection comme suit :

Variables d'état : Le régulateur d'intersection à feux peut être considéré comme un système dynamique ayant des variables d'état qui décrivent l'état actuel du système. Par exemple, les variables d'état peuvent inclure :

- Le temps restant pour chaque phase de feu *rouge*, *orange* et *vert* pour chaque direction de circulation.
- Le nombre de véhicules présents dans chaque file d'attente pour chaque direction de circulation.
- Le temps écoulé depuis la dernière fois que chaque feu a changé d'état.

Équations de transition : Les équations de transition décrivent comment les variables d'état évoluent au fil du temps. Par exemple, les équations peuvent inclure :

- Pour chaque phase de feu *vert*, le temps restant diminue à chaque pas de temps jusqu'à ce qu'il atteigne zéro, moment où la phase change à l'état suivant (*orange* ou *rouge*).
- Pour chaque file d'attente, le nombre de véhicules augmente à chaque pas de temps en fonction du taux d'arrivée des véhicules et diminue en fonction du taux de départ.

Fonctions de coût : Les fonctions de coût décrivent les objectifs du système et comment ils sont mesurés. Nous citons à titre indicatif les fonctions de coût suivantes :

- Le temps d'attente moyen des véhicules dans chaque file d'attente.
- Le nombre de véhicules traversant l'intersection par unité de temps.
- Le temps moyen de passage des véhicules à travers l'intersection.

En utilisant ce principe, il est possible de construire un modèle mathématique simulant le comportement du régulateur d'intersection à feux dans le but d'optimiser certaines performances décrites par les fonctions objectifs choisies.

1.5.4 Contrôleurs des feux de signalisation

1.5.4.1 Contrôleurs à temps fixe

Les contrôleurs à temps fixe des feux de signalisation sont des systèmes électroniques programmables qui permettent de contrôler la durée des feux de signalisation dans une intersection de manière prédéfinie. Ces contrôleurs sont souvent être utilisés dans les zones à faible trafic, où le trafic est prévisible et où il est possible de prévoir les moments de congestion.

Ce type de contrôleurs considère les cycles comme fixes pendant une durée donnée et les plans de synchronisation des phases son inchangeables une fois déployés. En effet, les feux de signalisation sont réglés pour changer de couleur à des intervalles de temps prédéterminés, indépendamment de la présence de véhicules ou de piétons. La fixation des temps des différents cycles se base principalement sur l'étude des données de l'historique du trafic. Toutefois, ces temps peuvent être changés d'une période à une autre. C.-à-d. en fonction de la période (*matin*, *midi*, *nuit*, etc.) et parfois du jour, le contrôleur adopte un plan de feux prédéfini. Habituellement, ces contrôleurs prédéfinissent trois principales périodes telles que : les périodes des pointes du matin, de l'après-midi, et le reste (hors pointes) (Faye, 2014).

Le plan de feux le plus simple consiste à répéter indéfiniment la même séquence de phases de durées fixes, toujours agencées dans le même ordre, de manière à constituer un cycle fixe (Gartner et al., 1975; Miller, 1963).

Bien que les contrôleurs à temps fixe soient une solution simple et efficace pour réguler la circulation dans une intersection, ils présentent également certains inconvénients :

- *Manque de flexibilité* : Les contrôleurs à temps fixe sont programmés pour fonctionner selon un plan de signalisation prédéfini, ce qui peut ne pas être adapté aux conditions de circulation en temps réel. En cas de changement de trafic imprévu, les feux de signalisation peuvent ne pas être optimisés pour les besoins actuels, ce qui peut entraîner des embouteillages et des temps d'attente plus longs pour les conducteurs.
- *Inefficacité dans les zones à trafic variable* : Dans les zones à trafic variable, comme les zones commerciales ou les zones industrielles, le trafic peut varier considérablement tout au long de la journée. Les contrôleurs à temps fixe ne peuvent pas s'adapter à ces variations et peuvent ne pas être optimisés pour répondre aux besoins du trafic à différents moments de la journée.

Des solutions plus avancées, telles que les systèmes de contrôle semi-adaptatifs ou adaptatifs en temps réel, peuvent offrir une solution plus efficace pour répondre aux besoins actuels du trafic.

1.5.4.2 Contrôleurs semi-adaptatifs

Les contrôleurs semi-adaptatifs sont une amélioration des contrôleurs à temps fixe. Ils se basent sur les données de trafic collectées en temps réel pour ajuster les plans de signalisation. En fonction des conditions du trafic actuelles, ces contrôleurs utilisent également des modèles de trafic pour prévoir les conditions de circulation futures. Ces

modèles prennent en compte les tendances de circulation passées et les données de trafic en temps réel pour prédire les périodes de pointe.

Les contrôleurs semi-adaptatifs sont souvent utilisés dans les zones à trafic mixte où le trafic est moins prévisible qu'à d'autres endroits. Ces systèmes peuvent détecter la présence de piétons et de cyclistes, ainsi que les tendances de trafic aux heures de pointe et en tenir compte dans les plans de signalisation.

Les avantages des contrôleurs semi-adaptatifs des feux de signalisation incluent :

- *Amélioration du trafic* : Les contrôleurs semi-adaptatifs peuvent aider à réduire les temps d'attente pour les véhicules en ajustant les plans de signalisation en fonction de l'état actuel du trafic. Cela peut également aider à réduire les congestions et les temps de trajet.
- *Réduction de la pollution atmosphérique et sonore* : Ils peuvent aider à réduire la pollution atmosphérique et sonore en minimisant le temps d'attente des véhicules aux intersections.
- *Flexibilité accrue* : Ils peuvent s'adapter à des situations de circulation variables, ce qui les rend plus flexibles que les contrôleurs à temps fixe.

Néanmoins, les contrôleurs semi-adaptatifs présentent également certains inconvénients, particulièrement :

- *Coûts plus élevés* : Les contrôleurs semi-adaptatifs peuvent être plus coûteux que les contrôleurs à temps fixe, en raison des coûts de la mise en place de capteurs et de l'installation de logiciels supplémentaires.
- *Complexité accrue* : Ils sont plus complexes que les contrôleurs à temps fixe, ce qui peut rendre la maintenance et la réparation des défaillances plus difficiles.

Bien que ces contrôleurs utilisent des observations en temps réel des flux de trafic, ils restent encore incapables d'optimiser des objectifs qui tiennent en compte plusieurs paramètres comme par exemple le temps total de déplacement ou la somme de toutes les files d'attente. À la limite, lorsque le trafic est important sur toutes les directions de l'intersection, les contrôleurs semi-adaptatifs fonctionnent comme des contrôleurs à temps fixe.

D'autre part, les contrôleurs semi-adaptatifs se basent sur des schémas de programmation anticipés pour ajuster les durées des feux en fonction des heures de pointe espérées, des périodes de faible trafic et d'autres conditions routières typiques. Toutefois, ces schémas de programmation ne prennent pas en compte les changements en temps réel du trafic, ce qui peut conduire à des retards graves et à des congestions.

C'est pourquoi les contrôleurs adaptatifs ont été proposés pour surmonter cette limite. Ils adoptent des algorithmes pour ajuster automatiquement les durées des feux en fonction de l'état du trafic en temps réel.

1.5.4.3 Contrôleurs adaptifs

Linguistiquement, le terme « *adaptatif* » est un adjectif qui dérive du verbe « *adapter* ». « *Adapter* » signifie ajuster quelque chose pour qu'il convienne à un certain contexte ou à une certaine situation. L'adjectif « *adaptatif* » décrit quelque chose qui est capable de s'adapter à des changements ou des situations variables, souvent de manière automatique ou programmée. En effet, les contrôleurs adaptatifs sont caractérisés par leur capacité à ajuster constamment les plans de synchronisation des phases pour qu'elles conviennent mieux aux objectifs visés et aux différentes variations de l'état du trafic. Ces contrôleurs s'appuient sur les données collectées de l'état actuel de la circulation pour améliorer la fluidité du trafic en ajustant les temps des phases en temps réel.

Dans la littérature, plusieurs travaux de recherche se sont intéressés à l'étude du problème de contrôle adaptatif des feux de signalisation donnant en effet naissance de plusieurs méthodes que nous pouvons classer quatre classes :

- *Méthodes de contrôle adaptatif basé sur des capteurs* (*méthodes réactives*) : cette classe de méthodes utilise des capteurs installés sur la chaussée pour détecter le nombre de véhicules et la vitesse de circulation et ajuster les temps de phase en conséquence.
- *Méthodes de contrôle adaptatif basé sur les modèles de trafic* : ces méthodes utilisent des modèles mathématiques pour prédire les conditions de circulation futures et ajuster les temps de phase en conséquence. Nous citons, à titre d'exemple, les méthodes : *TRANSYT* (Robertson, 1968) , *TRANSYT-7F* (Wallace et al., 1988) et *SIGOP III* (Lieberman et al., 1983), *SCOOT* (Robertson and Bretherton, 1991).
- *Méthodes de contrôle adaptatif issues de l'intelligence artificielle* : ces méthodes utilisent des algorithmes d'intelligence artificielle comme les algorithmes qui se basent sur la logique floue (Hawi et al., 2017; Hartanti et al., 2019), Les algorithmes génétiques (Xiao-Feng Chen and Zhong-Ke Shi, 2002), les algorithmes d'apprentissage automatique (Bingham, 2001), etc.
- *Méthodes de contrôle adaptatif basé sur les communications véhicules-infrastructure* : ce type de méthodes utilise des technologies de communication sans fil pour permettre aux systèmes de contrôle de feux de signalisation de communiquer avec les véhicules et adapter les temps de phase en fonction de la présence de véhicules à proximité. De nombreuses recherches ont été menées pour étudier la problématique des communications véhicule-véhicule ou véhicule-infrastructure parmi lesquelles

nous citons par exemple : (Yan, 2012; Kato et al., 2002), (Chisalita and Shahmehri, 2002) et (Gradinescu et al., 2007).

1.5.4.4 Contrôleurs adaptatifs avec coordination de signaux

Classiquement, la coordination des différents signaux se concrétise dans la majorité des cas en fixant un même temps de cycle pour chaque intersection, voire en fixant une même durée de phase pour chaque intersection, mais avec décalage. Toutefois, ce type de coordination reste encore insuffisant du fait qu'il ne tienne pas en compte les variations permanentes de la circulation. C'est ainsi que des études récentes, comme celles présentées dans cette thèse, suggèrent que les intersections pourraient fonctionner de manière collaborative et adaptative et qu'une synchronisation des feux pourrait être mise en place non seulement sur les axes principaux, mais également sur une large zone urbaine. Parmi les techniques utilisées pour synchroniser les feux de signalisation, nous citons la technique "*Vague verte*" (en anglais : *Greenwave*) (Lu et al., 2023). Cette dernière est souvent utilisée pour décrire un système de gestion du trafic qui vise à synchroniser les feux de circulation pour créer une "*vague*" de feux *verts* pour les véhicules qui se déplacent dans une direction particulière. Cela peut contribuer à réduire la congestion et à améliorer le flux de circulation. Bien que cette *Greenwave*, présente des avantages pour la gestion du trafic, elle présente également certaines limites à savoir : la coordination des feux pour une "*vague verte*" sur une route principale peut avoir des conséquences sur les intersections transversales. Les véhicules qui traversent la "*vague verte*" peuvent être confrontés à des feux rouges aux intersections perpendiculaires, ce qui peut entraîner des retards. De plus, les contrôleurs à base de *Greenwave* traditionnels peuvent manquer d'adaptabilité en temps réel. Ils peuvent ne pas être suffisamment réactifs aux changements rapides dans les conditions de circulation, tels que des accidents soudains, des conditions météorologiques extrêmes ou des événements spéciaux. D'autre part, *Greenwave* ne permet l'optimisation du flux que dans les routes unidirectionnelles. Pour surmonter ces limitations, d'autres techniques de coordination peuvent être adoptées, telles que : les contrôleurs adaptatifs avec coordination de signaux (en anglais "*adaptive controllers with signal coordination*"), le contrôle adaptatif de gestion du trafic centralisés, etc.

Les contrôleurs adaptatifs avec coordination de signaux permettent de contrôler plusieurs intersections souvent adjacentes. Les contrôleurs des différentes intersections adaptent dynamiquement leurs durées des cycles des feux de manière coordonnée en visant en effet la réduction de la congestion, l'amélioration de la sécurité et la minimisation de l'émission de gaz de carbone. En effet, chaque contrôleur ajuste la durée du cycle de ses feux en fonction aussi bien de l'état du trafic dans son intersection que des états des autres intersections. Plusieurs méthodes ont été proposées pour

résoudre ce problème, notamment les approches basées sur les systèmes multi-agents, l'apprentissage automatique, la théorie des jeux, la planification automatique, la programmation linéaire, etc.

L'état actuel de la recherche sur les contrôleurs adaptatifs intelligents avec coordination des feux de signalisation témoigne d'une diversité d'approches novatrices visant à optimiser la gestion du trafic routier. Wu et al. (2022) a adopté une approche décentralisée utilisant les *SMA* avec la théorie des jeux évolutionniste. Dans de telle approche chaque intersection est modélisée comme un agent qui vise non seulement de rendre le flux de véhicules plus fluide à l'intersection, mais aussi la réduction des temps de déplacement pour tous les véhicules.

À la lumière des performances remarquables d'apprentissage automatique, de nombreuses études ont proposées pour le contrôle coopératif des feux de signalisation à savoir : (Yan and Shang, 2022; Kolat et al., 2023). D'autre part, dans l'étude (Chu et al., 2020a), les auteurs ont proposé une méthode basée sur l'acteur critique pour coordonner les intersections dans un environnement partiellement observable. Zhou et al. (2021) ont proposé une stratégie coopérative consciente de la région basée sur le réseau d'attention graphique (en anglais *GAT : Graph Attention Network*), incorporant ainsi les informations spatiales des intersections voisines.

1.6 Conclusion

En conclusion de ce chapitre sur les généralités de la théorie du trafic routier, nous avons pu aborder les principaux concepts et modèles qui sous-tendent l'étude de la circulation routière. Nous avons vu comment la *densité*, la *vitesse* et le *débit* de trafic peuvent être mesurés pour analyser les performances des réseaux routiers, ainsi que les conséquences de la congestion routière. Nous avons également exploré les différentes méthodes de mesure et de modélisation du trafic, notamment les modèles *macroscopiques*, *mésoscopiques* et *microscopiques*. Enfin, nous avons présenté le principe de la gestion des intersections à feux de signalisation ainsi que les différentes approches proposées pour le contrôle des feux de signalisation afin d'optimiser la fluidité et la sécurité de la circulation. Nous avons également présenté le principe général de ces approches qui sont souvent groupées en trois grandes classes à savoir : les approches à temps fixe, les approches semi-adaptatives et les approches adaptatives.

Revenant sur le point central de cette thèse, il est à noter que nos travaux s'inscrivent dans le domaine des approches adaptatives, avec une orientation marquée vers des solutions fondées sur l'apprentissage par renforcement (en anglais *RL : Reinforcement Learning*). *RL* est l'une des techniques d'apprentissage automatique qui permet à un

agent de prendre des décisions en interagissant avec un environnement pour maximiser une récompense. Dans le contexte de la gestion du trafic, cette technique peut être utilisée pour le contrôle des feux de signalisation afin d'optimiser la fluidité du trafic et de réduire les temps d'attente pour les usagers de la route. Dans le chapitre suivant, nous explorerons les principes fondamentaux de l'apprentissage par renforcement et sa mise en œuvre pour le contrôle des feux de signalisation. Nous discuterons également des avantages de cette approche par rapport aux méthodes traditionnelles et des défis à relever pour son déploiement dans le monde réel.

Chapitre 2 : Apprentissage par renforcement pour le contrôle du trafic routier

2.1 Introduction

Les dernières améliorations technologiques ont augmenté la qualité du transport. De nouvelles approches basées sur les données font apparaître une nouvelle direction de recherche pour tous les systèmes basés sur le contrôle, par exemple dans les transports, la robotique, l'*IoT* et les systèmes électriques. La combinaison des applications basées sur les données avec des systèmes de transport joue un rôle clé dans les systèmes de transport récents. De nombreuses études ont démontré que l'optimisation de la gestion des feux de signalisation permet d'améliorer l'efficacité du transport urbain et, par conséquent, la qualité de vie des habitants de la ville.

D'autre part, *RL* est une branche de l'*IA* qui permet à un agent d'apprendre à prendre des décisions en interagissant avec son environnement. En effet, à l'inverse à l'apprentissage supervisé, en *RL* on ne dispose pas de données préalables d'entrainement, mais le modèle acquiert ces données directement en interagissant avec l'environnement en les mémorisant dans une mémoire. Cette branche est de plus en plus utilisée dans de nombreux domaines tels que la robotique, le contrôle de processus, les jeux, la publicité en ligne, contrôle du trafic, etc. Ceci, a lui, attiré l'attention d'une

myriade de chercheurs de diverses disciplines, notamment ceux travaillant sur les problématiques liées au domaine de la gestion et contrôle du trafic routier.

Dans ce chapitre, nous allons présenter les fondements du *RL* et les différentes approches algorithmiques permettant d'optimiser les décisions du système intelligent. Nous verrons également comment *RL* peut être appliqué à des problèmes complexes de contrôle et de prise de décision. Nous commencerons par présenter les concepts clés du *RL*, notamment le modèle de l'agent environnement, la notion de récompense, la politique et la fonction valeur. Nous aborderons ensuite les différents algorithmes d'apprentissage par renforcement tels que *Q-Learning*, *SARSA*, *Deep Q-networks*, etc. Nous nous concentrerons ensuite sur l'application de *RL* pour optimiser la fluidité du trafic routier en agissant sur les feux de signalisation. La gestion et contrôle intelligent des feux de signalisation représente, sans aucun doute, l'un des problèmes complexes de prise de décision, en temps réel, qui a attiré l'attention d'une multitude de chercheurs. Nous verrons comment les approches à base de *RL* peuvent aider à optimiser la fluidité du trafic en minimisant entre autres le temps d'attente des véhicules, les émissions de CO_2, etc.

2.2 Aperçu sur l'apprentissage automatique

L'apprentissage automatique (en anglais *ML : Machine Learning*) est un axe de recherche de l'*IA* qui permet aux machines d'apprendre et de s'améliorer à partir de données. L'objectif de *ML* est de développer des algorithmes qui peuvent apprendre à résoudre des problèmes sans avoir besoin d'être explicitement programmés pour chaque tâche. Au lieu de cela, les algorithmes sont conçus pour apprendre à partir de données, à partir de la reconnaissance de motifs et de la généralisation.

ML peut être divisé en plusieurs catégories, notamment l'apprentissage *supervisé*, l'apprentissage *non supervisé* et l'apprentissage *par renforcement* (Figure 2.1). Dans l'apprentissage *supervisé*, le modèle est entraîné à partir de données étiquetées, c'est-à-dire des données qui ont déjà été classées ou catégorisées. Le modèle apprend alors à prédire les étiquettes pour de nouvelles données. Dans l'apprentissage *non supervisé*, le modèle est entraîné à partir de données non étiquetées, afin de découvrir des structures et des motifs dans les données. Le *RL*, quant à lui, est une technique d'apprentissage dans laquelle un agent apprend à prendre des décisions en interagissant avec un environnement.

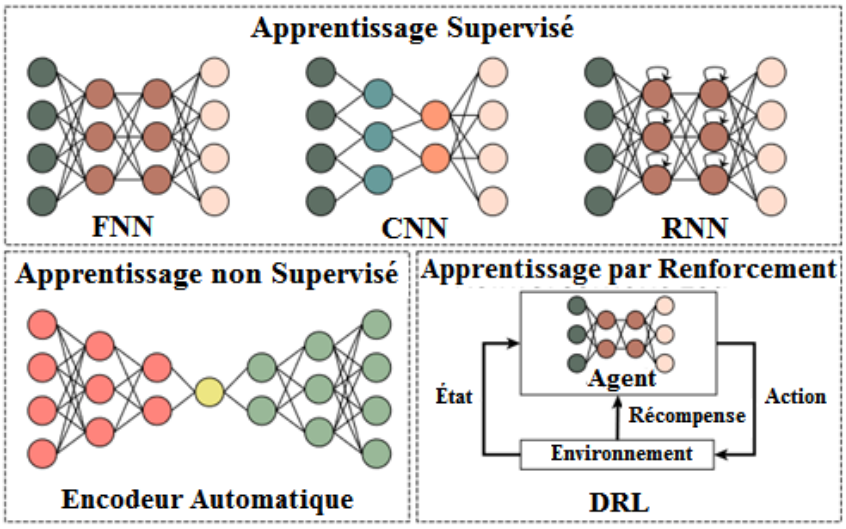


Figure 2.1 Catégories d'apprentissage automatique.

ML est de plus en plus utilisé dans de nombreux domaines, tels que la reconnaissance d'image, la reconnaissance de la parole, la prédiction de tendances, la détection de fraude, la recommandation de produits et la gestion de la relation client. Les algorithmes de *ML* sont également largement utilisés dans le domaine de la gestion du trafic routier, notamment pour la gestion et contrôle des feux de signalisation. Dans les sections suivantes, nous allons nous concentrer sur les fondements théoriques de *RL* ainsi que sur leur utilisation dans notre contexte du travail.

2.3 Fondements théoriques de l'apprentissage par renforcement

RL représente l'une des trois catégories d'apprentissage automatique (*les deux autres sont l'apprentissage supervisé et l'apprentissage non supervisé*) qui utilise la rétroaction bruyante et l'expérience passée pour résoudre des problèmes (Sutton and Barto, 2018). Il est également basé sur la théorie de l'apprentissage par essai-erreur, qui suppose que les agents apprennent, à optimiser leur comportement, à partir de leurs interactions avec leur environnement en essayant différentes actions et en observant les résultats. Les agents apprennent à maximiser une récompense, c'est-à-dire une mesure de la performance de l'agent dans l'environnement (Sutton and Barto, 2018). Plus précisément, *RL* utilise des algorithmes qui permettent à un agent d'apprendre à prendre des décisions optimales dans un environnement en interagissant avec celui-ci (Figure 2.2). À chaque étape de l'interaction, l'agent observe l'état de l'environnement, prend une action et reçoit une récompense. L'agent utilise alors cette information pour mettre à jour son modèle de l'environnement et améliorer sa politique de décision.

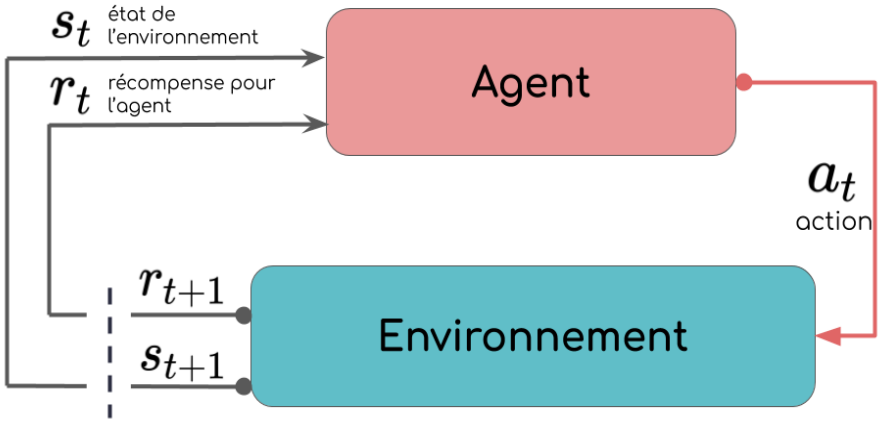


Figure 2.2 Illustration de l'interaction agent-environnement dans l'apprentissage par renforcement (Sutton and Barto, 2018).

En d'autres termes, les fondements théoriques de *RL* sont basés sur les éléments suivants :

- *L'agent ou l'apprenant ;*
- *L'environnement avec lequel l'agent interagit ;*
- *La politique que l'agent suit pour prendre des décisions ;*
- *La récompense que l'agent calcul lorsqu'il entreprend des actions ;*

L'agent est une entité autonome qui interagit avec son environnement pour atteindre des objectifs spécifiques. Il est généralement équipé d'un système de récompenses qui lui permet d'estimer la qualité de ses actions et de modifier son comportement en conséquence. L'agent représente l'élément qui prend des décisions, qui agit. Il vise à apprendre à prendre les bonnes décisions en maximisant une récompense globale au fil du temps. Les décisions sont appelées actions, elles sont définies antérieurement. L'agent évolue, dans le temps, dans un contexte défini par un ensemble d'états (noté S= {s_1, s_2, …}). A chaque action effectuée, l'agent influence le prochain état de l'environnement.

L'environnement est le monde ou le système dans lequel l'agent interagit pour apprendre à prendre les meilleures décisions. Cet environnement peut être simulé ou réel, et il peut être très simple ou très complexe. Dans un contexte d'apprentissage par renforcement, l'environnement est caractérisé par un ensemble d'états possibles (noté S= {s_1, s_2, …}) que l'agent peut observer à tout moment afin de sélectionner une action parmi plusieurs actions possibles (notée A= {a_1, a_2, …}) et d'évaluer une récompense notée r_i. Par conséquent, L'agent effectue des actions en fonction de son état actuel s_i, et l'environnement répond avec une nouvelle observation s_{i+1} et une récompense r_{i+1}. Dans de nombreux cas, l'environnement est modélisé sous forme de processus de décision

markovien (en anglais *MDP : Markovian Decision Process*) (Shani et al., 2005), où les *états*, les *actions* et les *récompenses* sont définis en fonction de la probabilité de transition entre les *états*. Cependant, il existe également d'autres formalismes pour représenter l'environnement, comme les arbres de décision et les processus de décision hiérarchiques (en anglais *HDP : Hierarchical Decision Process*) (Chancey, 1991).

La politique représente la stratégie que l'agent adopte pour choisir des actions dans son environnement. À partir de ses interactions avec son environnement, un algorithme d'apprentissage par renforcement calcule une politique (notée $\pi : S \rightarrow A$), c'est-à-dire une fonction qui à chaque *état* préconise une *action* à exécuter, dont on espère qu'elle maximise les *récompenses*. La politique peut être déterministe ou stochastique, selon que l'agent choisit toujours la même action pour un *état* donné ou s'il choisit une *action* au hasard en fonction de la probabilité définie par la politique.

Il est utile de préciser que *RL* s'appuie aussi sur deux concepts clés : *exploration* et *exploitation*. L'exploration consiste à prendre des actions aléatoires ou inconnues pour découvrir de nouvelles informations sur l'environnement ou la tâche à accomplir. L'exploitation, quant à elle, consiste à prendre les actions qui semblent les plus prometteuses en fonction des connaissances acquises jusqu'à présent. En effet, un compromis entre l'exploration et l'exploitation doit être trouvé afin d'établir un équilibre entre les nouvelles actions et les actions apprises. Lorsque l'agent se concentre trop sur l'exploration, il peut passer trop de temps à prendre des actions aléatoires et ne jamais atteindre les récompenses optimales. D'un autre côté, s'il se concentre trop sur l'exploitation, il peut manquer des opportunités d'explorer de nouvelles actions qui pourraient s'avérer plus efficaces.

Plusieurs approches proposées dans littérature pour équilibrer l'exploration et l'exploitation dans *RL*. L'une des plus courantes est l'approche *ε-greedy*, qui consiste à prendre une action aléatoire avec une probabilité ε et la meilleure action connue avec une probabilité $1\text{-}\varepsilon$. De cette façon, l'agent peut explorer de nouvelles actions avec une certaine probabilité tout en exploitant les actions les plus prometteuses la plupart du temps. D'autres approches incluent *l'optimisation bayésienne*, qui utilise des modèles probabilistes pour trouver les actions les plus prometteuses, et l'exploration *guidée par la curiosité*, qui encourage l'agent à explorer de nouvelles actions en fonction de ce qu'il ne sait pas encore.

2.4 Algorithmes d'apprentissage par renforcement

Les algorithmes d'apprentissage par renforcement peuvent être classés en deux grandes catégories : les algorithmes à base de modèle et les algorithmes sans modèle.

Néanmoins, certains algorithmes d'apprentissage par renforcement peuvent utiliser une combinaison de ces deux approches. Il s'agit de ceux que nous appelons les algorithmes d'apprentissage par renforcement hybrides qui peuvent utiliser des modèles pour certains *états* ou *actions*, et des approches sans modèle pour d'autres.

Avant de présenter les détails des différents algorithmes *RL* proposés dans la littérature, il est important de noter que le processus de renforcement de base est modélisé comme un processus décisionnel markovien (en anglais *MDP : Markov Decitional Process*), que nous détaillons dans la section suivante.

2.4.1 Processus Décisionnel de Markov

En théorie de la décision et de la théorie des probabilités, un *MDP* est un modèle stochastique où un agent prend des décisions et où les conséquences de ses actions sont aléatoires («Processus de décision markovien», 2023). Un *MDP* est un cadre mathématique bien adapté pour optimiser les processus de prise de décision dans des environnements incertains et dynamiques (Bellman, 1957). Il est utilisé dans *RL* pour déterminer la meilleure politique à suivre pour maximiser une récompense donnée.

Le *MDP* modélise l'environnement comme une série d'états possibles, dans lesquels l'agent peut se trouver, et les actions qu'il peut prendre à partir de chaque état. Lorsque l'agent exécute une action, l'environnement passe à un nouvel état, et une récompense est retournée en fonction de l'état actuel et de l'action décidée. L'objectif de *RL* est d'adopter la meilleure politique, c'est-à-dire la séquence d'actions qui maximise la récompense à long terme.

Le problème *RL* peut être formalisé mathématiquement par *MDP* comme étant un quintuple (S, A, P, R,γ) comprenant, respectivement, l'espace d'états S, l'espace d'action A, la probabilité de transition P, la fonction de récompense R et le facteur de d'évaluation γ, où leurs définitions sont données comme suit :

- Ensemble d'état S : Au pas de temps t, l'agent observe l'état $s_t \in S$.
- Espace des actions possibles A : Au pas de temps t, l'agent choisi une action $a_t \in A$.
- Fonction de transition P : Elle définit la probabilité de transition. Comment les états changent entre des périodes de temps successives t et $t+1$ en fonction des actions possibles par l'agent cette fonction est notée $P(s_{t+1} \mid s_t, a_t)$.
- Fonction de récompense notée $R(s_t, a_t, s_{t+1})$: Au pas de temps t, l'agent obtient une récompense r_t pour une action a_t sur l'état s_t lors de la transition vers l'état suivant s_{t+1}.
- Facteur de dévaluation$\gamma \in [0,1]$: qui représente la différence d'importance entre les récompenses à plus ou moins long terme. L'objectif d'un agent est de trouver une

politique qui maximise la somme des récompenses cumulées, où le facteur γ contrôle l'importance des récompenses immédiates par rapport aux récompenses futures. Si l'on considère que la récompense est reçue sur une longue période, alors un facteur de dévaluation γ peut être incorporé pour refléter l'actualisation. La récompense cumulée attendue au pas de temps *t* est défini comme suit :

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k} \tag{2.1}$$

Choisir une valeur de γ plus proche de *1* signifie que les actions de l'agent dépendent davantage de la récompense future. Alors qu'une valeur plus proche de *0* entraîne des actions qui se préoccupent principalement de la récompense instantanée r_t.

2.4.2 Algorithmes à base de modèle

Les algorithmes de *RL* base de modèle utilisent une représentation explicite du modèle de l'environnement. Cela signifie qu'ils connaissent la dynamique exacte de l'environnement, c'est-à-dire les règles de transition d'état et les récompenses associées à chaque état et à chaque action. Ces algorithmes peuvent utiliser cette information pour simuler des états futurs et choisir la meilleure action en fonction de cette simulation. Les algorithmes à base de modèle sont souvent utilisés lorsque la dynamique de l'environnement est connue et que la simulation de l'environnement est peu coûteuse.
Plusieurs algorithmes à base de modèle ont été proposés dans la littérature à savoir :

- *L'algorithme de la programmation dynamique ;*
- *L'algorithme de Monte-Carlo ;*

2.4.2.1 Algorithme de la programmation dynamique

L'algorithme de la programmation dynamique est une technique d'optimisation qui peut être appliquée à *RL* pour trouver la meilleure politique à suivre dans un environnement donné. Il peut être utilisé pour résoudre des problèmes de renforcement à petite échelle en stockant la valeur de la récompense attendue pour chaque état et chaque action possible dans une table. Ensuite, l'algorithme utilise cette table pour calculer la valeur de la récompense attendue pour chaque état en utilisant la formule de *Bellman*. En effet, en utilisant la formule de *Bellman*, qui est une équation récursive, on peut calculer la valeur de l'état actuel en utilisant les valeurs de la récompense attendue pour les états suivants et en prenant la meilleure action possible permettant de maximiser la somme des récompenses attendues futures.

La programmation dynamique est favorisée notamment pour résoudre des problèmes d'apprentissage avec un petit nombre d'états et d'actions. Néanmoins, pour des problèmes plus complexes avec un grand nombre d'états et d'actions possibles,

d'autres algorithmes, comme par exemple les algorithmes de *Monte Carlo* sont plus favorisés.

2.4.2.2 Algorithme de Monte-Carlo

L'algorithme de *Monte-Carlo* utilise des simulations pour calculer les valeurs d'état-action. Il est souvent appliqué pour résoudre des problèmes à grande échelle où la dynamique est mal connue. Le déroulement de l'algorithme de *Monte-Carlo* appliqué dans le *RL* pourra être résumé en ce qui suit :

1. *Formulation du problème* : Le problème est formulé comme un *MDP* qui décrit la dynamique de l'environnement, les actions possibles de l'agent, les récompenses obtenues et les états possibles dans lesquels l'agent peut se trouver.
2. *Génération d'épisodes* : L'algorithme de *Monte-Carlo* génère des épisodes en faisant interagir l'agent avec l'environnement. Un épisode est une séquence d'états, d'actions et de récompenses qui commence à l'état initial et se termine lorsque l'agent atteint un état final.
3. *Estimation de la valeur de l'état* : *Monte-Carlo* utilise les épisodes générés pour estimer la valeur de l'état. La valeur de l'état est la récompense attendue que l'agent puisse obtenir à partir de cet état en suivant une politique donnée.
4. *Mise à jour de la politique* : *Monte-Carlo* utilise les estimations de valeur pour mettre à jour la politique de l'agent. La politique de l'agent est une fonction qui mappe les états possibles aux actions possibles. La politique est mise à jour pour maximiser la valeur de la récompense cumulative attendue.
5. *Itérations* : Les étapes 2 à 4 sont répétées jusqu'à ce que la politique converge vers une politique optimale.

Dans le contexte de cette thèse, qui est caractérisé par un environnement dynamique et à grande échelle, la plupart des solutions proposées, notamment celles qui se basent sur *RL* reposent sur les algorithmes sans modèle. Cela est justifié par la grande difficulté que les chercheurs trouvent pour proposer des modèles prédéterminant tous les états et actions possibles du trafic. La classe des algorithmes sans modèle est bien présentée dans la section suivante.

2.4.3 Algorithmes sans modèle

Bien que les algorithmes à base de modèle démontrent leur performance pour résoudre les problèmes de petites tailles, ils restent encore limités pour les raisons suivantes :

- *Ils peuvent être plus sensibles aux erreurs de modélisation ;*
- *Ils peuvent être plus coûteux en temps de calcul ;*
- *Il peut y avoir des difficultés pour construire un modèle précis de l'environnement.*

En effet, les algorithmes d'apprentissage sans modèles sont plus couramment utilisés que les algorithmes à base de modèle, notamment pour résoudre les problèmes

complexes dont il est difficile de décrire un modèle précis de l'environnement. En outre, ils sont souvent plus flexibles et peuvent apprendre directement à partir de l'interaction avec l'environnement.

Les algorithmes *RL* sans modèle n'utilisent pas de représentation explicite du modèle de l'environnement. Ils ne connaissent pas la dynamique exacte de l'environnement et doivent apprendre à partir des interactions avec celui-ci. Ces algorithmes sont souvent utilisés lorsque la dynamique de l'environnement est complexe et difficile à modéliser, ou lorsque la simulation de l'environnement est coûteuse.

De nombreux algorithmes sans modèle ont été proposés dans la littérature que nous pouvons classifier en trois grandes classes (Sutton and Barto, 2018; Arulkumaran et al., 2017) à savoir : (i) *Les algorithmes basés sur les valeurs* qui utilisent une fonction de valeur pour évaluer les différentes actions dans un environnement donné. (ii) *Les algorithmes basés sur les politiques* qui cherchent à optimiser directement une politique qui décide quelle action prendre dans un état donné. Ils utilisent une mesure de performance, telle que la récompense cumulative, pour guider la recherche de la politique optimale. (iii) *Les algorithmes critiques d'acteurs* (en anglais "*actor-critic algorithms*") qui combinent des éléments des deux précédents. Ils utilisent une fonction de valeur pour évaluer les actions et cherchent également à optimiser une politique. Ils sont souvent utilisés dans des environnements plus complexes.

Parmi les algorithmes *RL* sans modèles les plus connues, que nous allons détailler dans le reste de cette section, nous citons : *Q-Learning* (Konda and Tsitsiklis, 1999), *SARSA (*l'acronyme de ***S**tate-**A**ction-**R**eward-**S**tate-**A**ction)* (Yang et al., 2019) et *DQN* (l'acronyme de ***D**eep **Q**-**N**etwork*) (Joo and Lim, 2021).

2.4.3.1 Q-Learning

La lettre '*Q*' indique la fonction qui estime la qualité d'une action effectuée dans un état donné de l'environnement. Le *Q-Learning* (Liao and Cheng, 2009; Chin et al., 2011) est l'algorithme *RL* le plus couramment utilisé (Wang et al., 2020). Il est appelé algorithme *hors politique* (en anglais : *Off-policy*) dans la mesure où il met à jour la valeur de *Q* de la paire *état-action* actuelle en fonction de la valeur *Q* maximale de la paire *état-action* suivante. Il convient de noter qu'au cœur de Q-*Learning* se trouve le calcul de *Q-table*, qui est définie comme *Q(s, a)* (équation 2.2). Les lignes de *Q-table* contiennent les valeurs *Q* des états tandis que les colonnes représentent des actions, qui mesurent à quel point il sera bénéfique lorsque l'état actuel est influencé par cette action (Figure 2.3). En effet, *Q-Learning* fonctionne en calculant l'action optimale qui maximise l'espérance des récompenses des états futurs, en prenant en compte un *facteur d'actualisation* (autrement dit *facteur d'apprentissage*) compris entre *0* et *1*. Ce facteur

d'actualisation permet de déterminer l'importance des états futurs dans l'étude. Un facteur proche de 0 ne considère que les états présents tandis qu'un facteur proche de 1 accorde plus d'importance aux étapes futures.

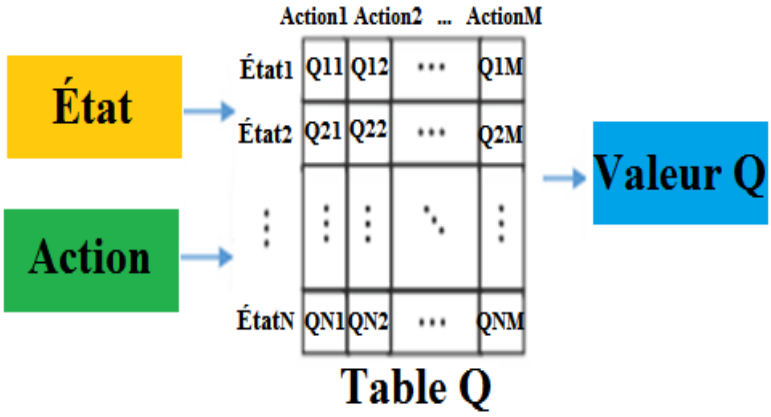


Figure 2.3 Architecture de Q-learning (Dalla Pozza et al., 2022).

Le *Q-Learning* permet à un agent d'apprendre une stratégie optimale dans un environnement dynamique. Il consiste à apprendre une fonction Q qui donne la valeur de chaque *état-action*, c'est-à-dire la somme des récompenses futures attendues en partant de cet état et en prenant cette action. La mise à jour de la fonction Q est basée sur l'équation de *Bellman*, qui exprime la relation de récurrence entre la valeur d'un état et la valeur de ses successeurs. Cette valeur action-état est initialement estimée arbitrairement par une fonction $Q : A \times S \rightarrow R$ comme suit :

$$Q(s,a) = R(s,a) + \gamma^* \max(Q(s',a')) \tag{2.2}$$

Où :

- $Q(s,a)$ est la valeur de l'état-action (s,a)
- $R(s,a)$ est la récompense obtenue en prenant l'action a à l'état s
- γ est un facteur d'actualisation $(0 \leq \gamma \leq 1)$ qui donne l'importance accordée aux récompenses futures par rapport à celles immédiates
- $\max(Q(s',a'))$ est la valeur maximale des états successeurs s' de l'état s après avoir pris l'action a'

Ensuite cette valeur est mise à jour à chaque étape de l'apprentissage en utilisant l'équation suivante :

$$Q(s,a) = Q(s,a) + a^*(R(s,a) + \gamma^* \max\left(Q(s',a')\right) - Q(s,a)) \tag{2.3}$$

Où :

- a est le taux d'apprentissage $(0 \leq a \leq 1)$ qui détermine l'importance accordée aux nouvelles informations par rapport aux anciennes

La mise à jour de la fonction *Q* est effectuée après chaque action effectuée par l'agent dans l'environnement. L'agent choisit son action en sélectionnant l'action ayant la plus grande valeur de *Q* pour l'état courant *s*. Cela peut être fait en utilisant la politique d'*exploration-exploitation*, qui consiste à choisir aléatoirement une action avec une probabilité ε (ε-*greedy*) et à choisir l'action ayant la plus grande valeur de *Q* avec une probabilité 1-ε. Le processus d'apprentissage continue jusqu'à ce que la fonction *Q* converge vers une solution stable. L'algorithme *Q-Learning* peut être décrit par le pseudo code suivant :

Algorithme 2.1. Q-learning

entrée : $\alpha > 0$ et un petit $\varepsilon > 0$, taux $\gamma > 0$

Initialisation
Initialiser de façon arbitraire Q[*s, a*] pour tous les états-actions possibles

Début
1) Pour *chaque épisode* **faire**

a) **Initialiser** l'environnement et l'état initial s.
b) **Répéter** (pour chaque étape de l'épisode)
   i) ***Sélectionner*** une action *a* à partir de l'état *s* en utilisant la politique dérivée de *Q* (ex.ε-greedy).
   ii) ***Exécuter*** l'action a
   iii) ***observer*** la récompense *r* et le nouvel état *s'*.
   iv) ***Mettre à jour*** la fonction *Q* en utilisant l'*équation* (*2.3*).
   v) ***Mettre à jour*** l'état courant s=s'.
   vi) ***Réduire*** la valeur de ε pour réduire l'exploration au fil du temps.
c) **jusqu'à** ce que *s* soit l'état terminal

Fin

Q-Learning est très efficace pour les problèmes où l'espace d'états est discret et fini. Il a été largement appliqué pour résoudre les problèmes *ATSC* dans de nombreuses recherches (Eom and Kim, 2020). En outre, pour les problèmes avec un espace d'états continu, il est possible d'utiliser une version continue du *Q-Learning* appelée Q-*Learning avec approximation de fonction* (en anglais : *Q-Learning with function approximation*), qui utilise une fonction d'approximation pour estimer la fonction *Q* (Melo and Ribeiro, 2007).

2.4.3.2 SARSA

L'algorithme *SARSA* est l'acronyme de (***S****tate-****A****ction-****R****eward-****S****tate-****A****ction*). Il s'agit d'une méthode *RL* sur politique (en anglais : *on-policy),* qui permet à un agent d'apprendre une politique optimale en interagissant avec son environnement. L'agent choisit une action en fonction de l'état courant, observe la récompense correspondante

et l'état suivant, puis met à jour sa fonction d'évaluation de la politique. L'algorithme *SARSA* est basé sur la méthode de la *descente de gradient* (en anglais *gradient descent*) et utilise une table de valeurs d'action pour chaque état possible. La méthode de *descente de gradient* est une méthode d'optimisation qui permet de minimiser une fonction de coût en ajustant itérativement les paramètres d'un modèle. L'idée de base de telle méthode est de calculer le *gradient* de la fonction par rapport à ses paramètres, puis de mettre à jour les paramètres dans la direction opposée du *gradient* pour minimiser la fonction. Plus précisément, à chaque étape de l'algorithme, on calcule le *gradient* de la fonction, puis on le multiplie par un petit pas appelé *taux d'apprentissage*, et on utilise cette quantité pour mettre à jour les paramètres. En répétant ce processus pour un grand nombre d'itérations, l'algorithme converge généralement vers un minimum local de la fonction.

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha\,[r_{t+1} + \gamma\, Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)] \qquad (2.4)$$

Où :

- $Q(s_t, a_t)$ est la valeur de l'état-action (s_t, a_t)
- r_{t+1} est la récompense obtenue en prenant l'action a à l'état s
- γ est un facteur d'actualisation ($0 \leq \gamma \leq 1$) qui donne l'importance accordée aux récompenses futures par rapport à celles immédiates
- s_{t+1} est le nouvel état obtenu après avoir effectué l'action a_t dans l'état s_t
- a_{t+1} est la prochaine action choisie par l'agent dans l'état s_{t+1}

L'algorithme *SARSA* peut être décrit par le pseudo code suivant :

Algorithme 2.2. SARSA

entrée : α > 0 et un petit ε > 0, taux γ > 0

Initialisation
Initialiser de façon arbitraire Q[*s, a*] pour tous les états-actions possibles

Début
2) Pour *chaque épisode* **faire**
  a) **Initialiser** l'environnement et l'état initial s.
  b) **Répéter** (pour chaque étape de l'épisode)
    i) ***Sélectionner*** une action *a* à partir de l'état *s* en utilisant la politique dérivée de *Q* (ex.ε-greedy).
    ii) ***Exécuter*** l'action a
    iii) ***observer*** la récompense *r* et le nouvel état *s'*.
    iv) ***Mettre à jour*** la fonction *Q* en utilisant l'*équation* (2.4).
    v) ***Mettre à jour*** l'état courant s=s'.
    vi) ***Réduire*** la valeur de ε pour réduire l'exploration au fil du temps.
  c) **jusqu'à** ce que *s* soit l'état terminal
Fin

2.4.3.3 DQN

L'algorithme *DQN* (l'acronyme de *Deep Q-Network*) est une méthode d'apprentissage par renforcement profond (en anglais *DRL : Deep Reinforcement Learning*) proposé par Mnih et al. (2015). Il combine également l'apprentissage par renforcement et les réseaux de neurones profonds pour apprendre une fonction *Q* de qualité d'action.

Le principe de base *DQN* est de construire une fonction *Q* qui estime la récompense attendue pour chaque action possible à partir de chaque état. La fonction *Q* est apprise en maximisant la récompense totale attendue de l'action, qui est définie comme la somme des récompenses futures actualisées. *DQN* utilise un réseau de neurones profonds pour estimer la fonction *Q*. Le réseau de neurones prend l'état actuel comme entrée et renvoie une valeur *Q* pour chaque action possible. La fonction *Q* est donc une approximation de la valeur de l'état-action qui prend en compte les récompenses futures potentielles. L'apprentissage de la fonction *Q* se fait à l'aide de l'algorithme *Q-Learning*. D'autre part, le réseau de neurones est entraîné à partir d'un ensemble d'expériences stockées dans une mémoire de relecture. Cette mémoire stocke une série d'observations passées (*état, action, récompense, état suivant*) que l'agent a rencontrées lors de son exploration de l'environnement. Le réseau de neurones est entraîné à minimiser l'erreur quadratique moyenne entre la sortie du réseau et la valeur cible calculée à partir des expériences stockées dans la mémoire de relecture.

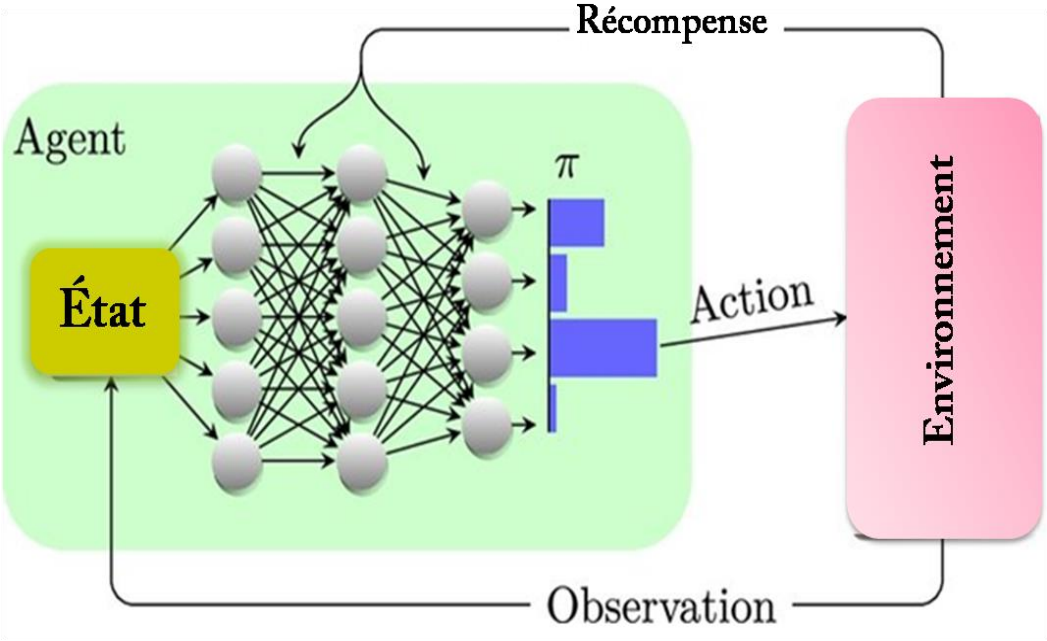


Figure 2.4 Architecture de DQN (Dalla Pozza et al., 2022).

Comme il est illustré sur la Figure 2.4, l'architecture de l'algorithme *DQN* adopte un réseau de neurones profond pour permettre l'apprentissage de la politique π que l'agent utilise pour effectuer une action sur l'environnement. Une récompense et les informations sur le nouvel état du système sont rendues à l'agent qui s'améliore et apprend sa politique en conséquence.

DQN se base principalement sur le calcul de plusieurs fonctions utilisant ainsi des équations importantes à savoir :

La *fonction de perte*, décrite par l'équation (2.5), qui représente l'erreur quadratique moyenne entre la valeur cible de la fonction Q et sa valeur prédite. La valeur cible est calculée en utilisant l'équation de *Bellman*, et la valeur prédite est la sortie du réseau de neurones.

$$L(s_t, a_t, r_{t+1}, s_{t+1}, \theta) = \left(r_{t+1} + \gamma \max_a Q(s_{t+1}, a, \theta) - Q(s_t, a_t, \theta)\right)^2 \qquad (2.5)$$

La *fonction de perte* est optimisée en ajustant les paramètres θ du réseau de neurones. Le paramètre θ dans l'algorithme *DQN* représente les *poids* et les *biais* du réseau de neurones profond qui est utilisé pour estimer la fonction Q. D'autre part, les paramètres θ sont ajustés de manière itérative au fur et à mesure que de nouvelles transitions sont collectées et que le réseau de neurones est mis à jour. Par conséquent, la mise à jour des paramètres θ se fait en utilisant la *descente de gradient stochastique* avec un taux d'apprentissage α comme suit :

$$\theta' = \theta - \alpha^* \nabla(L(\theta)/\nabla\theta) \qquad (2.6)$$

Où

- θ' sont les nouveaux paramètres,
- α est le *taux d'apprentissage*,
- $L(\theta)/$ est la *fonction de perte*,
- $\nabla(L(\theta)/\nabla\theta)$ est le gradient de la *fonction de perte* par rapport à θ.

Le taux d'apprentissage *a* contrôle la taille de chaque pas de mise à jour, tandis que le gradient de la *fonction de perte* par rapport aux paramètres θ indique la direction dans laquelle les paramètres doivent être ajustés.

Par ailleurs, l'algorithme *DQN* peut être difficile à stabiliser en raison de la nature instable de l'algorithme d'optimisation et de l'interaction continue entre l'agent et l'environnement. Pour y remédier, plusieurs techniques ont été proposées dans la littérature comme :

- *Reprise d'expérience* (en anglais *Experience Replay*) : Cette technique s'appuie principalement sur une mémoire de stockage (buffer de reprise) notée *M*, qui sert

également à mémoriser les expériences de l'agent pendant le processus d'entrainement. Une expérience représente le comportement de l'agent à un pas de temps donné. Du fait que la taille de *M* est limitée, il est utile d'adopter une technique de remplacement de mémoire (Long-Ji, 1992). En effet, lorsque la mémoire *M* est pleine, il semble nécessaire de remplacer certaines des transitions les plus anciennes pour libérer de l'espace pour de nouvelles transitions. La technique de remplacement de mémoire la plus couramment utilisée est la méthode du *FIFO (First-In-First-Out*). Une illustration de la technique de reprise d'expérience dans *RL* est bien décrite sur la Figure 2.5.

Algorithme 2.3. DQN avec Reprise d'expérience

Initialiser la mémoire de lecture *M* à la capacité *N*
Initialiser la fonction action-valeur *Q* avec des poids θ
Initialiser la fonction action-valeur cible Q' avec des poids $\theta' = \theta$

Début
  Pour *épisode = 1 à M* **faire**
    Initialiser la séquence s1 = {x1} et prétraiter la séquence $\phi_1 = \phi(s_1)$
    Pour *t = 1 à T* **faire**
      Avec probabilité ε sélectionner une action aléatoire a_t
      Sinon sélectionnez $a_t = argmax_a\ Q(\phi(s_1), a;\ \theta)$
      Exécutez l'action a_t et observez la récompense r_t et x_{t+1}
      Définir $s_{t+1} = s_t, a_t, x_{t+1}$ et pré procédés $\phi_{t+1} = \phi(s_{t+1})$
      Stocker la transition $(\phi_1, a_t, r_t, \phi_{t+1})$ dans *M*
      Définir $y_j = \begin{cases} r_j & \text{si l'episode se termine à l'etape } j+1 \\ r_j + \gamma max_{a'}\ Q'(\phi_{t+1}, a';\ \theta') & \text{sinon} \end{cases}$
      Mettre à jour la fonction *Q* en utilisant l'*équation* (*2.5*).
      Toutes les étapes *C* réinitialisées $Q' = Q$
    Fin Pour
  Fin Pour
Fin

Une variante de cette technique baptisée *reprise d'expérience prioritaire* (en anglais *Prioritized Experience Replay)*, où les transitions sont stockées avec une priorité qui est basée sur leur importance pour l'apprentissage de l'agent. Les transitions les plus importantes sont échantillonnées plus fréquemment et ont donc une probabilité plus élevée d'être conservées dans *M*.

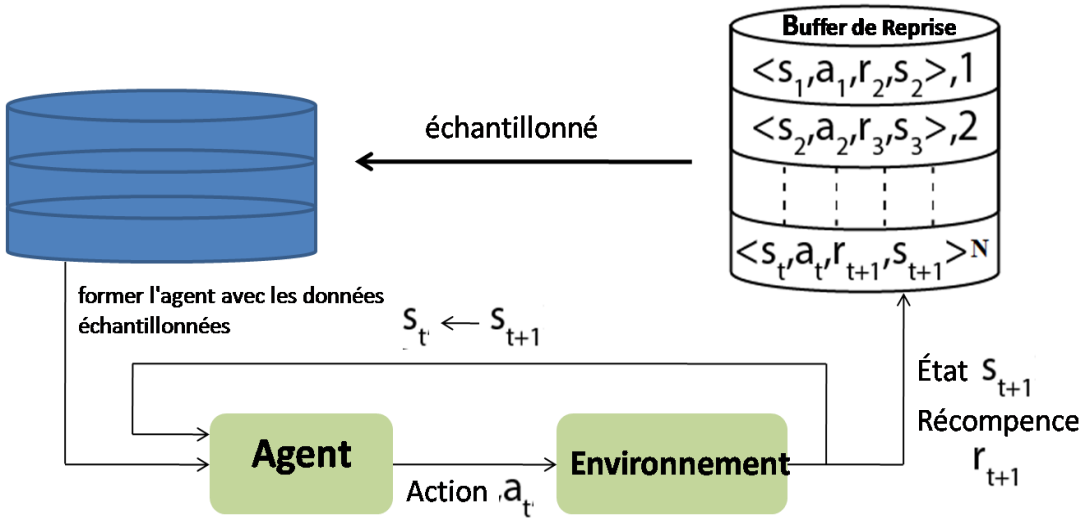


Figure 2.5 Reprise d'expérience (Lee and Lee, 2020).

- *Réseau cible* (en anglais *Target Network*) : La création d'un second réseau de neurones cible qui a les mêmes paramètres que le réseau de neurones principal, mais avec des poids figés pour une certaine période (Gao et al., 2017). Cela permet d'utiliser les prévisions précédentes pour la mise à jour des poids, tout en réduisant la corrélation entre les valeurs cibles et les valeurs prévues.
  Le réseau de neurones principal est le réseau qui est mis à jour à chaque étape de l'apprentissage. Il prend l'état actuel de l'environnement en entrée et produit une estimation de la valeur *Q* pour chaque action possible. La valeur *Q* est la mesure de la qualité de chaque action dans cet état et est utilisée pour guider la politique de l'agent.
  Le réseau de neurones cible est utilisé pour produire des valeurs cibles pour l'apprentissage. Il est identique au réseau de neurones principal au début de l'apprentissage, mais il est mis à jour de manière beaucoup plus lente. En effet, il est copié périodiquement depuis le réseau principal à intervalles fixes, après un certain nombre d'itérations d'apprentissage. Les poids du réseau cible sont figés entre chaque copie, ce qui permet de stabiliser l'apprentissage.
  L'utilisation d'un réseau de neurones cible est importante pour éviter les oscillations ou les divergences lors de l'apprentissage par renforcement profond.
- *Double Q-Learning* : La séparation de la sélection d'action et de l'estimation de la valeur en utilisant deux réseaux de neurones différents, l'un pour sélectionner l'action et l'autre pour estimer la valeur de cette action (Gu et al., 2020).

  A chaque étape d'apprentissage, l'algorithme sélectionne l'action *a'* en utilisant le premier réseau de neurones, puis utilise le second réseau de neurones pour estimer la valeur *Q* de cette action. La formule de mise à jour devient alors :

$$Q(s,a) \leftarrow Q(s,a) + \alpha[\, r + \gamma\, Q'\big(s', argmax\, Q(s',a')\big) - \, Q(s,a)] \qquad (2.7)$$

Où:

- $Q'\big(s', argmax\, Q(s',a')\big)$ *est l'estimation de la valeur Q de la meilleure action a' dans l'état s'. Cette valeur est calculée en utilisant le second réseau de neurones.*

En utilisant deux réseaux de neurones distincts pour la sélection et l'estimation de la valeur *Q*, *Double Q-Learning* permet d'éviter le biais de maximisation et d'améliorer les performances de l'algorithme *Q-Learning* classique. Le pseudo code suivant décrit l'algorithme de *Double Q-Learning.*

Algorithme 2.4. Double DQN

Entrée : mini-lot *k*, étapes η, période de relecture *K* et taille *N*, exposantes α, β, *T*
Initialiser la mémoire de lecture $M = \emptyset$, $\Delta = 0$, $p_1 = 1$
Observer S_0 et choisir $A_0 \sim \Pi_\theta(S_0)$

Début
  Pour *t = 1 à T* **faire**
    Observer S_t, R_t, γ_t
    Stocker la transition $(S_{t-1}, A_{t-1}, R_t, \gamma_t, S_t)$ dans *M* avec priorité maximale
      $p_t = max_i <_t p_i$
    if t = 0 mod *K* **then**
      Pour *j = 1 à k* **faire**
        Exemple de transition $j \sim P(j) = p_i^{\alpha} \,/\, \sum_i p_i^{\alpha}$
        Calculer le poids d'échantillonnage $w_j = \frac{(N \,.P(j))^{-\beta}}{max_i\, w_i}$
        Mettre à jour la fonction δ_j =*Q* en utilisant l'*équation* (*2.7*).
        Mettre à jour la priorité de transition $p_j \leftarrow |\delta_j|$
        Accumuler le changement de poids $\Delta \leftarrow \Delta + w_j \,.\, \delta_j \,.\, \nabla_\theta\, Q(S_{t-1}, A_{t-1})$
      Fin Pour
      Mettre à jour $\theta \leftarrow \theta + \eta \cdot \Delta$, réinitialiser $\Delta = 0$
     Toutes les étapes *C* réinitialisées $Q' = Q$
    Fin **if**
    Choisir une action $A_t \sim \Pi_\theta(S_t)$
  Fin Pour
Fin

- *Dueling Network* : L'algorithme *Dueling Network* est un algorithme d'apprentissage par renforcement profond utilisé pour estimer la fonction d'action-valeur *Q* dans un environnement de décision séquentielle. Il a été introduit en 2016 par Wang et al. (2016). La particularité de tel algorithme est qu'il utilise un réseau de neurones qui sépare l'estimation de la valeur de l'état et la valeur de l'action, en utilisant deux couches de neurones distinctes. Cette séparation permet d'estimer séparément l'importance de l'état et de l'action, ce qui peut améliorer la stabilité et la qualité de l'apprentissage. Le réseau de neurones utilisé par *Dueling Network* a la même

structure que les réseaux de neurones profonds classiques utilisés dans *RL*, mais avec deux sorties distinctes. La première sortie représente la valeur de l'état, tandis que la seconde sortie représente l'avantage de chaque action par rapport à la moyenne des avantages de toutes les actions possibles. La valeur *Q* de chaque action est ensuite calculée en combinant la valeur de l'état et l'avantage de l'action de la manière suivante :

$$Q(s,a) = V(s) + A(s,a) - \frac{1}{N} \sum A(s,a') \tag{2.8}$$

où :

- s *est l'état,* a *est l'action,*
- $V(s)$ *est la valeur de l'état s,*
- $A(s,a)$ *est l'avantage de l'action a dans l'état s,*
- $\frac{1}{N} \sum A(s,a')$*est la moyenne des avantages de toutes les actions possibles dans l'état s,*
- *N est le nombre total d'actions possibles dans l'état s.*

Dueling Network utilise ensuite cette fonction *Q* pour entraîner le réseau de neurones en utilisant l'algorithme de descente de gradient stochastique (en anglais *SGD : Stochastic Gradient Descent*) avec une fonction de perte appropriée. La fonction de perte est définie comme la différence entre la valeur cible de la fonction *Q* (c'est-à-dire la somme de la récompense et de la valeur *Q* de l'état suivant) et la valeur *Q* prédite par le réseau de neurones.

Lorsqu'il est appliqué à la gestion des feux de signalisation, l'algorithme *Dueling Network* peut contrôler efficacement le temps de chaque feu en fonction de l'état actuel de la circulation. Le modèle peut ainsi apprendre à allouer le temps de manière adaptative en fonction des fluctuations du trafic et des changements de la situation du réseau routier.

2.4.4 Synthèse

Dans le contexte de la présente thèse, de nombreux chercheurs ont appliqué avec succès les algorithmes précédents à savoir *SARSA* (Kekuda et al., 2021; Thorpe and Anderson, 1996), *Q-Learning* (Araghi et al., 2013), etc. à la fois pour une simple intersection et pour plusieurs intersections. Cependant, le cadre entièrement centralisé où un seul agent contrôle toutes les intersections coûte des ressources informatiques considérables et souffre de problèmes d'évolutivité. En outre, la contrainte de la dimension dans les réseaux de trafic à grande échelle pose de sérieux problèmes, cela signifie que plus il y a d'intersections, de routes et de véhicules impliqués, plus le problème de gestion du trafic devient complexe et difficile à résoudre de manière efficace en utilisant un seul agent centralisé. Pour utiliser *RL* dans un réseau de trafic à grande échelle, le cadre entièrement décentralisé où un seul agent ne contrôle qu'une

seule intersection offre une meilleure évolutivité et nous conduit naturellement à considérer l'apprentissage par renforcement multi-agents (en anglais *MARL : Multi-Agent Reinforcement Learning*) où les agents interagissent non seulement avec l'environnement, mais également avec d'autres agents. La stratégie la plus simple de *MARL* est celle des agents indépendants (*IA*), où chaque agent ne prend en compte que sa propre situation et maximise sa propre récompense (Wang et al., 2021b).

2.5 Apprentissage par renforcement multi-agents

L'apprentissage par renforcement multi-agents (Eom and Kim, 2020; Zhang et al., 2022) est un axe de recherche en *ML* qui concerne la problématique de recherche liée à l'entraînement de plusieurs agents autonomes à prendre des décisions dans un environnement partagé. Dans le cadre du *MARL*, chaque agent doit apprendre à maximiser sa récompense individuelle tout en prenant en compte les actions des autres agents et leur impact sur l'environnement partagé.

Les problèmes de *MARL* sont souvent plus complexes que les problèmes de *RL* classiques, car les agents doivent apprendre à interagir les uns avec les autres de manière coopérative ou compétitive. Les défis principaux sont notamment la coordination des activités des différents agents, l'apprentissage de politiques efficaces en présence d'autres agents et la gestion des interactions complexes entre les agents.

Il existe plusieurs approches pour *MARL*, allant des approches centrées sur la coordination entre agents (comme *l'apprentissage par renforcement multi-agents coopératif*) à des approches basées sur la compétition (comme *l'apprentissage par renforcement multi-agents compétitif*). Ces approches peuvent être utilisées pour résoudre différents types de problèmes, tels que les jeux à plusieurs joueurs, la gestion et le contrôle de trafic, la gestion des ressources, etc. En d'autres termes, l'apprentissage *compétitif*, *coopératif* et *distribué* sont des variations de *MARL* qui se différencient par la façon dont les agents interagissent entre eux (Buşoniu et al., 2010).

Contrairement à *RL* traditionnel où un seul agent interagit avec un environnement, *MARL coopératif* (Cui and Zhang, 2021) implique plusieurs agents qui coopèrent pour maximiser une récompense globale. Dans ce contexte, chaque agent a sa propre politique d'action, qui détermine les actions qu'il prend en réponse à l'état de l'environnement. Les agents peuvent communiquer entre eux pour partager des informations ou pour coordonner leurs actions. Il est utile de noter que l'un des principaux défis de telle approche (*MARL coopératif*) est de trouver un compromis entre la maximisation de la récompense globale et l'optimisation des politiques individuelles de chaque agent. L'une des approches les plus couramment adoptées pour lever de tel

défi est le *DQN* ou l'utilisation de méthodes basées sur la théorie des jeux (Abdelghaffar et al., 2016).

À l'inverse à *MARL coopératif*, dans *MARL compétitif* (Deka and Sycara, 2021), les agents cherchent à maximiser leur propre récompense individuelle en essayant de surpasser les autres agents. Dans ce cadre, chaque agent a sa propre politique d'action, qui détermine les actions qu'il prend en réponse à l'état de l'environnement. Les agents peuvent interagir directement entre eux, ce qui signifie que les actions d'un agent affectent les états futurs de l'environnement pour les autres agents. Ainsi, l'un des défis majeur de *MARL compétitif* est de trouver un équilibre entre l'exploration et l'exploitation. Les approches les plus courantes pour résoudre ce problème sont la conception d'algorithmes d'apprentissage spécifiques, tels que le *Policy Gradient* (Mousavi et al., 2017) ou le *Proximal Policy Optimization* (*PPO*) (Schulman et al., 2017).

D'autre part, *MARL distribué* (Heredia and Mou, 2019) implique des agents qui apprennent en interagissant avec leur environnement mais aussi en communiquant et en échangeant des informations entre eux. Les agents doivent apprendre à s'adapter à un environnement dynamique et incertain, et à prendre des décisions qui prennent en compte les actions des autres agents. Un exemple courant de *MARL distribué* est les voitures autonomes qui communiquent pour éviter les collisions.

2.5.1 Revue de la littérature : Approches basées sur *MARL* pour le contrôle de trafic à multiple intersections

Dans le domaine de recherche de cette thèse, un grand nombre de recherches ont annoncé la combinaison de *MAS* avec *RL*, proposant ainsi des approches à base de *MARL*, qui constitueraient des solutions assez puissantes pour un contrôle global de plusieurs intersections (Kim and Jeong, 2019). Selon les différentes structures d'informations, deux classes d'approches basées *MARL* ont été considérées dans la littérature : *MARL centralisés* et *MARL décentralisés*. Les approches *centralisées* entrainent un agent central global pour contrôler les actions de toutes les intersections (Wang et al., 2021a). Bien que ces approches produisent des résultats acceptables sur les petits réseaux routiers, elles restent très difficiles à appliquer sur les réseaux à grande échelle (Wei et al., 2019). Face à ce problème, différentes approches *décentralisées* ont été proposées dans lesquelles un ensemble d'agents *RL* indépendants contrôlent plusieurs intersections. Chaque agent *RL* s'entraîne à contrôler une seule intersection en observant et en percevant seulement des fragments du réseau routier (Liu et al., 2021).

Dans notre contexte, il est évident de noter que l'impact environnemental des décisions de tout agent dépend des décisions prises par les autres agents. De plus, le

niveau de coopération des agents influence, sans aucun doute, les performances globales du système. Ainsi, l'action de tout agent devrait être coordonnée avec les autres afin d'atteindre les résultats attendus. De nos jours, de nombreux chercheurs ont proposé de nombreuses solutions, comme Mannion et al. (2016) qui ont développé des algorithmes basés sur *MARL* tout en comparant leurs performances avec un algorithme de contrôle à temps fixe. Qu et al. (2020) ont proposé une nouvelle solution basée sur *MARL* fondée sur un équilibre mixte de stratégie Nash régionale. Les résultats de l'expérience dans un réseau de trafic de grille montrent l'efficacité de cette solution. Kuyer et al. (2008) ont développé une méthode de contrôle de signalisation routière coordonnée basée sur des diagrammes de collaboration. Quant au travail de recherche proposé par Wang et al. (2021a), un nouveau cadre appelé *CGB-MARL* (*Cooperative Group-Based Multi-Agent Reinforcement Learning*) est introduit pour coordonner la signalisation routière dans un réseau routier à grande échelle. Ce cadre repose principalement sur un système coopératif de véhicules-infrastructure. De plus, un nouvel algorithme *CGB-MAQL (Cooperative Group-Based Multi-Agent Q-Learning*) est développé dans ce travail de recherche, dont les résultats expérimentaux démontrent l'efficacité de ce dernier pour contrôler de multiples cas d'intersection, notamment en termes de la diminution de la congestion et la protection de l'environnement. Dans le travail de recherche de Bakker et al. (2010), un algorithme basé sur des graphes de coordination a été proposé pour décrire les dépendances entre les agents *RL* voisins afin d'améliorer la coopération entre eux, notamment dans les réseaux à grande échelle. El-Tantawy et al. (2013) proposent une méthode efficace dans laquelle les évaluations effectuées par chaque agent sont limitées à son voisinage, et chaque agent choisit à tout moment l'action qui maximise la fonction de récompense dans son voisinage.

Depuis l'introduction de l'algorithme *Q-Learning*, de nombreuses approches basées sur *MARL* ont montré leur efficacité pour résoudre les problèmes de contrôle de la signalisation routière (Abdulhai et al., 2003; Camponogara and Kraus, 2003; Lu Shoufeng et al., 2008; Salkham et al., 2008; Joo et al., 2020). Au départ, cet algorithme a été utilisé dans diverses solutions sans coordination entre les agents (Schneider et al., 1999). Par la suite, il a été adopté en considérant l'aspect synergique entre les agents où l'agent local est impacté par les voisins adjacents grâce à un signal de rétroaction intégré dans la mise à jour de la fonction *Q* locale.

Dans cette direction, diverses solutions ont été proposées dans la littérature, montrant l'efficacité de cette idée. Ozan et al. (2015) ont introduit une méthode pour optimiser les phases de signalisation prédéfinies dans un environnement multi-intersections en utilisant l'algorithme *Q-Learning*. Xu et al. (2013) ont proposé une solution qui ajuste l'algorithme de sélection d'actions dans le *Q-Learning* en intégrant

une estimation bayésienne de la probabilité de sélection d'actions pour d'autres agents. Ge et al. (2019) ont proposé un réseau de *Deep Q-Learning* coopératif avec transfert de valeur *Q* (en anglais *QT-CDQN : Cooperative Deep Q-Network with Q-value Transfer*) pour contrôler plusieurs intersections. Dans une telle proposition, l'apprentissage de la politique de processus de chaque agent est principalement influencé par les dernières actions de ses voisins. Ainsi, les *Q-valeurs* optimales des voisins sont transmises à la fonction de perte du réseau *Q*. Dans Liu and Ding (2022), les auteurs ont proposé une approche *DRL* distribuée, qui consiste en un apprentissage local et un consensus global. Les résultats expérimentaux ont démontré la supériorité de cette approche par rapport à la stratégie de contrôle à temps fixe, à l'apprentissage centralisé et aux algorithmes d'apprentissage local. Dans le même contexte, Wu et al. (2022) ont proposé deux nouveaux algorithmes basés *MARL* exploitant l'équilibre de *Nash* et *RL*. Les résultats expérimentaux ont montré que ces algorithmes ont des performances cohérentes et exceptionnelles dans la réduction du temps d'attente moyen au niveau global. Kim and Jeong (2019) a proposé une approche coopérative de contrôle de signalisation routière en utilisant la prédiction de flux de trafic pour plusieurs intersections. Un modèle basé *MARL*, permettant à chaque agent de partager ses informations de trafic avec ses voisins, est défini. Cette approche peut aider à dériver la valeur *Q* optimale globale en prédisant le flux entrant des autres intersections. Wei et al. (2019) ont proposé un nouveau modèle appelé *Colight*, qui utilise des réseaux d'attention graphique dans le cadre de l'apprentissage par renforcement pour faciliter la communication. Les résultats expérimentaux ont montré que, en apprenant la communication, *Colight* peut atteindre des performances supérieures à celles d'autres nombreuses méthodes proposées dans la littérature. Un autre travail de recherche proposé par Li et al. (2021), qui a proposé une nouvelle méthode fondée sur *MARL*, appelée *KSDDPG* (*Knowledge Sharing Deep Deterministic Policy Gradient*) pour atteindre un contrôle optimal en améliorant la coopération entre les signaux de circulation. Dans le travail de recherche de Tan et al. (2019), les auteurs ont supposé que la valeur *Q globale* peut être décomposée en un certain nombre de valeurs *Q locales* et que les différentes valeurs *Q globales* sont mises à jour de manière centralisée. Dans Chu et al. (2020a), un nouveau protocole de communication différentiable, appelé *NeurComm*, est proposé pour permettre à des agents *RL* indépendants d'améliorer la valeur Q globale. En effet, les observations locales et les messages des voisins constituent la base sur laquelle chaque agent apprend sa politique de contrôle.

Par ailleurs, dans leur travail, (Wang et al., 2021b) ont proposé une nouvelle méthode fondée *MARL*, appelée (*Co-DQL : Cooperative Double Q-Learning*), qui présente plusieurs caractéristiques remarquables. Elle utilise la méthode *DQL* (*Double Q-Learning*)

indépendante hautement évolutive basée sur des estimateurs doubles et la politique dite borne supérieure de confiance (en anglais *UCB : Upper Confidence Bound*) (Radović and Erceg, 2021), qui peut éliminer le problème de surestimation existant dans le *Q-Learning* indépendant traditionnel tout en assurant l'exploration. *Co-DQL* est appliqué au contrôle des feux de signalisation et testé sur divers scénarios de flux de trafic simulés. Les résultats montrent que *Co-DQL* surpasse les algorithmes *MARL* décentralisés en termes de plusieurs métriques de trafic définies dans leur travail.

Parmi les travaux récents dans ce contexte, nous distinguons le travail de (Wang et al., 2023a), dont les auteurs ont proposé un algorithme basé *MARL* avec réseau acteur-attention-critique pour le contrôle des feux de signalisation. Dans cet algorithme baptisé (*MAAC-TLC : Multi-Agent deep reinforcement learning with Actor-Attention-Critic network for Traffic Light Control*), chaque agent introduit un mécanisme d'attention dans le processus d'apprentissage, de sorte qu'il ne prête pas attention à toutes les informations des autres agents de manière indiscriminée, mais se concentre uniquement sur les informations importantes des agents qui jouent un rôle important dans le processus, afin de garantir que toutes les intersections peuvent apprendre la politique optimale.

Malgré les recherches que nous avons discutées dans cette section, qui prouvent l'utilité de l'aspect coopération dans le contexte de la gestion et le contrôle du trafic routier en temps réel, l'efficacité d'une telle coopération pour le contrôle multi-intersections reste encore une direction de recherche ouverte à l'investigation. À notre connaissance, il existe peu de recherches sur le *DRL*, qui adoptent le *DQN* avec la valeur *Q*, pour le contrôle coopératif de la gestion du trafic dans les réseaux routiers urbains à multiple intersections.

Nous proposons, dans cette thèse de Doctorat, trois contributions dont les deux premières apportent des solutions au problème de contrôle adaptatif des feux de signalisation pour les intersections isolées. Quant à la dernière, qui représente notre plus importante contribution, propose une nouvelle approche coopérative basée sur le *DRL* pour le contrôle adaptatif des feux de signalisation dans un réseau routier à multiple intersections. Une telle contribution prend en compte les informations du trafic des intersections voisines en partageant les *récompenses*, les *états* et les valeurs d'*actions*. Par conséquent, la valeur *Q* optimale globale de plusieurs intersections est estimée en fonction de ces valeurs partagées. En effet, le réseau routier en question est d'abord modélisé comme un *MARL*. Chaque agent contrôle une intersection via une *DQN* et transfère les dernières récompenses, états et actions obtenus de ses voisins vers sa propre fonction de perte pendant le processus d'apprentissage. En fait, les chapitres suivants détailleront bien les spécificités de ces trois contributions.

2.6 Conclusion

En conclusion, le domaine de *RL* montre un grand potentiel pour le contrôle du trafic routier, offrant une alternative prometteuse aux méthodes traditionnelles basées sur la planification et la surveillance. Les approches de *RL* ont prouvé leur efficacité pour résoudre des problèmes complexes de contrôle de trafic, tels que la gestion des intersections, la régulation du flux de véhicules et la réduction des congestions routières.

Dans ce chapitre, nous avons exploré le domaine de *RL* appliqué au contrôle du trafic routier, en mettant l'accent particulier sur le *DRL*. Nous avons également présenté les méthodes fondées *RL*, qui peuvent être classées en deux grandes classes à savoir : les méthodes basées sur les modèles et les méthodes sans modèle. Enfin, nous avons dressé un état de l'art sur les approches basées *MARL* pour le contrôle du trafic dans un réseau routier à multiple intersections.

En d'autres termes, nous pouvons conclure que le *DRL* appliqué au *TSC* est un domaine prometteur en pleine expansion. Toutefois, il est crucial de faire des choix de conception judicieux lors de la création d'un agent *RL* pour le *TSC*. Ce chapitre a présenté une synthèse des études de la littérature sur le sujet, y compris les approches d'état, de représentation des actions et des récompenses, ainsi que les conceptions d'agents *RL*. Nous avons également discuté des approches de contrôle coopératif des feux de circulation dans un réseau routier, qui ont le potentiel d'améliorer considérablement l'efficacité du trafic à une intersection en prenant en compte les conditions de circulation des intersections voisines.

Toutefois, il reste des défis à relever, tels que la garantie de la sécurité des usagers de la route, la gestion dynamiques des feux de signalisation et la gestion des interactions entre les différents agents. Pour y surmonter, il est nécessaire de poursuivre les recherches en matière *RL* et *MARL*, en développant des méthodes plus avancées pour le contrôle du trafic routier, en tenant compte de la complexité et de la dynamique du système de transport. C'est dans cette perspective que le travail de recherche de cette thèse de Doctorat se positionne en développant des approches innovantes pour le contrôle intelligent dynamique des feux de signalisation pour les intersections aussi bien isolées que multiples. Les détails de telles approches sont bien décrits dans les chapitres suivants.

Chapitre 3 : Approche réactive pour le contrôle adaptatif des feux de signalisation dans les intersections isolées

3.1 Introduction

L'optimisation du trafic routier dans les intersections isolées est un enjeu majeur pour les villes modernes. Par conséquent, les feux de signalisation sont souvent utilisés pour réguler le flux de véhicules et de piétons, mais leur contrôle est souvent statique et ne prend pas en compte les variations du trafic en temps réel. Cela peut entraîner des temps d'attente excessifs, des embouteillages, des congestions routières et des consommations accrues de carburant qui affectent négativement la propreté de l'environnement (Papageorgiou et al., 2003). Afin de résoudre ces problèmes, des approches adaptatives pour le contrôle des feux de signalisation ont été proposées, qui peuvent ajuster la durée des feux en fonction des conditions actuelles du trafic. Dans ce chapitre, nous proposons une approche réactive pour le contrôle adaptatif des feux de signalisation dans les intersections isolées, qui utilise des techniques de contrôle en boucle fermée pour s'adapter aux changements de trafic en temps réel. Nous présentons également les résultats d'une simulation pour évaluer l'efficacité de notre approche par rapport à d'autres approches proposées dans la littérature comme celles décrites dans les travaux de recherche de Rida and Hasbi (2019) et Yousef et al. (2010). L'implémentation de ces approches à l'aide de la plateforme *Netlogo* ainsi que l'exécution de plusieurs

expérimentations, en variant les scénarios du trafic montrent notre approche avec des résultats prometteurs.

3.2 Problématique des intersections isolées à feux de signalisation

Les intersections isolées, où la circulation se croise en deux directions, sont un élément clé des réseaux de transport urbain. Dans ces intersections, les feux de signalisation sont utilisés pour réguler le flux de trafic et minimiser les retards pour les automobilistes. Cependant, les systèmes de régulation des feux de signalisation actuels ont montré leurs limites en termes d'adaptabilité au trafic réel et de performances globales. En effet, les méthodes traditionnelles de contrôle de feux de signalisation basées sur un cycle de temps fixe ne tiennent pas compte de la variation de la demande de trafic en fonction de l'heure de la journée et de l'évolution des conditions de circulation en temps réel. Cette méthode peut conduire à des problèmes tels que des temps d'attente excessifs aux feux de signalisation, une congestion importante et une augmentation de la pollution de l'air.

La mise en place de systèmes de contrôle adaptatif des feux de signalisation dans les intersections isolées est donc une solution prometteuse pour améliorer la gestion du trafic dans les zones urbaines. Ces systèmes permettent une régulation dynamique des feux de signalisation en fonction de la situation de trafic réel et des conditions environnementales. Ils sont capables de s'adapter en temps réel aux fluctuations de la demande de trafic, réduisant ainsi les temps d'attente et améliorant la fluidité du trafic.

Notre problématique dans ce chapitre est donc de développer une approche réactive pour le contrôle adaptatif des feux de signalisation dans les intersections isolées, en vue d'améliorer l'efficacité et les performances du système de régulation de trafic dans les zones urbaines. Cette approche devrait être capable d'ajuster les paramètres de contrôle en temps réel en fonction des conditions de trafic, en utilisant des données en temps réel pour minimiser les temps d'attente et maximiser la fluidité du trafic. L'objectif d'une telle approche est donc la régulation des feux de signalisation plus efficace, offrant ainsi une alternative prometteuse aux méthodes à cycle fixe. L'efficacité ici concerne la maximisation du nombre de véhicules traversant l'intersection et l'optimisation du temps d'attente moyen (*AWT*) des véhicules.

La modélisation d'une intersection isolée à quatre directions implique généralement quatre voies d'approche qui convergent vers une zone de carrefour centrale dite zone de conflit. Les voies d'approche peuvent être désignées par des noms de rue ou des directions cardinales, tels que *nord*, *sud*, *est* et *ouest*. Chaque voie d'approche peut être équipée d'une ou plusieurs voies de circulation, en fonction du volume de trafic. La

zone de conflit peut être un *rond-point*, une *place circulaire*, ou simplement *une intersection à angles droits*. Les feux de signalisation sont généralement installés pour contrôler le trafic dans la zone de conflit. Les feux peuvent être configurés pour fonctionner en mode "*vert-rouge*" ou en mode "*vert clignotant-rouge*", en fonction du volume de trafic et des exigences locales. Les voies de départ sont généralement situées à la sortie de la zone de l'intersection

La Figure 3.1 schématise notre modèle adopté pour décrire une intersection isolée à quatre directions (*N* (*nord*), *S* (*sud*), *E* (*est*), *W* (*ouest*)), dont chacune conduit vers deux voies à double sens (*FR* (*avancer/tourner à droite*) et *L* (*tourner à gauche*)). Ainsi, chaque véhicule traversant l'intersection peut choisir l'une des quatre directions suivante *DR* = {*N*, *S*, *E* et *W*} et une voie *LN* = {*FR* et *L*}.Dans un tel modèle, il est possible d'identifier la voie où un véhicule circule par le couple *DR* et *LN*. En effet, en combinant les valeurs de tel couple (*DR*, *LN*), on peut dénombrer huit voies possibles à savoir: {*WFR*, *WL*, *EFR*, *EL*, *NFR*, *NL*, *SFR*, *SL*}.

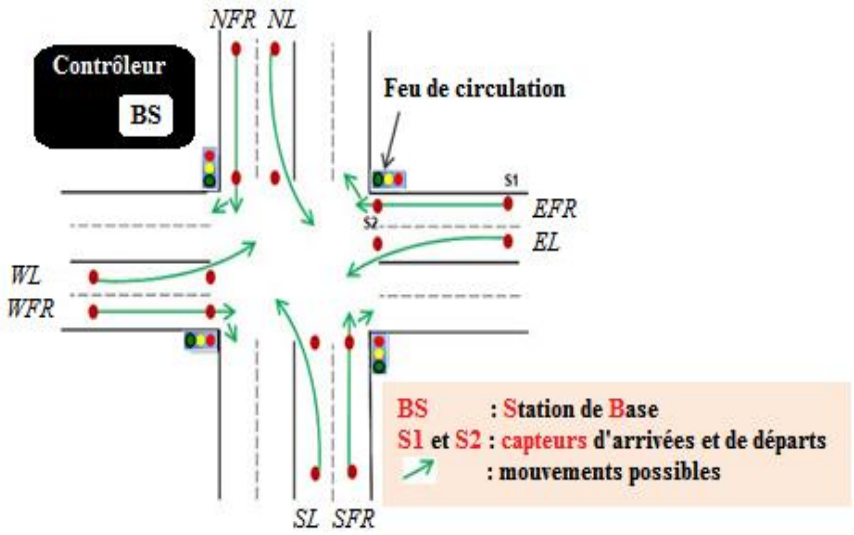


Figure 3.1 Un modèle d'intersection isolé (Faye et al., 2012).

Néanmoins, il est extrêmement important de considérer les cas de conflits qui leur paraissent nécessaires pour éviter les inter-blocages. Ainsi, la matrice des directions de conflit peut être résumée dans le tableau 3.1. Chaque colonne du tableau indique une direction dans l'intersection et son statut. Notez que s'il n'y a pas de conflit entre deux voies, la valeur 1 est attribuée, sinon la valeur 0 est attribuée. Par exemple, la direction *EL* dans la quatrième colonne est inopérante lorsque l'une ou l'autre des directions *WL* dans la deuxième ligne ou *EFR* dans la troisième ligne fonctionne. Ainsi, pour éviter les situations conflictuelles, lorsque les véhicules traversent l'intersection, le contrôleur doit verrouiller toutes les voies en conflit en utilisant les feux rouge. Sur la base des valeurs

du tableau 3.1, l'un des douze cas différents de feux verts peut être sélectionné pour une durée bien déterminée (Figure. 3.2).

Voie	*WFR*	*WL*	*EFR*	*EL*	*NFR*	*NL*	*SFR*	*SL*
WFR	—	**1**	**1**	*0*	*0*	**1**	*0*	*0*
WL	**1**	—	*0*	**1**	*0*	*0*	**1**	*0*
EFR	**1**	*0*	—	**1**	*0*	*0*	*0*	**1**
EL	*0*	**1**	**1**	—	**1**	*0*	*0*	*0*
NFR	*0*	*0*	*0*	**1**	—	**1**	**1**	*0*
NL	**1**	*0*	*0*	*0*	**1**	—	*0*	**1**
SFR	*0*	**1**	*0*	*0*	**1**	*0*	—	**1**
SL	*0*	*0*	**1**	*0*	*0*	**1**	**1**	—

Tableau 3.1 Matrice des directions de conflit.

De plus, pour identifier en temps réel le nombre de véhicules dans les voies, deux capteurs (notés *S1* et *S2*) sont placés aux deux extrémités de chaque voie (Figure 3.1). Pour acquérir une sélection correcte d'échantillons, on suppose que la distance *D* entre *S1* et *S2* doit être suffisante. L'un est installé à l'intersection et l'autre est à une distance donnée. Au total, seize nœuds de capteurs sont placés sur les huit voies pour détecter le flux de trafic.

Tous les capteurs communiquent et transfèrent les informations de trafic à la station de base qui calcule la longueur des files d'attente pour chaque direction et son temps d'attente moyen afin de contrôler le flux de trafic.
Par conséquent, face à l'évolution dynamique de l'environnement de trafic, le problème est transformé pour décider quel cas doit ensuite obtenir le feu vert et combien de temps il doit durer. L'objectif principal est d'optimiser *AWT* à l'intersection, en construisant les séquences de phases. Plusieurs solutions ont été proposées dans la littérature pour résoudre ce problème, et elles sont basées sur une variété de paramètres, y compris les temps d'attente, les longueurs de files d'attente, la vitesse et autres. La plupart des approches de la littérature privilégient attribuer le feu vert à la phase possédant le plus grand nombre de véhicules dans sa fille d'attente. A l'opposé de ces méthodes, nous proposons une solution basée sur le choix de phase en fonction à la fois de l'*AWT* et du nombre de véhicules dans les files d'attente. L'approche proposée cible principalement la sélection de la phase qui réduit au maximum la longueur moyenne des files d'attente (*AQL*) et l'*AWT*.

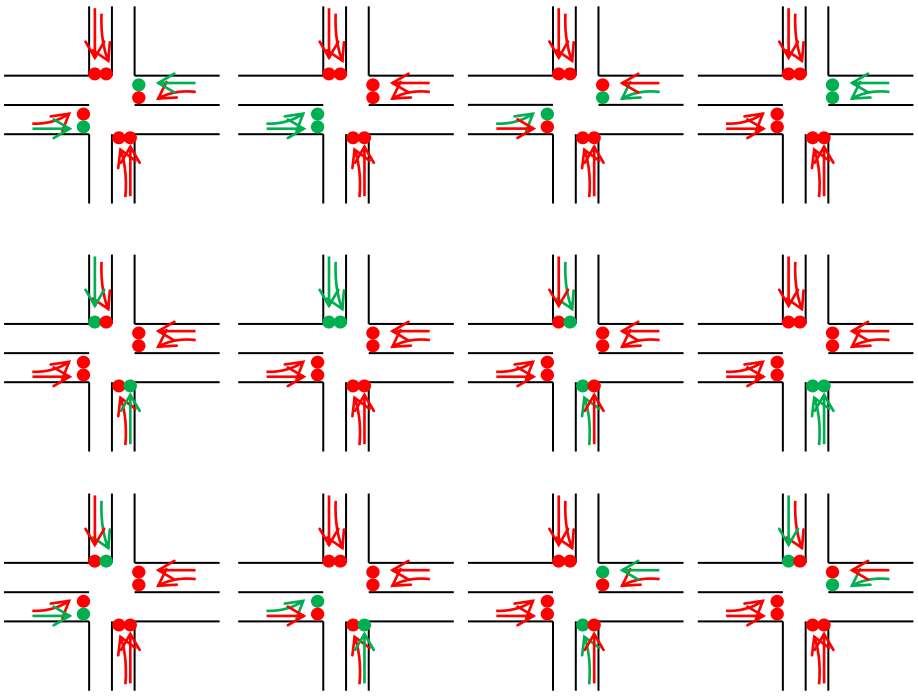


Figure 3.2 Toutes les configurations possibles des phases et feux rouge vert.

Notons que *T* indique la durée du cycle de trafic, *G* symbolise la période de *feu vert* d'une phase en secondes et *R* représente la période de *feu rouge* en secondes. En conséquence, la relation entre *T*, *R* et *G* est donnée par (Équation 3.1).

$$R = T - G \tag{3.1}$$

On suppose que *G* est inférieur au temps maximal prédéterminé nécessaire aux feux verts noté *Tmax*. Le *Tmax* a fait l'objet de plusieurs études comme (Chakraborty, 2014; Johri et al., 2012). De plus, *G* est déterminé en fonction des valeurs de *Q* et du temps nécessaire à un véhicule pour franchir l'intersection (noté *tpass*).

$$G = Q * t_{pass} \tag{3.2}$$

Toutes les informations de trafic sont collectées par le contrôleur distribué sur le côté d'intersection. Ainsi, il est possible de mesurer dynamiquement le nombre de véhicules (longueur) dans toutes les files d'attente ainsi que l'*AWT* à l'intersection. La longueur de la file d'attente *j* pour une direction au $t^{ième}$ cycle est noté $Q_j(t)$. Soit $Rate_{arrival,\ j}(t)$ et $Rate_{departure,\ j}(t)$ représentant respectivement, le taux d'arrivée et le taux de départ pour la file d'attente *j* au $t^{ième}$ cycle. Par conséquent, le taux de croissance de la longueur de la file d'attente pour la file d'attente *j* au $t^{ème}$ cycle qui est noté $a_j(t)$, qui est calculé comme suit :

$$\alpha_j(t) = Rate_{arrival,\,j}(t) - Rate_{departure,\,j}(t) \quad (3.3)$$

$$Q_j(t) = Q_j(t-1) + \alpha_j(t) * G + Rate_{arrival,} * R \quad (3.4)$$

Où : $Q_j(t\text{-}1)$ représente les véhicules restants du cycle précédent.

Pour sélectionner la prochaine phase concernée par le futur *feu vert*, les longueurs de file d'attente : Q_i et Q_j ainsi que le Temps d'Attente (*WT*) : WT_i et WT_j , pour l'ensemble des deux mouvements *i* et *j* sélectionnés parmi les douze cas précédents, sont calculés . Ainsi, la priorité est donnée à la phase dont la valeur la plus élevée est la somme calculée comme suit :

$$S = Q + WT \quad (3.5)$$

Où :

Le temps d'attente de la phase *k* (WT_k) est calculé entre l'heure courante et le dernier temps d'arrêt par la formule suivante :

$$WT_k = T_{current} - T_{stopping} \quad (3.6)$$

3.3 Algorithme de contrôle adaptatif proposé

Nous décrivons dans cette section le principe de fonctionnement de l'algorithme adaptatif proposé. Ce dernier s'exécute, comme le montre clairement la Figure 3.1, sur la station de base (en anglais *BS : Base station*) installée sur le réseau routier. Les données, collectées à partir des différents capteurs sont utilisées pour calculer dynamiquement la longueur probable de la file d'attente pour le prochain cycle ainsi que pour planifier les différents feux de signalisation. Comme spécifié précédemment, l'objectif cible de l'algorithme est d'optimiser *AWT*. L'algorithme considère trois importantes étapes :

i) Détection en temps réel des informations de circulation ;
ii) Sélection de la phase concernée par le prochain cycle de feu vert ;
iii) Détermination de la durée du feu vert.

Par ailleurs, il est utile de préciser qu'il est extrêmement important de tenir compte les véhicules d'urgence lors de la conception d'algorithme adaptatif de contrôle des feux de signalisation car ils peuvent avoir un impact significatif sur la fluidité du trafic. Les ambulances, les camions de pompiers et les voitures de police ont souvent besoin de se déplacer rapidement pour répondre à des situations d'urgence. Lorsque ces véhicules sont bloqués dans la circulation, cela peut entraîner des retards critiques pour les interventions d'urgence et mettre en danger la vie des personnes concernées. En outre, les véhicules d'urgence ont souvent la priorité sur les autres véhicules sur la route.

L'inclusion de ces véhicules dans l'algorithme adaptatif permettrait de réduire le temps d'attente pour les véhicules d'urgence et d'améliorer ainsi leur efficacité dans les situations d'urgence, tout en assurant la sécurité de tous les usagers de la route. Par conséquent, considérer les véhicules d'urgence dans l'algorithme adaptatif de contrôle des feux de signalisation est une étape importante pour assurer une gestion efficace et sûre du trafic. Par conséquent, l'algorithme attribue le *feu vert* d'une durée TG_{emerg} à la voie sur laquelle se présente le véhicule d'urgence. TG_{emerg} est choisi de manière à donner le temps nécessaire aux véhicules d'urgence de traverser l'intersection. Cela dépend du nombre de véhicules qui les précèdent. Dans le cas où plusieurs véhicules d'urgence sont présents sur des voies différentes, la priorité est donc attribuée à celle avec *S* maximum. Le pseudo code de l'algorithme proposé peut être décrit comme suit :

Algorithme 3.1. Algorithme Adaptatif Proposé

Entrée:

- *$Q_j(t)$: Nombre total de véhicules en attente au feu de signalisation sur la $j^{ème}$ voie active au cours d'un $t^{ème}$ cycle;*
- *Voies: Ensemble de voies actives à l'intersection {WFR, WL, EFR, EL, NFR, NL, SFR, SL};*
- *T: Durée du cycle de circulation;*
- *TG_{emerg}: Le temps vert nécessaire pour permettre au véhicule d'urgence de traverser l'intersection.*
- *EmergPresent: Au moins un véhicule d'urgence est présent {vrai, faux};*

Sortie:

- *Séquence de feu vert, durée de phase.*

Debut

/ la fonction de contrôle de présence de véhicules d'urgence revient vraie si au minimum un véhicule d'urgence est présent et la priorité est donnée à la $k^{ème}$ voie pendant la durée TG_{emerg} */*

1. *EmergPresent* = ***Vérification de la présence du véhicule d'urgence*** (*TG_{emerg} , k*);
2. *Si* **(***EmergPresent***)** *Attribuer le feu vert à la $k^{ème}$ voie pour une durée du TGemerg ;*
3. *Sinon {*
4. *Calculer Q pour chaque phase* (*équation* (3.4))
5. *Calculer S pour chaque phase* (*équation* (3.5))
6. *Calculer le temps vert G* (*équation* (3.2))
7. *Si* ($G > T_{max}$) $G = T_{max}$;
8. *Affecter le feu vert suivant à la phase avec S maximum pour une durée G ;*
9. *}*

Fin.

3.4 Résultats de la simulation et discussion

Pour mesurer les performances de notre approche, une implémentation à l'aide du simulateur *Netlogo* est réalisée. *NetLogo* est un langage de programmation et un environnement de modélisation pour le développement de système multi-agents. Les expérimentations ont été menée sur plusieurs scénarios, en comparant les résultats avec les algorithmes : à *temps fixe*, proposé par Rida and Hasbi (2019) qui donne la priorité aux voies de la plus petite file d'attente, et proposé par Yousef et al. (2010) qui attribue la priorité aux voies de la plus grande files d'attente. Les performances sont également mesurées en calculant les valeurs de temps d'attente, de débit de trafic et de nombre de véhicules arrêtés.

Toutes les données des instances sont générées aléatoirement. Le tableau 3.2 décrit certaines données principales de simulation. La zone de simulation est considérée comme environ 30 x 30 patches (*Les patches sont un type spécial d'agents stationnaires dans NetLogo qui composent le monde d'un modèle. On peut les considérer comme les carrés qui composent un échiquier*). Le temps total de simulation est fixé à 2000 ticks (*Un tick est une mesure de temps dans les modèles NetLogo*) ce qui représente 2000 s. La simulation se termine lorsque tous les véhicules ont atteint leur destination, Dans l'expérience de simulation, les douze directions du modèle basé sur l'intersection sont actives. Ainsi, pour imiter les variantes de circulation réelles, lors de l'utilisation des réseaux routiers urbains, la génération dynamique des débits de circulation est variable d'un cycle à l'autre. En outre, un résumé de la variation relative *temps d'attente*, *Débit* et *Nombre de véhicules arrêtés* des trois algorithmes est représenté graphiquement dans les Figure 3.3, Figure 3.4 et Figure 3.5.

Paramètres	**Valeurs**
Zone d'intersection	30 x 30 patches
Unité de temps	Ticks
Temps d'exécution	2000 ticks
Vitesse maximale du véhicule	0.5 patch / ticks
Accélération du véhicule	0.006 patch /ticks
Densité du véhicule	10%−60% par chaque direction
Longueur de voie	24 patches

Tableau 3.2 Données de simulation.

Comme le montre la figure 3.3, on peut percevoir que, sur la courbe d'*AWT*, l'algorithme proposé produit de bonnes résultats en les comparant à l'algorithme à *temps fixe* et l'algorithme de Rida and Hasbi (2019) de manière significative. Cependant, ses performances sont plus-au-moins pareilles à celles de l'algorithme proposé par Yousef et al. (2010). De plus, à propos des valeurs des écarts, nous percevons qu'elles accroissent de manière significative, notamment lorsque les valeurs temporelles varient entre 666s et 1661s (Figure 3.3).

D'autre part, la Figure 3.4 présente les résultats des différents algorithmes, en termes de *débit du trafic*, et qui semblent assez proches. Toutefois, notre algorithme montre une amélioration relative par rapport au débit fourni par l'algorithme fixe. De manière similaire à l'analyse précédente, l'évolution du *nombre de véhicules arrêtés* illustrée sur Figure 3.5 montre que tous les algorithmes, concernés par cette étude, produisent sensiblement les mêmes performances.

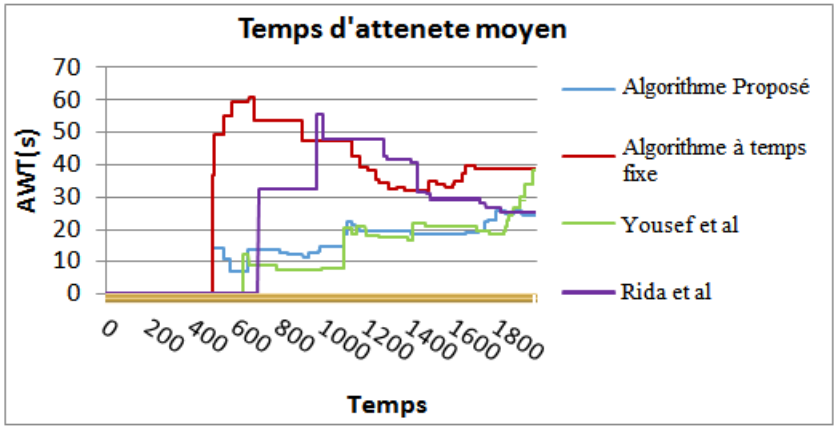


Figure 3.3 Temps d'attente moyen de déférent algorithme de contrôle des feux.

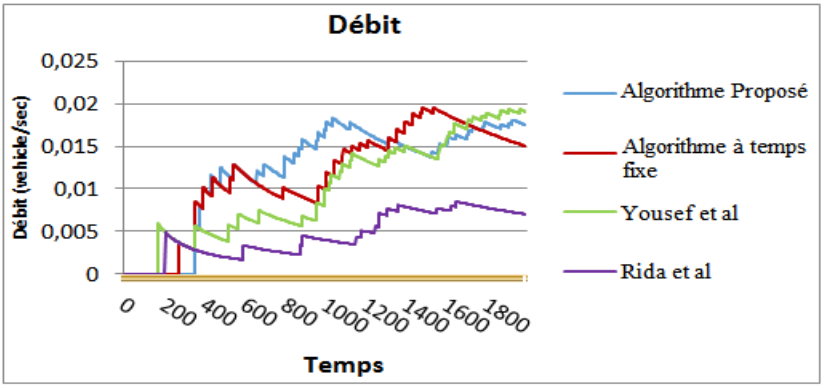


Figure 3.4 Débit de déférent algorithme de contrôle des feux.

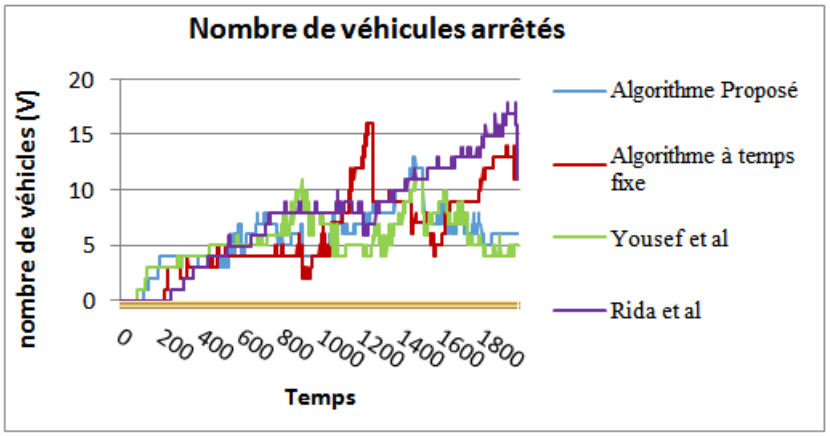


Figure 3.5 Nombre de véhicules arrêtés à une intersection de déférent algorithme de contrôle des feux.

3.5 Conclusion

En conclusion, l'approche réactive pour le contrôle adaptatif des feux de signalisation dans les intersections isolées est une méthode prometteuse pour améliorer la fluidité et la sécurité du trafic routier. Cette approche repose sur l'utilisation de capteurs pour détecter les flux de trafic et ajuster les temps de feux de signalisation en temps réel. Les résultats de cette étude expérimentale, menée pour mesurer les performances de notre approche réactive proposée, en faisant des comparaisons avec d'autres algorithmes de la littérature, ont montré la qualité de cette proposition. À cet égard, la principale contribution de l'algorithme proposé est qu'il peut effectivement sélectionner le meilleur réglage possible pour le séquencement des *feux vert* pour améliorer la qualité du trafic. Néanmoins, des recherches complémentaires sont nécessaires pour évaluer l'applicabilité de cette approche dans des environnements urbains plus complexes et pour évaluer son impact sur la sécurité routière. En somme, l'approche réactive pour le contrôle adaptatif des feux de signalisation dans les intersections isolées représente une avancée majeure dans le domaine du transport intelligent et de la mobilité urbaine durable.

L'approche réactive pour le contrôle adaptatif des feux de signalisation dans les intersections isolées présente certaines limites, notamment en termes de capacité à anticiper les changements futurs de la demande de trafic et à optimiser les temps d'attente des véhicules sur une période plus longue. De plus, cette approche peut être limitée par la difficulté à ajuster rapidement les temps d'attente en réponse à des changements imprévus ou à des situations d'urgence. C'est là qu'interviennent les approches intelligentes à base de *RL*. Ces approches permettent de prendre en compte

un plus grand nombre de facteurs, tels que les conditions météorologiques, les accidents de la route et les événements spéciaux, pour adapter le contrôle des feux de signalisation de manière plus efficace et prédictive. Le *RL* permet à l'algorithme de s'adapter de manière autonome à l'évolution des conditions de trafic, en tirant parti des données en temps réel pour optimiser les temps d'attente et de déplacement des usagers de la route sur une période plus longue.

En somme, si l'approche réactive est efficace dans des situations simples, les approches intelligentes à base de *RL* peuvent permettre une optimisation plus avancée de la gestion du trafic routier, en prenant en compte un plus grand nombre de facteurs et en étant plus prédictives. Ces approches peuvent offrir des résultats plus satisfaisants pour les usagers des réseaux routiers et contribuer à réduire les temps d'attente et les embouteillages dans les intersections isolées et dans les environnements urbains plus complexes.

Dans le chapitre suivant, nous proposerons donc notre deuxième contribution de cette thèse. Il s'agit d'une nouvelle approche adaptative basée sur *DRL* pour gérer le trafic au niveau les intersections isolées.

Chapitre 4 : Approche intelligente basée *DRL* pour le contrôle adaptatif des feux de signalisation dans les intersections isolées

4.1 Introduction

Le problème de la congestion du trafic est un phénomène critique qui affecte la majorité des environnements des villes à travers le monde. L'amélioration des infrastructures a été la principale solution pour faire face à ce problème. Néanmoins, cela n'est souvent pas toujours possible, notamment lorsque les contraintes financières sont trop fortes. En conséquence, diverses approches alternatives ont été envisagées, qui sont souvent centrées sur l'amélioration de l'efficacité des infrastructures existantes en utilisant des systèmes de contrôle des feux de circulation aux intersections. Les solutions de la première génération adoptent des feux tricolores à temps fixe pour contrôler le trafic au niveau des zones de conflit sans considérer l'état actuel du réseau routier (Miller, 1963). Ces méthodes ont plusieurs inconvénients. Tout d'abord, Elles ne sont pas adaptatives aux changements dynamiques du trafic, ce qui peut entraîner des temps d'attente prolongés pour les usagers de la route lorsque le trafic est faible ou des congestions importantes lorsque le trafic est élevé. De plus, les temps de cycle des feux de signalisation doivent être prédéterminés à l'avance, ce qui peut être difficile à optimiser pour des intersections avec des conditions de trafic variables. En outre, elles peuvent entraîner une augmentation de la consommation d'énergie et de la pollution de l'air. Ces inconvénients soulignent la nécessité de développer des méthodes de contrôle des feux de signalisation plus intelligentes et adaptatives (Touhbi et al., 2017) pour améliorer l'efficacité du trafic et réduire les émissions de gaz carbone. D'autre part, le grand progrès qu'a connu le domaine de l'*IA* et de l'*IoT* a joué un rôle primordial dans

la mise en service de nombreux *ITS* permettant d'améliorer la qualité des réseaux de trafic (Haddad et al., 2021).

Par ailleurs, avec le succès des techniques *RL* et *DL* en *IA*, les chercheurs ont montré un intérêt accru pour l'utilisation des techniques *DRL* pour résoudre les problèmes *TSC*. En effet, ces techniques peuvent aider à optimiser la circulation routière en temps réel. En utilisant des données en temps réel sur la circulation, l'agent peut apprendre à ajuster les feux de signalisation de manière à minimiser les temps d'attente et à maximiser le débit de trafic. De plus, l'apprentissage par renforcement peut s'adapter à des situations imprévues, telles que des accidents de la route ou des travaux de construction, en temps réel, ce qui permet d'optimiser la circulation dans ces situations également. C'est ainsi que de nombreuses approches basées sur les *DRL* ont été proposées dans la littérature, prouvant l'intérêt des communautés scientifiques et économiques à profiter des avantages de ces techniques pour améliorer les performances des réseaux de trafic (Genders and Razavi, 2016; Vidali et al., 2019; Wan and Hwang, 2018; Wei et al., 2018; Zeng et al., 2018; Liang et al., 2019).

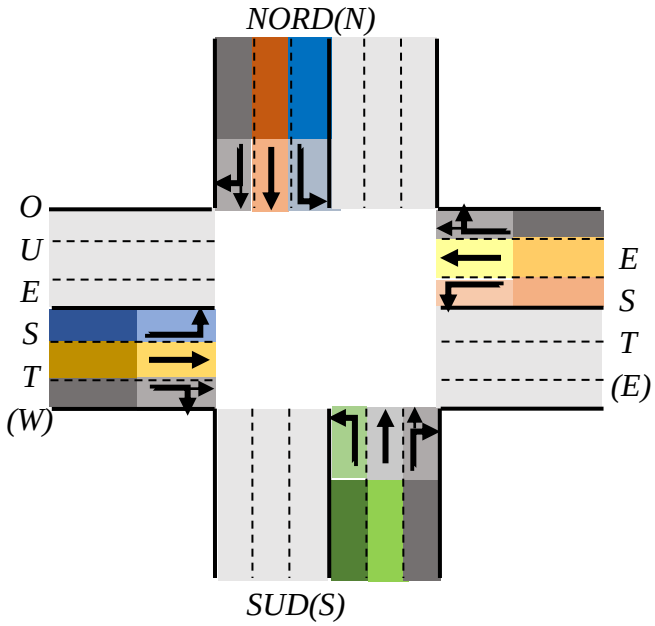
Dans ce chapitre, nous proposons une approche intelligente basée sur le *DRL* pour le contrôle adaptatif des feux de signalisation dans les intersections isolées. Cette approche utilise un agent *DRL* implémentant la méthode *Double Deep Q-Network* (*DDQN*) pour apprendre et optimiser le comportement du contrôleur des feux de signalisation en fonction des conditions de trafic actuelles (Haddad et al., 2022b). Dans cette optique, ce chapitre est structuré en plusieurs sections. Tout d'abord, nous discutons de la problématique des intersections isolées à feux de signalisation et des défis associés au contrôle de ces intersections. Ensuite, nous présentons en détail l'algorithme de contrôle adaptatif proposé. Nous montrons comment l'approche *DRL* peut être utilisée pour apprendre le comportement de contrôle des feux de signalisation en fonction des conditions de trafic actuelles. Enfin, nous présentons les résultats expérimentaux de notre approche et discutons de ses performances.

4.2 Description du problème et objectifs

Notre problématique de recherche de cette approche proposée peut être décrite comme un modèle intégrant un environnement (noté *E*) comprenant une intersection isolé composé de trois voies et un agent intelligent (noté *G*) gérant le flux de trafic dans ce carrefour à l'aide d'un feu tricolore (*Rouge*, *Vert* et *Jaune*). Les lettres *N*, *S*, *W* et *E* désignent respectivement les directions nord, sud, ouest et est. Pour chaque direction, on considère trois voies indiquant tous les sens possibles. Dans chaque voie, deux capteurs (notés *Capteur1* et *Capteur2*) sont placés aux deux extrémités pour compter en

temps réelle nombre de véhicules présents dans la voie. La gestion de l'accès concurrent à la zone de conflit de l'intersection est également assurée par un système de contrôle à feux de signalisation tricolores. Ce dernier est basé sur un système de cycles de feux de signalisation réguliers qui régulent la circulation en fonction des phases de circulation. Le fonctionnement des feux de signalisation à trois couleurs est relativement simple. Les feux sont composés de trois lumières colorées : *rouge*, *jaune* et *vert*. Chaque lumière est allumée en fonction de la phase de circulation en cours à l'intersection. Lorsque la lumière *rouge* est allumée, cela signifie que la circulation doit s'arrêter. Les véhicules doivent attendre jusqu'à ce que la lumière *verte* s'allume, indiquant qu'ils peuvent continuer à avancer en toute sécurité. La lumière *jaune* est généralement utilisée pour indiquer aux conducteurs de ralentir et de se préparer à s'arrêter avant que la lumière *rouge* ne s'allume. Le temps que chaque lumière reste allumée est déterminé par une unité de contrôle centrale, qui peut être programmée pour changer les temps de cycle des feux de signalisation en fonction l'état du trafic à l'intersection.

Il est utile de noter que les phases de circulation font référence aux différentes périodes de temps pendant lesquelles les feux de signalisation à une intersection donnée permettent la circulation de certains types de véhicules ou de piétons. Les phases sont une partie essentielle de la stratégie de contrôle adaptatif des feux de signalisation, car elles permettent aux feux de signalisation de s'adapter aux besoins du trafic en temps réel. Les phases sont généralement définies en fonction des types de mouvements de trafic autorisés, tels que le mouvement direct (avancer tout droit), le mouvement à gauche, le mouvement à droite et les mouvements réservés aux piétons. Par exemple, une phase peut autoriser uniquement le mouvement direct des véhicules allant *d'est* en *ouest* et interdire tout autre mouvement de trafic. Une autre phase peut permettre uniquement le mouvement à gauche pour les véhicules allant du *nord* au *sud* et les *piétons* traversant dans la même direction. Notre approche proposée considère un système de contrôle à huit phases (Figure 4.1). Ces phases sont réparties en deux cycles de quatre phases. Chaque cycle est conçu pour traiter un ensemble spécifique de mouvements de trafic.



Phase	Mouvements
1	
2	
3	
4	
5	
6	
7	
8	

Figure 4.1 Notre modèle basé sur une intersection isolée avec 8 phases.

L'objectif de l'approche proposée est d'optimiser le contrôle des feux de signalisation dans une intersection isolée en utilisant une méthode basée sur le *DRL*. Plus spécifiquement, cette approche vise à utiliser les données de trafic en temps réel pour entraîner un agent à prendre des décisions de contrôle des feux de signalisation de manière à minimiser les temps d'attente et à maximiser le débit de trafic dans une intersection isolée. En résumé, l'objectif est d'améliorer la fluidité de la circulation, de réduire les temps d'attente, les émissions de gaz et d'augmenter la sécurité routière dans cette intersection isolée.

4.3 Formulation du problème

Le problème de contrôle adaptatif des feux de signalisation peut être formulé comme un problème de prise de décision en temps réel, où l'objectif est de maximiser la fluidité du trafic tout en minimisant les temps d'attente pour les usagers de la route. Pour résoudre ce problème, un agent d'apprentissage par renforcement est entraîné à partir d'interactions avec l'environnement, c'est-à-dire les mouvements des véhicules et des piétons, les fluctuations de volume de trafic, etc. L'agent prend des décisions en temps réel pour ajuster les temps de phase des feux de signalisation afin de minimiser les temps d'attente et d'optimiser la fluidité du trafic.

La formulation du problème de contrôle adaptatif des feux de signalisation tricolores en utilisant les techniques *DRL* implique la définition d'un espace d'états, d'un espace d'actions et d'une fonction de récompense appropriés. L'espace d'états représente l'état

actuel du trafic à l'intersection, tel que le nombre de véhicules et de piétons qui traversent l'intersection, les temps d'attente actuels, etc. L'espace d'actions représente les actions possibles que l'agent peut prendre, telles que l'ajustement des temps de phase des feux de signalisation. La fonction de récompense est utilisée pour guider l'agent vers une politique de contrôle optimale en lui fournissant une rétroaction positive ou négative en fonction des résultats de ses actions. En effet, notre problématique dans ce contexte pourra être formulée comme suit :

- *Espace d'états (S)* : Il s'agit de l'ensemble des variables qui décrivent l'état actuel du système. Dans le cas des feux de signalisation tricolores, cela pourrait inclure des variables telles que le nombre de véhicules dans les files d'attente et de piétons qui traversent l'intersection, la durée des temps d'attente actuels, etc. L'espace d'états peut être formalisé mathématiquement comme une liste de variables d'état s = [s1, s2, ..., sn], où n est le nombre de variables d'état. Les actions ici attribuent des droits de passage pour chaque flux de trafic spécifié. Ainsi, suite à toute action l'état du trafic possède l'état s_t à l'état s_{t+1}.
- *Espace d'actions (A)* : Il s'agit de l'ensemble des actions possibles que l'agent peut prendre à chaque étape de décision. Dans le cas des feux de signalisation tricolores, cela pourrait inclure des actions telles que l'ajustement des temps de phase des feux de signalisation, l'activation ou la désactivation de feux de signalisation, etc. L'espace d'actions peut être formalisé mathématiquement comme une liste d'actions a = [a1, a2, ..., am], où m est le nombre d'actions possibles.
- *Fonction de transition (T)* : Elle décrit comment l'état du système évolue lorsque l'agent prend une action donnée. Elle peut être formalisée mathématiquement comme *T(s, a, s')*, où *s'* est l'état résultant de l'application de l'action *a* à l'état *s*.
- *Fonction de récompense (R)* : Elle fournit une mesure quantitative de l'efficacité d'une action donnée par l'agent. Dans le cas des feux de signalisation tricolores, cela pourrait inclure des récompenses positives pour la réduction des temps d'attente et des récompenses négatives pour la congestion du trafic. La fonction de récompense peut être formalisée mathématiquement comme *R(s, a)*.
- *Politique de contrôle (π)* : Elle décrit comment l'agent choisit une action donnée à partir de l'état actuel du système. La politique peut être formalisée mathématiquement comme $\pi(s) = a$, où *a* est l'action choisie par l'agent à l'état *s*.

L'objectif du contrôleur adaptatif des feux de signalisation est de trouver une politique de contrôle optimale (notée π^*), qui maximise la récompense cumulative au fil du temps. L'agent contrôleur de l'intersection isolée interagit avec son environnement à des pas de temps discrets, tout en essayant d'optimiser certaines métriques à savoir : Temps d'attente moyen (en anglais *AWT : Average Waiting Time*), Longueur de file d'attente moyenne (en anglais *AQL : Average Queue Length*), Consommation de

carburant moyenne (en anglais *AFC : Average Fuel Consumption*) et Émissions moyennes de CO_2 (en anglais *AEC : Average Emission* CO_2). A chaque pas de temps (noté *t*), l'agent observe l'état de l'environnement s_t puis sélectionne une action a_t.

Au début de l'entraînement, l'agent ne connaît pas les bonnes actions à prendre dans chaque situation. Il explore donc l'environnement en choisissant des actions au hasard et en observant les résultats. Ensuite, il utilise les expériences passées pour mettre à jour sa politique, afin de sélectionner des actions plus adaptées aux situations. Au fil du temps, l'agent apprend à choisir les actions qui maximisent la récompense R_t. R_t est défini comme :

$$R_t = \sum_{k=0}^{\infty} \gamma^k \, r_{t+k} \tag{4.1}$$

Où γ représente un facteur d'actualisation permettant de contrôler l'importance des récompenses immédiates par rapport aux récompenses futures.

Notre modèle considère une fonction de récompense définie comme une variation du temps d'attente cumulé des véhicules (noté *W*) et de l'émission de CO_2 (noté *Ec*) entre deux étapes successives. L'équation 4.2 décrit la formule de cette fonction :

$$Reward = W_t - W_{t+1} + Ec_t - Ec_{t+1} \tag{4.2}$$

Où W_t, W_{t+1} et Ec_t, Ec_{t+1} représentent le temps d'attente total cumulé et l'émission CO_2 de toutes les voitures à l'intersection. Ces paramètres peuvent être définis comme :

$$W_t = \sum_{v \in V_t} W_t^v \tag{4.3}$$

$$Ec_t = \sum_{v \in V_t} Ec_t^v \tag{4.4}$$

Où $\boldsymbol{V_t}$ est l'ensemble des véhicules sur les voies d'approche dans la simulation à l'instant $_t$ et $\boldsymbol{w_t^v}$et $\boldsymbol{Ec_t^v}$est le temps d'attente et les émissions de CO_2 du véhicule v à l'instant t respectivement.

4.4 Approche proposée

Rappelons que l'approche que nous proposons ici s'appuie sur l'algorithme *DDQN* qui combine les réseaux de neurones profonds et l'algorithme *Q-Learning*. Par conséquent, le *DDQN* utilise deux réseaux de neurones pour estimer la fonction de valeur Q : le réseau "*online*" et le réseau "*target*". La mise à jour des poids des deux réseaux de neurones suit l'équation (2.7).

Plusieurs éléments essentiels sont à considérer pour mettre en œuvre notre approche à savoir (Figure 4.2) tels que : L'*environnement* qui représente l'intersection routière pour laquelle nous essayons de concevoir un système de contrôle adaptatif de feux de signalisation. Il est composé de plusieurs éléments tels que les véhicules, les piétons, les feux de signalisation, les voies, etc. L'état de l'*environnement* est observé par l'agent en temps réel. L'état peut inclure des informations sur la *densité* du trafic, l'état des différentes files d'attente, la durée écoulée depuis la dernière fois que les feux de signalisation ont été modifiés, la durée écoulée depuis que les véhicules ont été autorisés à passer, etc. Quant à l'*agent* décideur, il est considéré comme étant le cerveau de l'ensemble de notre système de contrôle adaptatif de feux de signalisation. Il est implémente l'algorithme *DRL* pour apprendre des actions et prendre les meilleures décisions en fonction de l'état actuel de l'*environnement*.

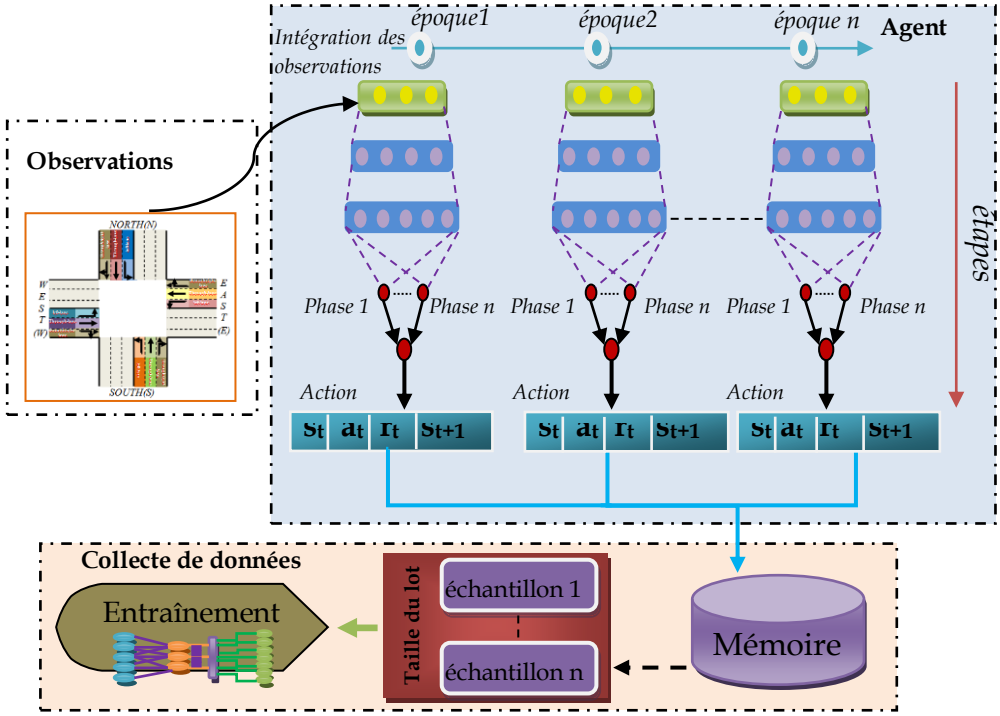


Figure 4.2 Structure du modèle proposé.

4.4.1 Stratégies d'exploration et d'exploitation

A chaque pas de contrôle, l'agent observe l'état du trafic à l'intersection et sélectionne une action à exécuter. Il n'est pas explicitement informé des actions à prendre, mais doit plutôt découvrir quelle action rapporte le plus de récompenses par essais et erreurs. Pour ce faire, l'agent utilise des stratégies d'*exploration* et d'*exploitation*. Pour équilibrer l'*exploration* et l'*exploitation*, notre modèle adopte l'approche *ε-greedy* (Watkins, 1989). Cette dernière permet à l'agent d'explorer de nouvelles actions avec une certaine

probabilité tout en exploitant les actions les plus prometteuses la plupart du temps. Elle consiste à prendre une action aléatoire avec une probabilité ε et à prendre la meilleure action connue avec une probabilité *1-ε*. La valeur de ε étant définie comme suit :

$$\varepsilon = 1 - \frac{epoch\_current}{epochs} \tag{4.5}$$

Où *epoch_current est l'épisode actuel et epochs est le nombre total d'épisodes.*

Il est utile de préciser que le choix de la valeur ε est un compromis important dans la stratégie *ε-greedy*. Si ε est trop élevé, l'agent effectuera trop d'actions aléatoires et ne pourra pas exploiter pleinement les connaissances acquises. En revanche, si ε est trop faible, l'agent sera moins enclin à explorer de nouvelles actions, ce qui pourrait le conduire à une stratégie sous-optimale. Le choix des valeurs de ε dépend du problème spécifique et de la marge de tolérance pour l'exploration (Figure 4.3). Dans certaines situations, une valeur faible de ε peut être suffisante, tandis que dans d'autres cas, une valeur plus élevée est nécessaire pour explorer efficacement l'espace des états et des actions.

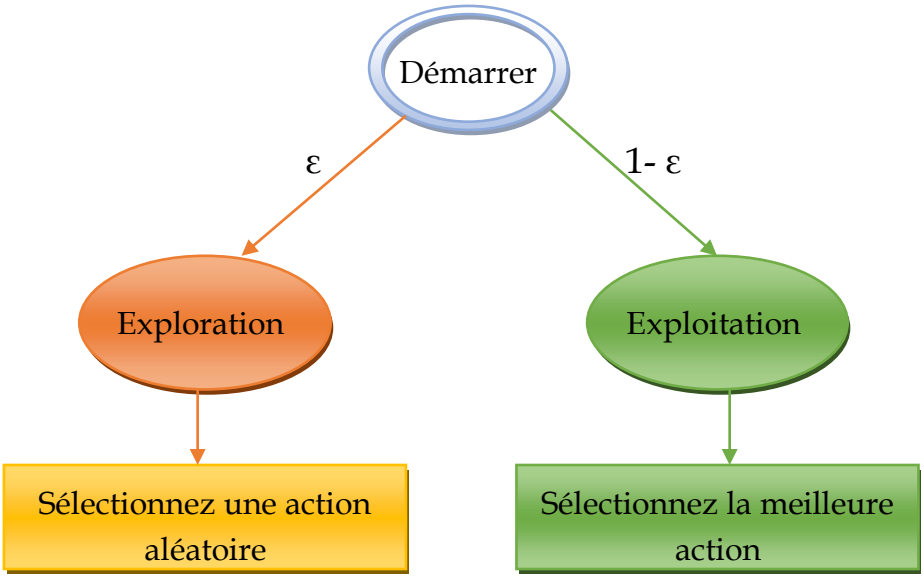


Figure 4.3 ε-greedy : équilibrage entre l'exploration et l'exploitation.

Dans notre approche, nous avons décidé d'adapter dynamiquement la valeur de ε au cours de l'apprentissage, de manière à diminuer linéairement ε au fil du temps. En effet, les valeurs de *début* et de *fin* et le nombre d'étapes pour atteindre la fin sont prédéfinis. En d'autres termes, la valeur de ε est initialement fixée à un niveau élevé, ce qui signifie que l'agent est plus enclin à choisir une action aléatoire plutôt que la meilleure action connue. Au fil du temps, la valeur de ε est réduite de manière linéaire jusqu'à atteindre

une valeur minimale prédéfinie. Cette réduction progressive de ε permet à l'agent d'explorer de nouvelles actions au début de l'apprentissage tout en apprenant à exploiter les actions les plus prometteuses au fil du temps. L'Algorithme 4.1 décrit le pseudo code de la stratégie *ε-greedy*.

Algorithme 4.1. ε-greedy

Donnés : Q : la valeur Q généré jusqu'à présent, ε : un petit nombre,
S : État actuel.
Résultat : Action sélectionnée
Fonction Sélectionnez-action (Q, S, ε)
n ← nombre aléatoire uniforme compris entre 0 et 1 ;
Si n < ε **alors**
A ← Sélectionnez une action aléatoire
Sinon
A ← Sélectionnez la meilleure action
Finsi
Retourner l'action sélectionnée A
Fin

Dans cette optique, plusieurs composantes indispensables au processus d'apprentissage doivent être définies, telles que l'*état*, l'*action* et la *récompense*. Pour définir l'état (s_t) à l'étape *t*, toutes les voies sont discrétisées en cellules, chaque cellule est représentée par une variable ayant une valeur de 1 ou 0, représentant la présence ou l'absence de véhicule, respectivement, ainsi que le niveau de congestion pour refléter le nombre de véhicules en attente. Les indices des voies les plus encombrées sont également utilisés pour que chaque direction choisisse une voie moins encombrée, ce qui est utilisé comme état pour contrôler le temps de phase. Enfin, en tant qu'action, à la fin de chaque étape, l'agent sélectionne la phase suivante à appliquer à l'intersection.

La récompense constitue un élément clé dans notre approche. Elle représente également la mesure numérique qui indique l'efficacité de la stratégie de contrôle des feux adoptée par l'agent. Elle est également définie en fonction de quatre mesures à savoir : *AWT*, *AQL*, *AFC* et *AEC*.

4.4.2 Architecture des réseaux de neurones

Comme mentionné dans la section précédente, l'approche proposée se base sur modèle utilisant deux réseaux de neurones : un réseau "*online*" et un réseau cible appelé

"*target*". Le réseau "*online*" est le réseau qui prend les décisions dans l'environnement, tandis que le réseau "*target*" est un réseau utilisé pour évaluer la qualité de la politique de décision prise par le réseau "*online*". Ces deux réseaux ont une structure similaire avec trois couches cachées, chacune contenant trois cent neurones, et une couche de sortie contenant huit neurones correspondants aux huit actions possibles (Figure 4.4). Chaque neurone est modifié par la fonction d'activation baptisée *Unité Linéaire Rectifiée* (en anglais *ReLU : Rectified Linear Unit*). Les entrées du réseau de neurones comprennent les états de l'environnement, tandis que les sorties du réseau de neurones sont les valeurs de la fonction d'action (*Q-value*) pour chaque action possible dans l'état actuel de l'environnement.

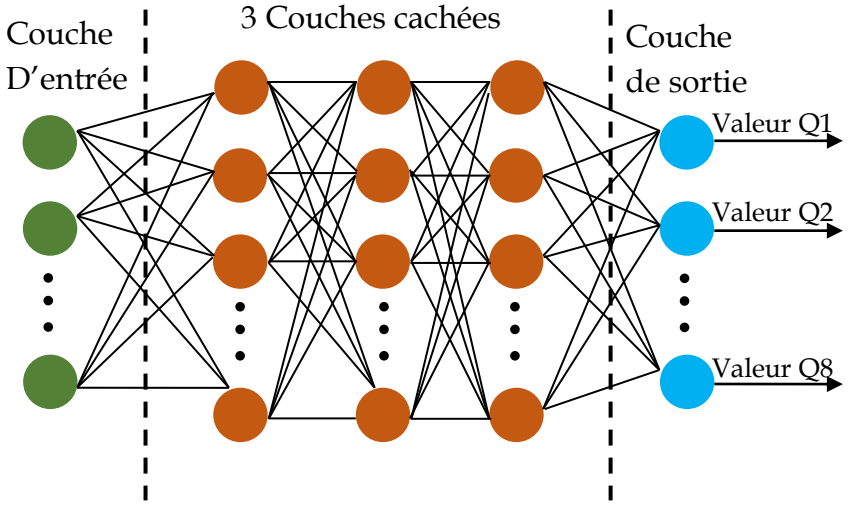


Figure 4.4 Architecture des réseaux de neurones online et target.

En outre, une technique appelée "*replay buffer*" est utilisée pour stocker les données d'apprentissage. Les données stockées dans le "*replay buffer*" (notée *M*) sont utilisées pour mettre à jour les deux réseaux de neurones, ce qui permet d'améliorer l'apprentissage et de stabiliser les mises à jour des réseaux.

4.4.3 Entrainement du modèle

En ce moment, il est temps de procéder à l'entraînement du modèle. Cette étape consiste à exécuter de nombreuses simulations, en utilisant différentes configurations de contrôle des feux de signalisation, afin de collecter des données d'entraînement. Ces données comprennent des états, des actions, des récompenses et des états suivants, qui sont utilisés pour entraîner notre modèle de contrôle de signalisation.

Pendant l'entraînement, le modèle est alimenté en données d'entrée qui représentent l'état courant de l'intersection et le contrôleur de signalisation choisit une action en

fonction de ces données d'entrée. La récompense associée à cette action est ensuite calculée en fonction de l'impact de l'action sur les performances du trafic. Cette récompense est utilisée pour ajuster les poids et les biais du modèle, de manière à ce que le contrôleur de signalisation apprenne à prendre des décisions qui maximisent les performances du trafic. Enfin, l'entraînement du modèle se poursuit jusqu'à ce que le contrôleur atteigne un niveau de performance satisfaisant.

L'algorithme 4.2 décrit le processus d'entraînement pour notre approche proposée pour contrôler les feux de signalisation d'une intersection isolée. Il prend en entrée plusieurs paramètres à savoir : *taille du lot* (notée $|B|$), *facteur d'évaluation* $\gamma \in [0,1]$ et *capacité de la mémoire* (notée M). L'objectif de l'algorithme est de trouver les valeurs Q optimales avec des poids entraînés.

Algorithme 4.2. Processus d'entrainement

Entrée : $|B|$, γ, M
Sortie : *Valeurs Q optimales avec des poids entraînés*
Initialisation
  Initialiser aléatoirement tous les paramètres entraînables θ
  Initialiser la mémoire m à la capacité M, taille de lot B
  Initialiser les valeurs Q de manière aléatoire pour l'agent
  Initialiser l'état initial s_t
Début
  Répéter *(pour chaque épisode) :*
    Répéter *(pour chaque étape d'un épisode):*
      Observer l'état actuel s_t
      Choisissez l'action a et exécutez-la par ε*-greedy en utilisant Algorithme* 4.1
      Effecter l'action a, observer r, s_{t+1} *dans équation* (4.2)
        si $|B| > M$
          Supprimer le plus ancien tuple t de la mémoire m
        fin
      Stocker le tuple $t = (s_t, a_t, r_t, s_{t+1})$ *dans la mémoire m*
      Expériences aléatoires de taille B par lot
      Mettre à jour le réseau θ *en utilisant équation* (2.7)
      Mettre $\theta^* = \theta$
      $s \leftarrow s_{t+1}$
    Jusqu'à *s est terminal*
  Jusqu'à *Le dernier épisode se termine;*
Fin

L'algorithme commence par une phase d'initialisation où tous les paramètres d'entrainement comme θ, M, S_o (*état initial*), etc. sont initialisés. Ensuite, à chaque étape de l'épisode, l'algorithme observe l'état actuel s_t et choisit l'action a_t en utilisant la

stratégie *ε-greedy.* L'action est ensuite exécutée, et la récompense r_t et l'état suivant s_{t+1}sont calculés. Par ailleurs, lorsque la mémoire *M* atteint sa capacité, le tuple le plus ancien est supprimé pour libérer de l'espace afin de stocker le nouveau tuple (noté $t = (s_t, a_t, r_t, s_{t+1})$). Après, le réseau θ est mis à jour en utilisant l'équation (2.7) qui correspond à la fonction de perte du réseau de neurones qui approxime la fonction de valeur *Q*.

Les paramètres du réseau θ sont copiés dans θ^* pour la mise à jour des poids cibles et l'état actuel s_t est mis à jour avec s_{t+1}. Enfin, l'algorithme continue la boucle principale jusqu'à ce que le dernier épisode se termine.

4.5 Résultats expérimentaux et discussions

Dans cette section, nous présentons les résultats de l'étude expérimentale que nous avons menée pour évaluer les performances de l'approche proposée afin de contrôler les feux de signalisation d'une intersection isolée. Nous avons mené des expérimentations en utilisant une configuration précise. Tout d'abord, nous avons utilisé la plateforme *SUMO* pour simuler et modéliser le comportement du trafic dans notre environnement d'intersection. Ensuite, nous avons mis en œuvre notre approche en utilisant *TensorFlow*, une bibliothèque d'apprentissage automatique renommée, qui nous a permis de construire et d'entraîner des réseaux de neurones profonds. Pour optimiser les réseaux de neurones, nous avons utilisé l'optimiseur[3] *Adam*[4] *(**Ada**ptive **M**oment Estimation)* (Kingma and Ba, 2014). Nous avons fixé le taux d'apprentissage à *0,0001* pour favoriser une convergence stable. L'agent *DRL* utilise une mémoire de taille *30 000* échantillons pour stocker les expériences passées et former un ensemble d'apprentissage. Lors de l'apprentissage, nous avons utilisé des échantillons de taille *B* égale à *50* et un facteur d'actualisation γ de *0,99* pour tenir compte des récompenses futures. Cette configuration a été soigneusement sélectionnée afin d'optimiser les performances de ladite approche. Les simulations ont été exécutées sur un ordinateur personnel (*PC : Personal Computer*) avec un processeur Intel *(i5-2310, @ 2.5GHz 2.5GHz*), processeur graphique (*GPU*) *Radeon*, une *RAM* de taille *6GB et running Mint 19.3*.

[3]Les ***optimiseurs*** sont des algorithmes ou des méthodes qui sont utilisés pour modifier ou ajuster les attributs d'un réseau de neurones tels que les poids des couches, le taux d'apprentissage, etc. afin de réduire la perte et d'améliorer à son tour le modèle.

[4]***Adam*** est un optimiseur largement utilisé dans l'apprentissage automatique pour l'entraînement des modèles de réseaux neuronaux. Il combine efficacement l'algorithme du gradient stochastique avec des mécanismes d'adaptation du taux d'apprentissage.

Le tableau 4.1 résume les éléments essentiels de la configuration adoptée par notre étude expérimentale.

Paramètres	Descriptions	Valeurs
Taille mémoire	Taille de la mémoire	30000
Taille du lot	Taille de l'échantillonnage par lots	50
Épisode	Taille du pas d'entraînement	200
Temps de simulation	Durée de la simulation	3000s
Optimiseur	Algorithme d'optimisation	Adam
Taux d'apprentissage α	Taille de pas dans la fonction de perte	0.0001
facteur d'actualisation γ	Poids qui multiplie la récompense future	0.99

Tableau 4.1 Hyper-paramètres de l'agent.

La génération de trafic dans une intersection simulée est considérée comme une étape critique car elle peut avoir un impact significatif sur les performances globales. Pour créer une simulation réaliste du trafic dans une intersection isolée, nous avons tenu compte plusieurs facteurs importants. Tout d'abord, la modélisation des véhicules dont nous avons considéré différents types de véhicules (voitures, camions, motos, etc.) avec des caractéristiques spécifiques telles que la vitesse maximale, l'accélération, etc. Ensuite, la répartition du trafic en fonction de l'heure de la journée. Par exemple, le trafic peut être plus dense pendant les heures de pointe du matin et du soir. D'autre part, la définition des itinéraires des véhicules semble extrêmement utile dans toute simulation de trafic réaliste. Nous avons défini un modèle de circulation qui permet entre autre de déterminer les itinéraires probables des véhicules en fonction de leur point de départ et de leur destination. Par conséquent, pour chaque véhicule, la source et la destination sont définies à l'aide d'un générateur de nombres aléatoires dont la graine est modifiée pour chaque nouvel épisode (Figure 4.5). Dans cette optique, pour maintenir un haut niveau de réalité, notre générateur de trafic suit une distribution arbitraire (Vidali et al., 2019).

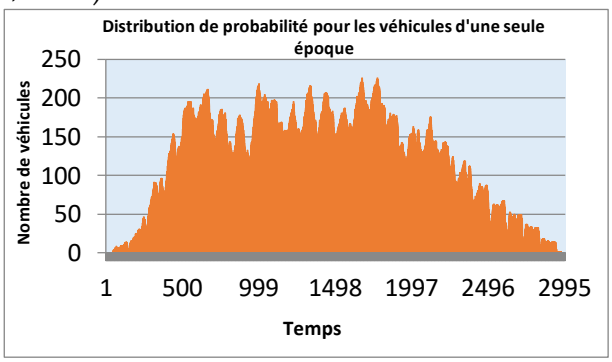


Figure 4.5 Distribution de génération de signaux de trafic sur une époque.

Le code XML suivant décrit notre modèle de circulation adopté pour mener les expérimentations sous SUMO :

Algorithme 4.3. Configuration xml utilisée dans SUMO

```
<routes>
        <vType accel="1.0" decel="4.5" id="car" length="5.0" minGap="2.5" maxSpeed="25"
        sigma="0.5" />

        <route id="WN" edges="West North"/>
        <route id="WE" edges=" West East "/>
        <route id="WS" edges=" West South "/>
        <route id="NW" edges=" North West "/>
        <route id="NE" edges=" North East"/>
        <route id="NS" edges=" North South "/>
        <route id="EW" edges=" East West "/>
        <route id="EN" edges=" East North "/>
        <route id="ES" edges=" East South "/>
        <route id="SW" edges=" South West "/>
        <route id="SN" edges=" South North "/>
        <route id="SE" edges=" South East "/>

  <vehicle id="veh0" type="car" route="WS" departLane="random" departSpeed="10"/>
  <vehicle id="veh1" type="car" route="EW" departLane="random" departSpeed="10"/>
  <vehicle id="veh2" type="car" route="SE" departLane="random" departSpeed="10"/>
  .
  .
  .
  <vehicle id="veh999" type="car" route=" ES" departLane="random" departSpeed="10"/>
</routes>
```

Nous avons évalué les performances de notre approche en termes de plusieurs métriques clés à savoir : *AWT, AQL, AFC* et *AEC.* Nous avons comparé également nos résultats avec ceux obtenus par des méthodes traditionnelles de contrôle des feux de signalisation comme : *DQTSCA* (*Deep Q-network Traffic Signal Control Agent*) (Genders and Razavi, 2016), l'algorithme de contrôle à *temps fixe* et *TSTMA* (*Traffic Signals Time Manipulation Algorithm*) (Yousef et al., 2010).

Tous ces algorithmes ont été évalués dans les mêmes conditions. Plusieurs expériences ont été lancées. Ainsi, il a été remarqué ce qui suit : alors que la demande de trafic augmente, les quatre mesures de l'algorithme de contrôle à temps fixe connaissent une croissance exponentielle. Cela s'explique par le fait que l'algorithme de contrôle à temps fixe ne tient pas compte des variations en temps réel de la demande de trafic, ce qui

limite son adaptation. En revanche, les Figure 4.6, Figure 4.7 et Figure 4.8 illustrent que l'approche proposée démontre les meilleures performances selon toutes les métriques évaluées, en plus elle semble stable tout au long de la période de simulation (Figure 4.5). Des discussions à propos des résultats obtenus sont bien présentées dans la section suivante.

4.5.1 Discussions

Dans cette section, nous présentons les résultats obtenus des expérimentations utilisant l'approche proposée basée sur les métriques de performance mentionnées précédemment tels que : *AWT*, *AFC*, *AEC* et *AQL*.

La figure 4.6 montre les performances d'entraînement de l'agent concernant les métriques de trafic pendant le processus de formation. Au début du processus de formation, il est raisonnable de s'attendre à de mauvaises performances lorsque l'agent explore et exécute des actions pour acquérir de l'expérience. Cependant, à mesure que le nombre d'épisodes d'entraînement augmente, les métriques diminuent rapidement et finissent par converger vers de petites valeurs. Ceci suggère que l'agent a réussi à apprendre une politique d'action efficace grâce à cet entraînement. De plus, il est observable que ces métriques se maintiennent à des petites valeurs stables après 100 épisodes.

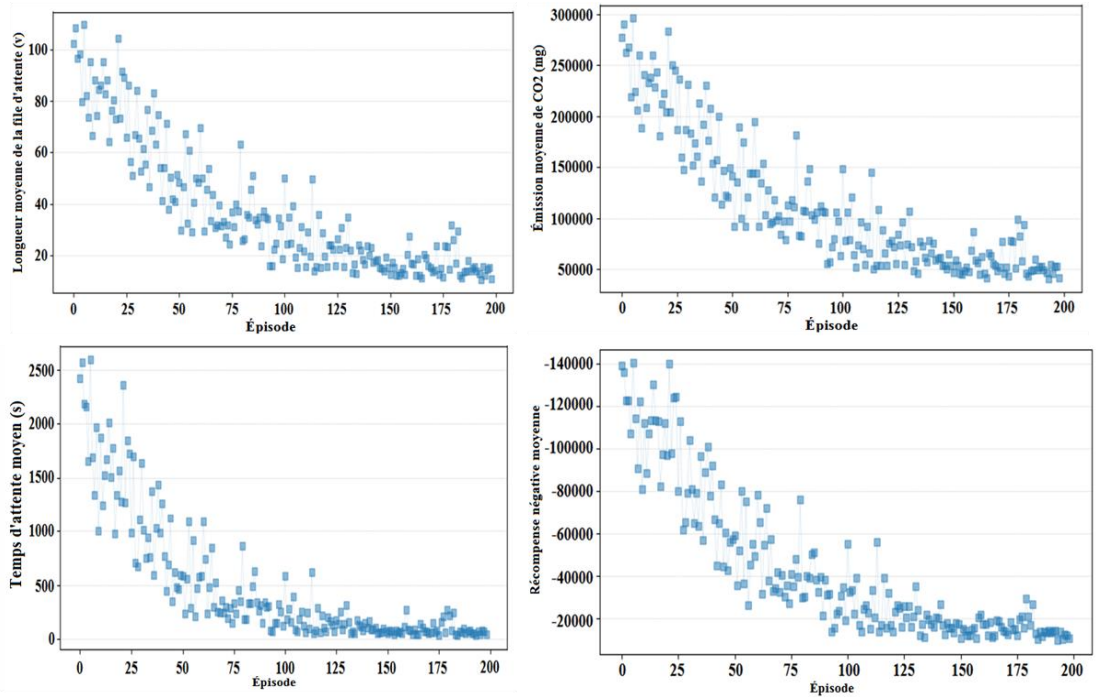


Figure 4.6 Les résultats de l'entraînement en termes de récompense cumulée, AWT, AEC et AQL.

Les résultats comparatifs des performances de quatre métriques sont présentés dans la figure 4.7, en utilisant les méthodes traditionnelles *DQTSCA*, l'algorithme de contrôle à temps fixe et *TSTMA*. Les résultats suggèrent que l'approche à temps fixe est moins performante que les autres méthodes. *DQTSCA* présente des performances similaires à celles de *TSTMA*, tout en nécessitant un coût de d'entrainement inférieur (de 0 à 1000s). Parmi ces méthodes, l'approche proposée a démontré les meilleures performances pour toutes les métriques, et elle semble également plus stable tout au long de la simulation.

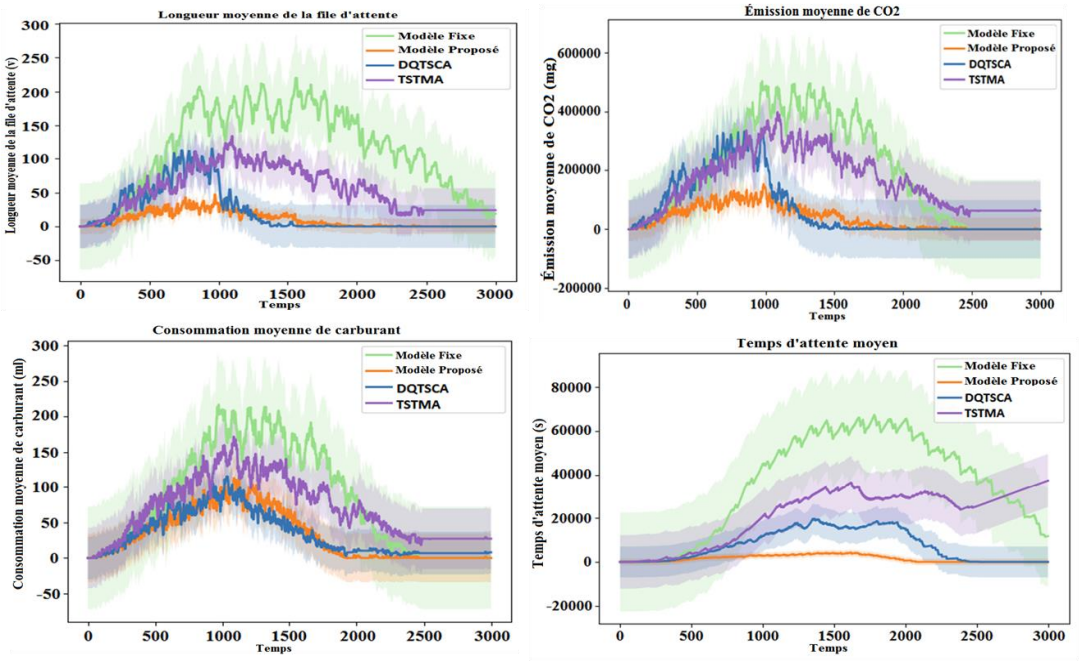


Figure 4.7 Graphiques de comparaisons de toutes les métriques adoptées.

Afin de tirer les pertinentes conclusions à propos de l'approche proposée, nous présentons les résultats de nos expériences sous forme de valeurs moyennes, calculées à partir de plusieurs types de flux de trafic à savoir : *500, 750, 1000* et *1500 véhicules/heure*. Ces résultats sont illustrés dans la Figure 4.8. Toutefois, l'approche proposée et *DQTSCA* ont été entrainées dans un système de circulation avec un débit de *1000 véhicules/heure.*

En conséquence, l'approche proposée est performante pour tous les paramètres du système de circulation, à l'exception du flux de trafic de *1500 véhicules/heure*. Il n'y a pas de différences significatives dans toutes les mesures entre les méthodes *Fixe* et *DQTSCA* en raison de la capacité acceptable de l'intersection à des taux de production de véhicules faibles. Cependant, lorsque le nombre de véhicules augmente avec une augmentation de la production de trafic (*1000 véhicules/heure*), notre approche surpasse les autres algorithmes de contrôle dans toutes les métriques. Dans la plupart des cas, en

particulier lors de fortes périodes de trafic (*1500 véhicules/heure*), notre approche de contrôle offre une fiabilité et une efficacité supérieures.

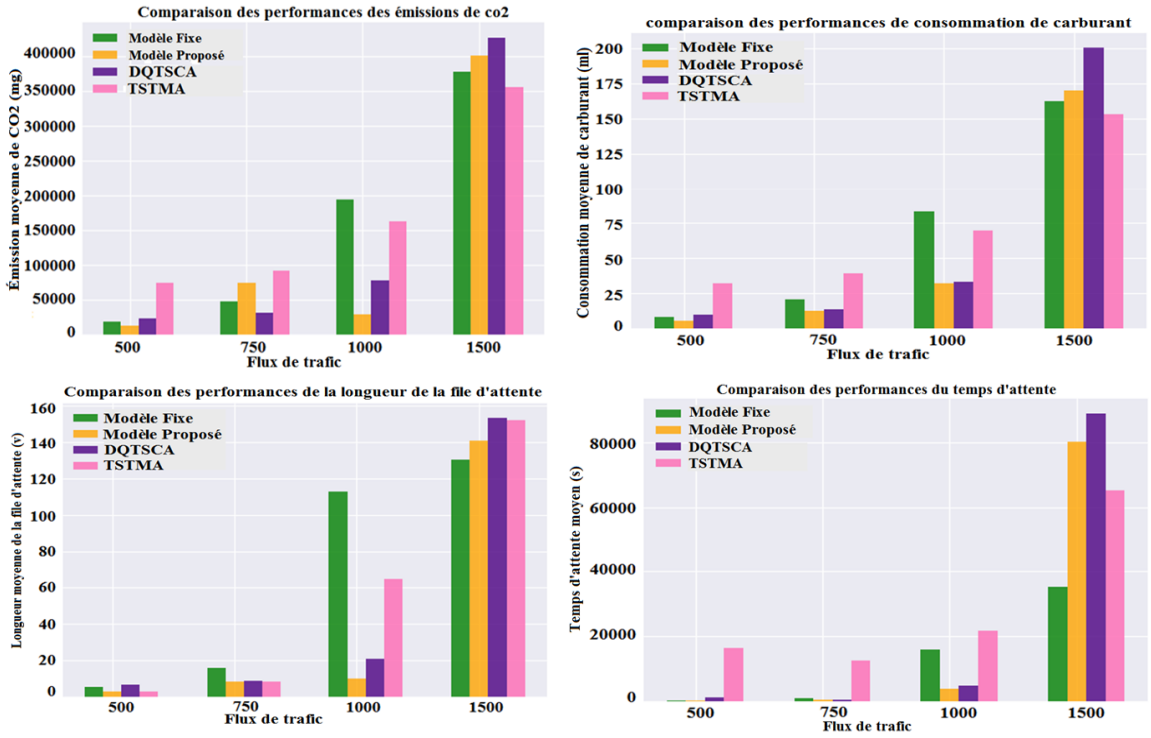


Figure 4.8 Comparaison des performances en termes de moyennes de toutes les métriques adoptées pour différents flux de trafic.

Le tableau 4.2 résume les résultats obtenus pour différentes métriques (*AWT, AQL, AFC* et *AEC*) en comparant les taux d'améliorations par rapport aux trois approches précédentes (*Fixe, DQTSCA* et *TSTMA)*. Pour l'*AWT*, notre approche propose une réduction de *30.01%* par rapport à l'approche *Fixe*, une réduction de *17.32%* par rapport à *DQTSCA* et une réduction de *40.08%* par rapport à l'approche *TSTMA*. Cela signifie que notre approche permet de réduire significativement le temps d'attente moyen par rapport aux deux autres approches. En ce qui concerne l'*AQL*, elle produit une réduction de 35.32% par rapport à l'approche *Fixe,* une réduction de *16.02%* par rapport à *DQTSCA* et une réduction de *20.54%* par rapport à l'approche *TSTMA*. Cela indique que notre approche parvient à réduire la longueur moyenne de la file d'attente de manière plus importante que les autres approches. Quant à l'*AFC*, l'approche proposée montre une réduction de *41.57%* par rapport à l'approche *Fixe,* une réduction de *13.84%* par rapport à *DQTSCA* et une réduction de *45.88%* par rapport à l'approche *TSTMA*. Ces résultats indiquent que notre approche permet de réduire de manière significative la consommation moyenne de carburant par rapport aux autres approches. Enfin, pour l'*AEC*, elle présente une réduction de *16.15%* par rapport à l'approche *Fixe,* une

réduction de *12.65%* par rapport à *DQTSCA* et une réduction de *39.79%* par rapport à l'approche *TSTMA*. Ces résultats mettent en évidence la capacité de notre approche à réduire les émissions moyennes de CO_2 de manière plus importante que les trois autres approches.

Métriques	Fixe	DQTSCA	TSTMA
AWT	30.01%	17.32%	40.08%
AQL	35.32%	16.02%	20.54%
AFC	41.57%	13.84%	45.88%
AEC	16.15%	12.65%	39.79%

Tableau 4.2 Le pourcentage réduit de notre approche par rapport aux autres approches.

En résumé, les résultats obtenus démontrent que notre approche se distingue dans toutes les métriques étudiées, en offrant des réductions significatives du temps d'attente moyen, de la longueur moyenne de la file d'attente, de la consommation moyenne de carburant et des émissions moyennes de CO_2. Ces résultats confirment l'efficacité et les avantages de notre approche de contrôle par rapport aux autres approches évaluées.

4.6 Conclusion

Dans ce chapitre, nous avons abordé la problématique d'*ATSC,* tout en proposant une nouvelle approche basée *DRL*. L'objectif de cette dernière est de réduire la congestion routière dans les intersections isolées. Il s'agit d'une approche intelligente capable de s'adapter aux changements dynamiques du trafic routier. Elle utilise un agent *DRL* implémentant la méthode *DDQN* pour apprendre et optimiser le comportement du contrôleur des feux de signalisation en fonction des conditions de trafic réelles. Tout d'abord, nous avons discuté la problématique des intersections isolées à feux de signalisation. Ensuite, nous avons détaillé le principe de fonctionnement de l'approche proposée pour apprendre à ajuster les durées des feux de signalisation en fonction des conditions de trafic actuelles. Enfin, nous avons présenté les résultats expérimentaux en discutant ses performances en les comparants avec d'autres méthodes de la littérature.

Il est également important de noter que malgré les résultats satisfaisant qu'a produits l'approche proposée, elle peut être limitée à la gestion d'intersections isolées, sans prendre en compte les interactions et la coordination entre les feux de signalisation de différentes intersections. De ce fait, il est essentiel de proposer des approches intelligentes pour le contrôle adaptatif coopératif des feux de signalisation dans le contexte de multiples intersections. En adoptant une approche coopérative, il devient

possible de coordonner les feux de signalisation dans plusieurs intersections, afin d'optimiser le flux de trafic à l'échelle du réseau routier. Les approches intelligentes peuvent exploiter des techniques avancées comme par exemple *MARL*. Cela permet de prendre en compte les conditions de trafic globales, les priorités de passage, les flux de trafic interconnectés et d'autres facteurs pour optimiser la fluidité et l'efficacité de l'ensemble du réseau.

En proposant de telles approches intelligentes, il devient possible de réduire la congestion, les temps d'attente et les émissions de CO_2, tout en améliorant la sécurité. Ce type d'approche permet de prendre en compte les interactions complexes et les effets systémiques qui peuvent se produire lorsque plusieurs intersections interagissent, ouvrant ainsi la voie à une gestion plus efficace et holistique du trafic urbain. Dans cette perspective, nous avons consacré le chapitre suivant pour présenter notre troisième approche proposée. Il s'agit d'une technique basée *DRL* pour le contrôle adaptatif coopératif des feux de signalisation dans les réseaux à intersections multiples.

Chapitre 5 : Approche intelligente basée *MARL* pour le contrôle adaptatif des feux de signalisation dans les réseaux à intersections multiples

5.1 Introduction

Dans le chapitre précédent, nous avons essayé de démontrer les bénéfices que nous pouvons gagner en adoptant des méthodes adaptatives et intelligentes pour le contrôle des feux de signalisation dans les intersections isolées. Nous avons particulièrement prouvé l'efficacité de telles méthodes à travers une étude expérimentale menée pour évaluer notre approche basée *DRL* pour le contrôle des intersections isolées. Les différentes métriques exposant les résultats des différentes expérimentations expliquent clairement les avantages significatifs que peuvent offrir telles approches pour gérer efficacement les flux de circulation dans les intersections isolées. En termes d'adaptabilité, les approches basées *DRL* permettent aux feux de signalisation de s'adapter de manière dynamique aux conditions de circulation changeantes. Elles sont capables d'apprendre et de prendre des décisions en temps réel pour optimiser le flux de trafic en fonction des conditions spécifiques à un instant donné. Cela permet une meilleure utilisation des ressources et une gestion plus efficace des flux de trafic. En plus, elles permettent au système de contrôle d'apprendre à partir de l'expérience, en s'entraînant en interagissant directement avec l'environnement. Cela permet une amélioration continue des performances du système de contrôle des feux de

signalisation. En utilisant des techniques d'exploration et d'exploitation, ces approches peuvent trouver des politiques d'action optimales qui améliorent les performances globales du système de contrôle des feux de signalisation.

Par ailleurs, le choix entre le contrôle pour intersection isolée et le contrôle pour intersections multiples dépend de plusieurs conditions du trafic, notamment la densité du trafic. En effet, quand la densité du trafic est relativement faible et que les intersections ne sont pas fortement interconnectées, le contrôle pour intersection isolée peut être suffisant pour gérer efficacement le trafic. Cela est souvent le cas dans les zones résidentielles ou les zones moins densément peuplées. D'autre part, si le flux de trafic est principalement asymétrique, avec une intersection qui connaît un trafic plus important que les autres, il peut être plus efficace d'adopter un contrôle adaptatif pour intersection isolée plutôt que d'essayer de coordonner plusieurs intersections. En outre, lorsque les intersections sont fortement interconnectées et que les interactions entre les flux de trafic aux différentes intersections sont complexes, il peut être nécessaire d'adopter un contrôle pour intersections multiples pour optimiser la coordination et améliorer les performances globales du réseau (Tan et al., 2020; Zhong, 2021).

En adoptant un objectif principal d'optimiser les performances globales du réseau de circulation, tels que les temps de trajet moyens, la fluidité du trafic ou la réduction des émissions, le contrôle pour intersections multiples peut être préférable car il permet une coordination et une optimisation globales de performance. En outre, le choix entre le contrôle pour intersection isolée et le contrôle pour intersections multiples peut également dépendre des ressources disponibles, telles que les infrastructures de communication, les systèmes de détection de trafic, les budgets et les contraintes opérationnelles. La mise en œuvre d'un contrôle pour intersections multiples peut nécessiter des investissements plus importants en termes de systèmes de communication et de coordination.

Dans cette optique, nous pouvons conclure que la décision de choix entre le contrôle pour intersection isolée et le contrôle pour intersections multiples peut varier en fonction des besoins spécifiques de chaque situation. Toutefois, l'utilisation d'approches pour contrôler plusieurs intersections de manière coopérative devrait être préférable et peut être bénéfique (El-Tantawy et al., 2013). D'autre part, assurer une coopération entre les contrôleurs de plusieurs intersections peut souvent être nécessaire pour améliorer les performances puisque la majorité des cas de congestion du trafic sont dus à l'interférence entre eux (Huo et al., 2020). C'est pourquoi les problèmes de contrôle multi-intersections ont, ces dernières années, attiré une attention croissante de

nombreux chercheurs (Boukerche et al., 2022; Liu et al., 2021; Lee et al., 2020; Liu et al., 2017). Néanmoins, dans certains cas, une combinaison des deux approches peut être utilisée pour tirer parti des avantages de chaque méthode et répondre aux exigences spécifiques du système de circulation.

Dans ce chapitre, nous présentons notre troisième contribution dans le cadre de cette thèse de Doctorat (Haddad et al., 2022a). Il s'agit d'une approche intelligente basée *DRL* pour le contrôle coopératif des feux de signalisation de plusieurs intersections adjacentes. Par conséquent, chaque intersection est considérée comme un agent *DRL*. En permettant la communication entre les agents, cette approche permet aux agents de partager leurs décisions et observations les uns avec les autres, ce qui conduit à un comportement synergique de l'ensemble des agents plutôt qu'à une série d'actions individuelles. Cette contribution prend en compte les informations sur le trafic aux intersections voisines en partageant les *récompenses*, les *états* et les valeurs des *actions*. Ainsi, la valeur *Q* optimale globale pour plusieurs intersections est estimée en se basant sur ces valeurs partagées. En effet, le réseau de trafic multi-intersections dans une région est d'abord modélisé comme un *SMA*. Chaque agent contrôle une intersection spécifique en utilisant un *DQN* et transfère les *récompenses*, les *états* et les *actions* les plus récents de ses voisins vers sa propre fonction de perte lors du processus d'apprentissage.

5.2 Formulation du problème et objectifs

Dans cette étude de recherche, nous considérons un environnement composé de multiples intersections signalisées (N > 1 intersections), et nous adoptons une approche basée *DRL* pour le contrôle des feux de circulation afin de réduire la congestion du trafic. Il est supposé que chaque contrôleur d'intersection est capable d'ajuster les paramètres de ses phases en fonction des conditions réelles du trafic. De plus, toutes les données de trafic sont collectées automatiquement et en temps réel à l'aide de capteurs installés dans les rues du réseau routier.

Dans la Figure 5.1(a), un modèle de réseau routier à quatre intersections ($N = 4$) est particulièrement considéré et chaque intersection est contrôlée par un contrôleur d'éclairage. Par conséquent, on note $IC = \{ic1, ic2, ic3, ic4\}$ un ensemble de *quatre* contrôleurs d'intersection dans les zones proches, qui coopèrent et se coordonnent pour contrôler la circulation. On suppose que chaque intersection a quatre côtés d'entrée-sortie, où chaque côté se compose de deux voies. Il y a deux flux de circulation circulant dans le même sens : voies extérieures (*gauche*) et intérieures (*droite : celle la plus proche du trottoir*). La voie extérieure est réservée aux véhicules qui tournent à gauche, la voie

intérieure est réservée aux véhicules qui circulent tout droit ou tournent à droite. A chaque voie, deux capteurs (notés *S1* et *S2*) sont placés pour détecter avec précision le flux de circulation des véhicules en temps réel. On suppose que la distance (notée *D*) entre *S1* et *S2,* qui représente la zone de surveillance, est suffisamment longue pour mesurer l'évolution de la file d'attente.

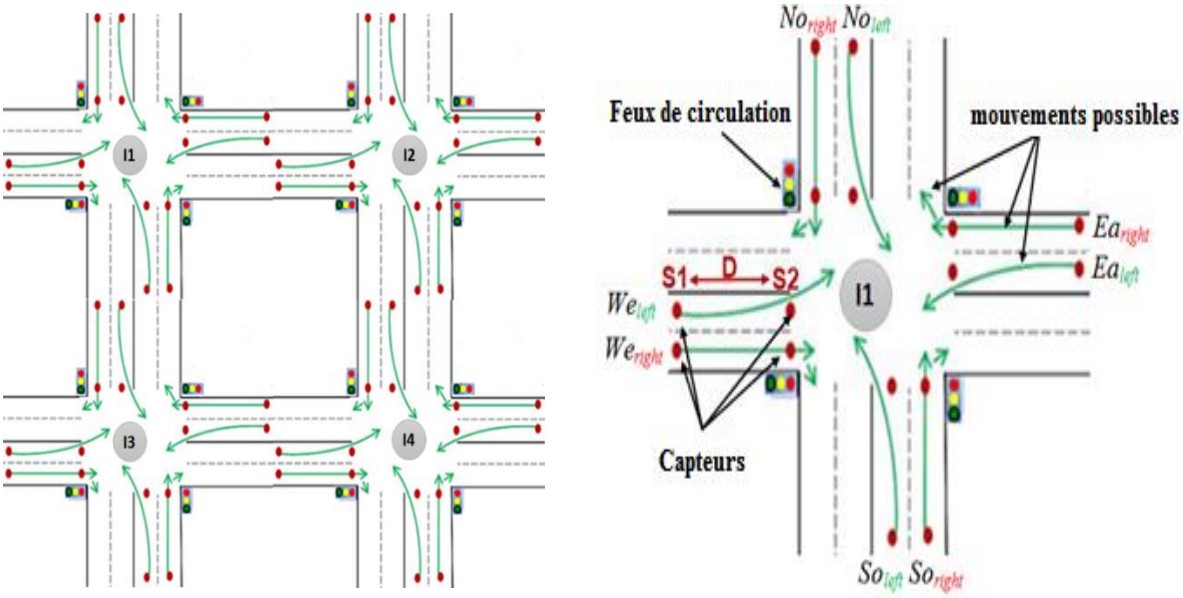


Figure 5.1(a) Modèle de réseau routier globale (avec N=4) (b) Un sous-système à une intersection du modèle de réseau routier.

La Figure 5.1(b) présente un sous-système à une intersection du modèle de réseau routier adopté. Il comprend plusieurs composants importants : contrôleur d'intersection (*IC* en anglais *Intersection Controller*), feux de circulation (*TL* en anglais *Traffic Light*) et pour chaque voie deux capteurs d'arrivée et de départ (*S1* et *S2*) sont misent en place. En considérant le véhicule passant à l'intersection, une voie peut être déterminée à la fois par le chemin et la direction. Quatre directions sont à considérer (*No* (*nord*), *So* (*sud*), *Ea* (*est*), *We* (*ouest*)) où chacune a deux chemins dans la direction entrante à savoir : (*Tourner à gauche* (*gauche*) et *Aller tout droit/Tournez à droite*). En conséquence, il y a au maximum huit voies opérationnelles dans le modèle, chacune étant définie par une combinaison chemin-direction : (No_{left}, No_{right}, Ea_{left}, Ea_{right}, So_{left}, So_{right}, We_{left}, We_{right}). Le contrôle d'accès aux zones de conflit est garanti par un système de plusieurs contrôleurs des feux de signalisation. En effet, lorsqu'une zone de conflit d'une intersection est occupée par des véhicules d'une voie, l'*IC* correspondant verrouille toutes les autres voies en opposition. Cela est assuré en utilisant trois lumières auxquelles une valeur temporelle est associée pour chaque phase. En note *T* : *la durée du cycle de circulation*, *G* : *la période de feu vert d'une phase en secondes*, *Rd* : *la période de feu rouge en secondes*, *Tpass* : *le temps nécessaire à un véhicule pour traverser une intersection*, et *Q* : *la file d'attente longueur*

pour une direction. On suppose que *G* ne dépasse pas le temps maximum prédéterminé nécessaire au feu vert (noté *Tmax*). Les valeurs *Rd* et *G* sont calculées par les équations suivantes :

$$Rd = T - G \tag{5.1}$$

$$G = Q * Tpass \tag{5.2}$$

Nous visons à travers notre contribution de proposer une nouvelle approche intelligente permettant de contrôler de façon coopérative plusieurs feux de signalisation afin de bien gérer le flux de circulation dans les réseaux routiers à intersections multiples. Dans cette perspective, nous cherchons à optimiser plusieurs paramètres à savoir : l'*AQL*, l'*AWT*. Cette contribution adopte un modèle à base de *MARL*. Elle repose principalement sur une méthode de partage de connaissances, permettant à chaque agent de recueillir explicitement des informations pertinentes sur l'état du trafic du réseau, y compris celles observées par d'autres agents. Grâce à cette méthode, chaque agent a la capacité d'apprendre à la fois de ses propres expériences et de l'expérience collective accumulée grâce à la coopération du système, en utilisant les expériences des autres agents pour optimiser sa propre stratégie. Par conséquent, chaque agent peut influencer l'environnement, ce qui entraîne des conséquences variées en fonction des actions exécutées par les autres agents.

Dans cette optique, nous basons notre contribution sur l'hypothèse stipulant que l'ensemble des agents contrôle le système de gestion des feux de circulation au moyen de leurs actions individuelles, sous la condition que les conséquences des actions d'un agent coïncident avec celles des autres. De plus, les agents adjacents partagent mutuellement les valeurs de leurs actions, incluant les *états*, les *actions* et les *récompenses*.

En utilisant le *Modèle du Processus Décisionnel de Markov Multi-Agents (MDPMA)*, notre problématique peut être décrite par un 6-uplet $\langle AG, S, A, P, R, \gamma \rangle$. Ici, *AG* représente l'ensemble des agents, *S* désigne l'ensemble des états, *A* englobe l'ensemble des actions conjointes exécutées par les agents, *P* est la matrice de probabilité de transition d'état, *R* englobe les récompenses anticipées et $\gamma \in [0, 1]$ représente le facteur d'actualisation utilisé pour évaluer l'importance des récompenses futures et immédiates. Il est important de noter que chaque agent $Agi \in AG$ possède son propre ensemble d'actions individuelles *Ai*. Par conséquent, à chaque instant *t*, un agent Agi observe l'état actuel S_{it}, puis choisit une action a_{it} en accord avec une politique modélisant la sélection d'actions pour cet agent. Notre contribution s'inscrit dans cette problématique en visant à décrire le fonctionnement de notre nouvelle approche intelligente, conçue pour améliorer la qualité du trafic dans les réseaux routiers à

plusieurs intersections. En effet, le principe de fonctionnement de ladite approche est bien détaillé dans la section suivante.

5.3 Approche proposée

Dans cette étude, les actions possibles (notée *Actions_Possibles*) sont définies par un ensemble défini comme suit :

Actions Possibles = {*NSG*, *EWG*, *NSLG*, *EWLG*}

Chaque action de set signifie la phase du signal. De plus, les notations *No*, *So*, *Ea*, *We*, *L*, *G* signifient respectivement *Nord*, *South, Est, Ouest, left* et *Green*. Par exemple, *NSG* représente la phase où le signal est *vert* pour le trafic allant du *nord* au *south*, et *EWG* correspond à la phase où le signal est *vert* pour le trafic allant de l'*Est* à *Ouest*. D'autre part, lorsqu'une action est sélectionnée, elle entraîne un changement d'état de l'environnement, passant de l'état actuel S_{it} à un nouvel état S_{it+1}. Par conséquent, l'agent perçoit le nouvel état S_{it+1} et reçoit une récompense R_{it+1} en fonction de la fonction de probabilité de transition d'état. Ainsi, l'agent Ag_i évolue dans son environnement et génère une séquence d'*états-actions-récompenses*, que l'on appelle une trajectoire, définie comme suit : S_{i1}, A_{i1}, R_{i2}, S_{i2}, A_{i2}, R_{i3}...

Au niveau d'un agent Ag_i, une politique notée π qui maximise la *récompense cumulée* attendue pourrait être formulée comme suit :

$$\pi: A_i x S_i([0,1]) \tag{5.3}$$

$$\pi(a, s) = Pr(a_t = a \mid s_t = s) \tag{5.4}$$

Où : Pr est la probabilité de transition (au pas de temps t) de l'état $S_t = s$ à l'état S_{t+1} sous l'action a.

En notant γ comme facteur d'actualisation déterminant l'importance des récompenses futures, la récompense cumulée attendue au pas de temps t noté R_t est définie comme suit :

$$R_t = \sum_{k=0}^{\infty} \gamma^k \, r_{t+k} \tag{5.5}$$

où la récompense r_t est définie comme une différence entre le temps d'attente actuel (à l'étape t) et prévu (à l'étape t+1) de tous les véhicules, formulée comme suit :

$$r_t = -(W_{t+1} - W_t) = W_t - W_{t+1} \tag{5.6}$$

Cette investigation se trouve corroborée au moyen de l'algorithme *Deep Q-learning* (Wan and Hwang, 2018), qui est une combinaison des deux aspects largement adoptés dans le domaine du *RL* qui sont le *Deep Neural Networks* (*DNN*) (Miikkulainen et al., 2017) et *Q-Learning* (Watkins and Dayan, 1992). De cette manière, la *fonction Q,* qui anticipe l'*état* et l'*action* liée à la *récompense* future escomptée, s'articule au moyen de la formulation suivante :

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha\,[r_t + \gamma \max_a Q_t\,(s_{t+1}, a') - Q_t(s_t, a_t)] \qquad (5.7)$$

où $Q_{t+1}(s_t,\ a_t)$ est la nouvelle valeur de la *fonction Q* après la modification à l'étape d'apprentissage st, tandis que s_{t+1} est l'état atteint lors de l'exécution de a_{t+1}. Le paramètre $\alpha \in [0, 1]$ correspond au taux d'apprentissage et r_t représente la récompense immédiate.

5.3.1 Coopération entre les agents adjacents

De nos jours, la *MARL* coopérative est un domaine de recherche très intéressant dont il a attiré l'attention de nombreux chercheurs de diverses disciplines. Les chercheurs dans le domaine de la gestion du trafic ont manifesté un intérêt considérable, en particulier, pour l'amélioration de la qualité du trafic en proposant des méthodes coopératives intelligentes de gestion des feux de circulation. Cette nouvelle tendance pourrait jouer un rôle primordial dans l'optimisation du trafic routier (Chu et al., 2020c; Huo et al., 2020; Kim and Jeong, 2019; Lee et al., 2020; Ma et al., 2021).

D'autre part, une idée importante stipulant que "*Lorsque plusieurs agents coopèrent entre eux, dans la plupart des cas, les utilités totales seront supérieures aux utilités de tous les agents sans coopération*" (Zhang and Zhang, 2020), Cette idée nous a incité à proposer une nouvelle contribution s'appuyant sur la coopération de plusieurs agents afin d'optimiser le trafic dans un réseau routier comportant plusieurs intersections. En effet, plusieurs agents participent à la gestion du trafic, chacun étant chargé de réguler la circulation à une intersection spécifique en ajustant les feux de signalisation associés. La coopération d'agents se manifeste par l'intégration, au cours du processus d'apprentissage, des mécanismes d'interaction permettant aux agents adjacents d'échanger les valeurs des paramètres tels que : *états, actions et récompenses* (Figure 5.2). Par conséquent, la sélection d'action par un agent dépend non seulement de ses propres valeurs d'*état*, d'*action* et de *récompense*, mais aussi de celles de ses voisins. C'est ainsi que le système de contrôle peut équilibrer le flux de trafic entre plusieurs intersections adjacentes, tout en améliorant les performances globales du réseau routier. La mise à jour de la fonction *Q* pour chaque agent Ag_i se réalise de la manière suivante :

$$Q_{t+1}^{i}(s_t^i, a_t^i) = Q_t^i(s_t^i, a_t^i; \theta_i) + \alpha(t)\left[r_t^i + \gamma \max_{a'} Q_t^i(s_{t+1}^i, a'; \theta_i^*) - Q_t^i(s_t^i, a_t^i; \theta_i)\right] + \sum_{j \in N_a} r_{t-1}^j \quad (5.8)$$

où N_a est le nombre d'intersections adjacentes et θ_i, θ_i^* sont respectivement les paramètres du réseau d'évaluation et du réseau cible.

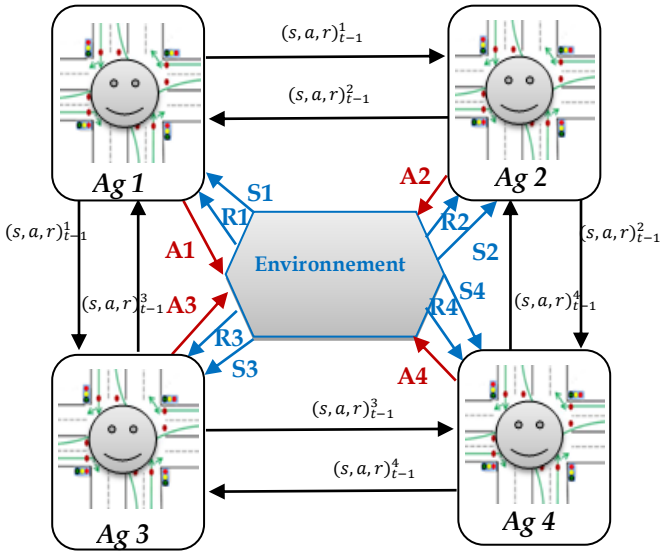


Figure 5.2 Structure du contrôleur de feux de circulation MARL coopératif proposé pour les quatre intersections.

Les valeurs optimales de la *fonction Q* des agents voisins, transmises à la fonction de perte du *réseau Q* pour l'apprentissage de la politique, facilitent la coopération en vue de contrôler les feux de signalisation pour plusieurs intersections. Pour chaque itération *i*, chaque agent met à jour les paramètres du réseau en calculant une fonction de perte en utilisant la méthode *MSE (Mean Squared Error*, en français, *Erreur Quadratique Moyenne*). Cette dernière est fréquemment utilisée comme mesure de l'écart moyen quadratique entre les valeurs prédites par un modèle et les valeurs réelles de l'ensemble de données. La formule de la *MSE* (θ_i) à l'itération *i* est la suivante :

$$MSE(\theta_i) = \frac{1}{m}\sum_{t=1}^{m}\{[r_t^i + \gamma \max_{a'} Q_t^i(s_{t+1}^i, a'; \theta_i^*) + \sum_{j \in N_a} r_{t-1}^j] - Q_t^i(s_t^i, a_t^i; \theta_i)\}^2 \quad (5.9)$$

où m représente la taille du lot, $max\, Q_t^i(s_{t+1}^i, a'; \theta_i^*)$ représente la valeur cible optimale pour toutes les actions sous l'état s_{t+1}^i. $Q_t^i(s_t^i, a_t^i; \theta_i)$ représente la sortie du réseau d'évaluation. Ainsi, la valeur cible notée y_t^i est calculée par l'équation suivante :

$$y_t^i = r_t^i + \gamma \max_{a'} Q_t^i\left(s_{t+1}^i, a'; \theta_i^*\right) \quad (5.10)$$

De plus, les fonctions d'activation sont considérées parmi les paramètres les plus importants dans le domaine de l'apprentissage en profondeur. Leur rôle crucial se manifeste principalement dans la détermination de la sortie d'un système, sa précision et son efficacité de calcul. De nombreuses fonctions d'activation ont été proposées dans la littérature, parmi lesquelles l'unité linéaire rectifiée (*ReLU*) (Glorot et al., 2011) a gagné une adoption répandue dans les travaux récents (Apicella et al., 2021). De plus, la capacité de la *ReLU* à accélérer l'entraînement du réseau de neurones sans compromettre significativement la précision de la généralisation (Krizhevsky et al., 2017) constitue l'un des facteurs qui a motivé l'adoption de cette fonction dans notre approche proposée.

5.3.2 Architecture globale

Chaque agent collecte toutes les données récentes du trafic à l'intersection correspondante tout en ajoutant les valeurs des *états*, des *actions* et des *récompenses* de ses agents voisins. Cela représente l'état courant (noté s_t) formant ainsi le vecteur d'entrée du *DNN*. La sortie du *DNN* est représentée par un vecteur de valeurs *Q* estimées pour différentes *actions* à appliquer à l'état S_t. Par la suite, l'agent prend une décision quant à la phase à appliquer à l'intersection. En conséquence, l'action correspondant à la valeur *Q* la plus élevée est soigneusement sélectionnée pour influencer le fonctionnement du réseau routier. Cette stratégie vise à optimiser les performances du système de contrôle du trafic en favorisant des actions associées à une plus grande valeur de *Q*, reflétant ainsi une meilleure efficacité et une gestion plus précise du trafic.

Comme illustré dans la Figure 5.3, l'approche proposée se distingue par sa capacité à gérer plusieurs intersections en utilisant un *DNN* entièrement connecté pour évaluer la valeur *Q* associée à chaque paire *état-action* à chaque intersection. Pour plus de précision, l'architecture du *DNN* se compose de deux couches cachées entièrement connectées, comprenant chacune *42 neurones* régulés par une fonction d'activation *ReLU*. Dans la couche de sortie, le nombre de neurones est équivalent à la taille de l'espace d'action de l'intersection correspondante, à savoir la configuration de la phase verte du feu de signalisation. Cette configuration permet une représentation détaillée des états du trafic, améliorant ainsi la précision et l'efficacité du contrôle du réseau routier.

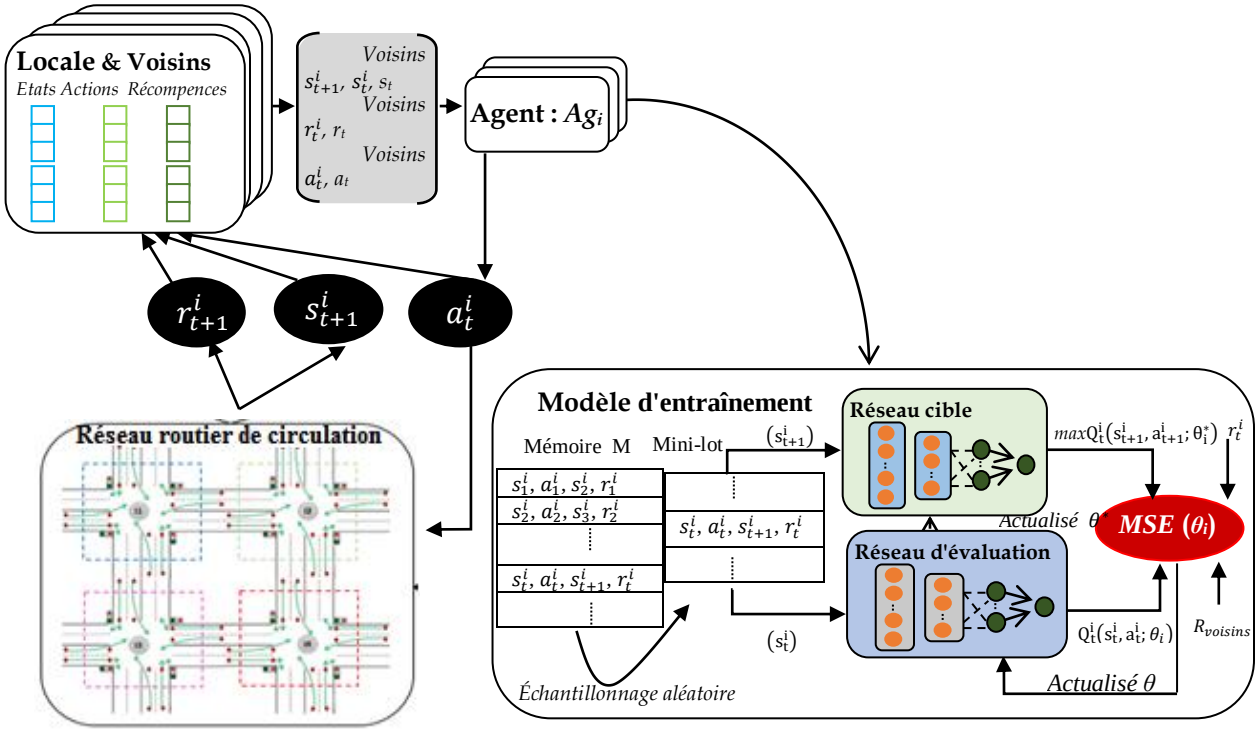


Figure 5.3 Architecture globale de l'approche proposée.

Il est utile de noter que la stabilisation des valeurs cibles améliore la convergence de l'apprentissage en réduisant la variance des cibles, ce qui peut rendre l'apprentissage plus robuste et accélérer la convergence. Par conséquent, l'approche proposée tient en compte de ce problème, en adoptant une stratégie qui se base sur l'utilisation de deux réseaux distincts, le réseau cible (en anglais : *target network*) et le réseau d'évaluation (en anglais : *evaluation network*). Le *réseau cible*, défini comme un réseau auxiliaire pour chaque agent contrôleur, consiste en une duplication directe du *réseau d'évaluation*. Il est utilisé afin d'estimer les valeurs *Q* cibles, nécessaires à la mise à jour des poids du *réseau d'évaluation*, et ce comme décrit par l'*équation* (*5.10*). Le *réseau d'évaluation* est le réseau principal utilisé pour prendre des décisions en fonction de l'état actuel et est constamment mis à jour au fur et à mesure de l'apprentissage. En effet, Il s'adapte continuellement aux nouvelles données d'apprentissage, reflétant ainsi les changements dans la politique apprise et permettant une meilleure prise de décision au fur et à mesure que l'agent explore l'environnement.

D'autre part, l'architecture globale adoptée utilise une mémoire notée *M*. Cette dernière joue un rôle crucial dans l'approche proposée en fournissant un mécanisme de stockage des expériences passées de l'agent. Son utilisation trouve ses motivations dans plusieurs aspects du processus d'entraînement. Tout d'abord, la mémoire *M* permet de conserver un historique des transitions antérieures, comprenant les *états*, les *actions*

prises, les *récompenses* obtenues, et les nouveaux états résultants. Cette archive d'expériences est essentielle pour pallier le manque de corrélation temporelle dans les données d'apprentissage, car les échantillons peuvent être réutilisés plusieurs fois, renforçant ainsi la stabilité de l'apprentissage. De plus, l'utilisation de la mémoire *M* favorise une exploration plus efficace de l'espace des *états* en permettant au *réseau d'évaluation* d'apprendre à partir d'une diversité d'expériences passées. Cela contribue à prévenir le phénomène d'oubli catastrophique qui peut se produire lorsque l'approche proposée apprend une nouvelle action, pour ensuite oublier comment effectuer une action précédemment apprise. En résumé, la mémoire tampon *M* offre un moyen efficace de gérer l'apprentissage par renforcement en stockant de manière organisée et réutilisable les informations clés issues des interactions passées de l'agent avec son environnement.

En ce qui concerne le calcul de perte, l'approche proposée adopte la fonction *MSE*. Cette dernière évalue la disparité entre les valeurs *Q* prédites par le *réseau d'évaluation* et les valeurs cibles calculées à l'aide du *réseau cible*. Elle est définie par l'équation (5.9), où la différence entre la valeur cible et la valeur prédite est élevée au carré pour chaque transition. Cette approche vise à minimiser la divergence entre les prédictions du modèle et les valeurs cibles, favorisant ainsi la convergence du *réseau d'évaluation*. L'utilisation de la *MSE* comme mesure de perte contribue à stabiliser l'apprentissage en réduisant la variance des cibles, améliorant ainsi la robustesse de l'algorithme et accélérant la convergence. En d'autre termes, pour tenir compte de l'impact des intersections voisines, les dernières valeurs de l'*état*, l'*action* et la *récompense* de chaque agent sont transmis à la fonction de perte $MSE(\theta_i)$.

5.3.3 Processus d'entraînement

L'approche proposée repose sur un processus d'entraînement élaboré, exploitant les principes de *MARL*. L'essence même de *MARL* réside dans la capacité des agents à apprendre et à ajuster leurs décisions en fonction des interactions avec un environnement dynamique. Dans ce contexte particulier, le processus d'entraînement revêt une importance cruciale, guidant chaque agent dans l'optimisation en temps réel de la gestion des feux de signalisation. Ce processus met en œuvre des mécanismes tels que des *réseaux d'évaluation* et des *réseaux cibles*, une mémoire tampon de relecture *M*, et une adaptation dynamique pour permettre à chaque agent de s'adapter aux conditions changeantes du trafic. Cette section détaille les composants fondamentaux de ce processus, démontrant comment chaque élément contribue à l'efficacité globale de l'approche, offrant ainsi une gestion adaptative des feux de signalisation aux intersections.

L'*Algorithme 5.1* décrit le processus d'entraînement adopté par l'approche proposée, mettant en œuvre plusieurs variables et paramètres qui sont préalablement initialisés. L'entraînement consiste en une séquence d'itérations, englobant des étapes successives telles que le choix d'une action, l'application de cette action, l'observation du réseau de trafic, le stockage de l'expérience, et la mise à jour du réseau de neurones profonds (*DNN*). Pour la sélection d'une action, l'algorithme utilise la méthode *ε-greedy* (Watkins, 1989), où la probabilité de choisir une action aléatoire est définie par ϵ, tandis que l'action choisie est celle ayant la plus grande valeur de la fonction Q dans l'état actuel (S_t). la méthode *ε-greedy* (voir *Algorithme 4.1*) équilibre l'*exploration*, permettant à l'agent d'améliorer sa compréhension des actions potentielles, et l'*exploitation*, sélectionnant l'action la plus prometteuse en fonction des estimations actuelles de la valeur d'action. Après avoir appliqué l'action à l'intersection, l'agent observe l'environnement, obtient la *récompense* et le nouvel *état*, et stocke cette expérience cruciale dans la mémoire *M*. Cette mémoire, limitée à *32 éléments* dans cette étude, joue un rôle essentiel en enrichissant l'historique d'interactions. Lorsque la mémoire *M* requiert des mises à jour périodiques, l'agent ajuste les poids du réseau (θi) en effectuant des tirages aléatoires dans *M* pour former des *mini-lots* d'échantillons. Pour chaque échantillon, la valeur cible est calculée, et l'algorithme de descente de gradient stochastique d'Adam (Kingma and Ba, 2014) est utilisé pour rétro-propager l'apprentissage dans le *DNN*, minimisant ainsi la fonction *MSE* et actualisant les valeurs de θ et θ^*.

Algorithme 5.1. Contrôleur de feux de circulation coopératif basé sur DRL pour plusieurs intersections

***Initialisation de touts les variables et parameters**:*

θ, θ^*, memoire de relecture M avec capacité limitée, ε,γ, N_episode, MaxSteps, taille du lot B

pour (*episode = 0; episode< N_episode; episode++*) {

pour (*chaque étape de l'épisode*) { *// le nombre d'étapes est limité à MaxSteps*

Observer s_t^i // état actuel de l'intersection i

a_t^i= sélection d'actions ε-greedy (s_t^i,ε)

Executer action a_t^i

Observer r_t^i // selon (*équation* (5.6))

Observer s_{t+1}^i // prochain état de intersection i

si(*la capacité de mémoire M_i est atteinte*){

Remplacer l'expérience la plus ancienne par l'actuelle ($exp_t^i = (s_t^i, a_t^i, s_{t+1}^i, r_t^i, R_t^{voisins})$)

}

sinon {

Ajouter le tuple $exp_t^i = (s_t^i, a_t^i, s_{t+1}^i, r_t^i, R_t^{voisins})$ to M_i

}

si *(longueur (Mi) >= B)*{

Échantillon aléatoire B d'expériences de Mi // noté E

pour *(chaque éxperience $e_k \in E$)* {

Calculer la valeur cible y_t^i // selon (*équation* (5.10))

}

Calculer MSE(θ_i), lors de la mise à jour de la valeur θ_i // selon (*équation* (5.9))

$\theta_i^* = \theta_i$

}

}

}

En outre, la performance du réseau entraîné, intrinsèquement liée à la fonction de perte, revêt une importance cruciale dans cette recherche. Pour évaluer la qualité de l'entraînement, cette étude valide les résultats à chaque étape du processus, notamment après chaque époque. Un total de 300 époques est utilisé pour comparer les résultats entre la perte d'apprentissage et la perte de validation. Comme illustré dans la figure 5.4, il est clair que la perte d'entraînement converge progressivement vers la perte de validation à mesure que le nombre d'époques augmente. Cette convergence indique un ajustement approprié du modèle, où ses performances sont excellentes sur les jeux de données d'apprentissage et de validation. En d'autres termes, le modèle entrainé démontre une qualité suffisante pour anticiper les résultats de manière appropriée dans

des conditions diverses, témoignant ainsi de son efficacité et de sa capacité à généraliser au-delà des données d'apprentissage. Cette évaluation systématique de la performance contribue à renforcer la fiabilité et la robustesse du modèle dans la prise de décision en temps réel dans des contextes complexes de gestion des feux de signalisation.

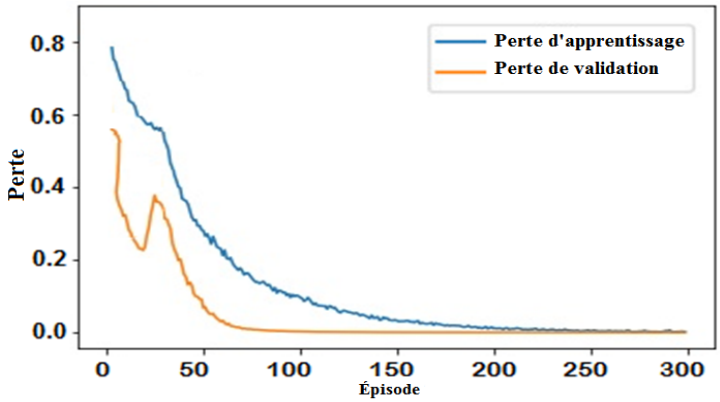


Figure 5.4 Perte d'apprentissage et la perte de validation.

5.4 Expérimentations

Pour mesurer les performances de l'approche proposée, de nombreuses expérimentations ont été menées en utilisant la plateforme *SUMO*. Cela implique l'utilisation du langage de programmation *Python* pour l'implémentation, avec des paramètres spécifiques pour le matériel et les scénarios. Ces expérimentations ont été exécutées sur une configuration matérielle comprenant un processeur *Intel i5-2310* ; quad-core cadencé à 2,5 GHz. La mémoire vive (*RAM*) allouée aux expériences était de *6 Go*, tandis qu'un *GPU Radeon (850 MHz)* était éventuellement mobilisé pour des tâches nécessitant des calculs parallèles, bien que nous n'ayons pas opté pour une utilisation spécifique du *GPU*.

Par ailleurs, les données des instances sont générées de manière aléatoire. Des comparaisons avec d'autres approches de l'état de l'art sont effectuées dans le cadre de deux scénarios distincts et de différents flux. Il s'agit de deux scénarios : (1) réseaux de grille 2×2 (quatre intersections) et (2) réseaux de grille 2×3 (six intersections) (voir Figure 5.5). Dans ces configurations, toutes les routes ont une longueur de *200 m*, le rayon de toutes les zones de surveillance est de *70 m*, et la longueur de tous les véhicules est fixée à *4 m*.

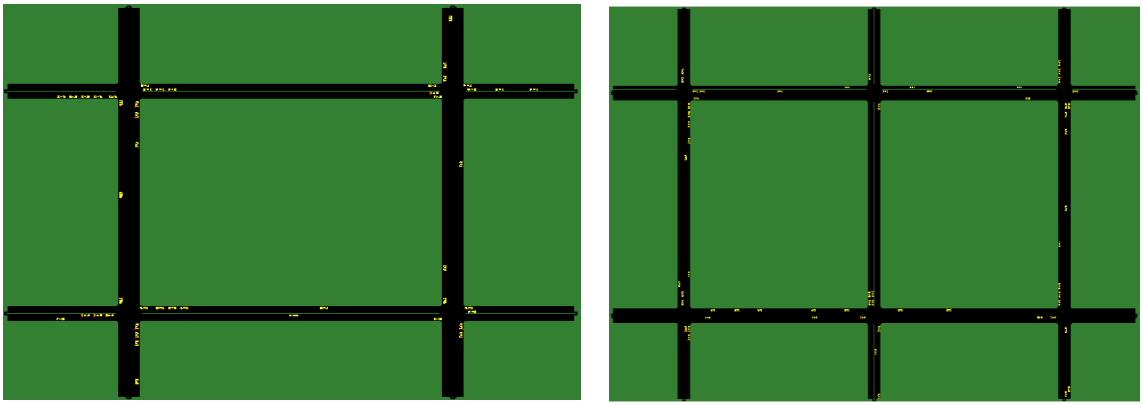


Figure 5.5 Structure du réseau routier pour les deux scénarios.

Le modèle proposé a été entraîné sur 300 épisodes en utilisant les données extraites de la mémoire de relecture *M*. Chaque épisode correspond à 4500 secondes du trafic simulé. Les tailles du lot |B| et M sont fixées respectivement à 32 et 20000. L'optimiseur utilisé est Adam, avec un taux d'apprentissage de 0,0001. Afin de permettre à l'agent de passer de l'*exploration* à l'*exploitation*, la valeur de ε diminue progressivement tout au long du processus d'entraînement, passant de 1 à 0,01. Des paramètres de simulation plus détaillés sont présentés dans le tableau 5.1.

Données de simulation			
Paramètres de simulation		***Paramètres du réseau routier***	
Paramètres	***Valeurs/ Description***	***Paramètres***	***Valeurs / Description***
Taille du lot |B|	*32*	*Intersections synthétisées*	*2×2, 2×3*
Taille de la mémoire de lecture	*20000*	*Région géographique de 2 ×2*	*350 m × 350 m*
Couches	*Dense*	*Région géographique de 2 ×3*	*350 m × 550 m*
Nombre de couches cachées	*2*	*Nombre de voies*	*2*
Fonction d'activation des couches cachées	*ReLU*	*Flux de trafic*	*RandomTrips*
Taux d'apprentissage α	*0.0001*	*Plateforme simulée*	*Sumo, keras*
facteur d'actualisation γ/ Fonction de perte	*0.99/MSE*	*Durée du jaune*	*4s*
Optimiseur	*Adam*	*Durée maximale du vert*	*30s*
Rayon du zone de surveillance	*70m*	*Vitesse du véhicule*	*[2-5](m/s)*
Episode	*300*	*Taille du véhicule*	*4m*
Départ ε	*1*	*Distance maximale entre les véhicules*	*0.5m*
Fin ε	*0.01*		

Tableau 5.1 Données de simulation.

5.5 Résultats et discussion

Dans cette section, nous présentons les résultats de nos expérimentations et engageons une discussion approfondie sur les implications et les interprétations de ces résultats obtenues. Sur la base de plusieurs métriques, notamment *AWT*, *AQL* et *AEC*, nous analyserons les performances de l'approche proposée par rapport à plusieurs approches de d'état de l'art à savoir : *QT-CDQN* (Ge et al., 2019), (Cooperative Deep Q-network with Q-value Transfer), *MADRL* (Liu et al., 2017), (Multi-Agent Deep Reinforcement Learning) et *CODRL* (Hussain et al., 2020), (COordinated Deep Reinforcement Learning), en mettant en lumière les points forts et les éventuels défis rencontrés.

Avant d'explorer les performances de notre approche, nous introduisons brièvement les méthodes de l'état de l'art que nous avons sélectionnées pour les comparaisons. Ces approches, à savoir *QT-CDQN*, *MADRL*, et *CODRL*, présentent des stratégies variées pour résoudre le problème de contrôle adaptatif du trafic. Nous examinons comment chacune d'entre elles aborde la coordination entre les intersections, la détermination des politiques de contrôle des feux de circulation, et l'utilisation de l'apprentissage par renforcement profond. En effet, l'approche *QT-CDQN* influence chaque agent en se basant sur les dernières actions de ses voisins. Elle recherche la stratégie optimale pour contrôler une intersection à l'aide d'un réseau *Q* profond. Quant à l'approche *MADRL*, l'algorithme *DQN* est employé dans cette approche pour déterminer la politique optimale de contrôle des feux de circulation. Aucun échange d'informations n'a lieu entre les agents, chacun prenant indépendamment des décisions selon un algorithme glouton. Enfin, l'approche *CODRL* s'appuie sur une réflexion où chaque intersection est modélisée par un agent. Chaque agent est formé via une technique *DRL* et utilise une *récompense* partagée pour l'ensemble des agents.

Étant donné que la densité de trafic influence de manière significative les performances de toute approche résolvant le problème *ATSC*. Par conséquent, nous avons décidé de varier ce paramètre selon trois niveaux clés : *faible*, *moyen* et *élevé*. Ces niveaux se traduisent par des productions de trafic de *50, 100* et *200* voitures par seconde, respectivement. Cette diversité permet d'évaluer la robustesse de notre approche face à des conditions de trafic variées. En effet, plusieurs tests de simulation ont été effectués, sur les deux scénarios (Figure 5.5), les résultats obtenus sont ainsi récapitulés dans le tableau 5.2.

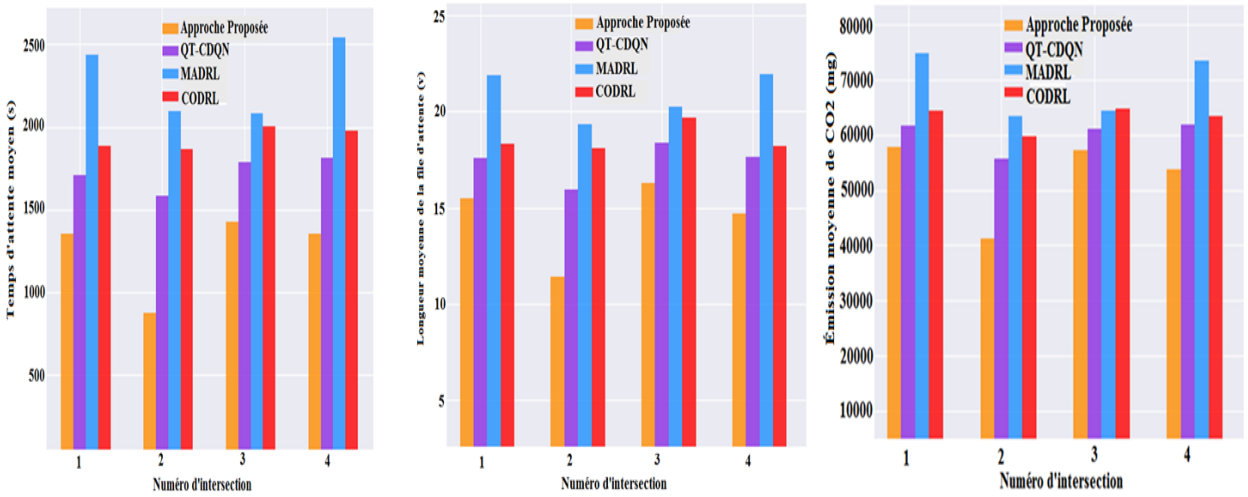
Les conclusions tirées de ces résultats démontrent que, tout au long de chaque épisode de simulation, l'approche que nous proposons présente des avantages significatifs dans divers cas pour les deux scénarios. En analysant l'ensemble des

données, il est notable que sur l'*AWT*, notre approche a réduit en moyenne de 26,8 % l'*AWT* de *QT-CDQN*, une moyenne de 47,28 % de l'*AWT* de *MADRL* et une moyenne de 27,29 % de l'*AWT* de *CODRL*. En ce qui concerne l'*AQL*, notre approche a réduit en moyenne de 20,54 % l'*AQL* du *QT-CDQN*, une moyenne de 35,21 % de l'*AQL* du *MADRL* et une moyenne de 21,74 % de l'*AQL* du *CODRL*. Enfin, sur l'*AEC*, elle a réduit en moyenne de 18,25 % l'*AEC* de *QT-CDQN*, une moyenne de 30,26 % de l'*AEC* de *MADRL* et une moyenne de 39,92 % de l'*AEC* de *CODRL*. Ces résultats soulignent l'efficacité de notre approche dans des conditions de trafic variées, établissant ainsi sa pertinence et son avantage par rapport aux trois approches de l'état de l'art.

Approches	**Métriques**	Premier Scénario (2x2)			Deuxième Scénario (2x3)		
		Faible	Moyen	Élevé	Faible	Moyen	Élevé
Approche Proposée	AWT	**1505,56**	**1442.74**	**4000.41**	**1037.52**	**2338.15**	**5639.11**
	AQL	**15.69**	**21.45**	**46.96**	**11.01**	**25.18**	**63.99**
	AEC	**47043.42**	**74967.58**	**169761.2**	**32045.03**	**78128.93**	**224423.6**
QT-CDQN	AWT	1574.04	3347.57	5536.98	1193.66	3718.08	7277.86
	AQL	16,38	33.65	56.34	15.21	34.41	72.48
	AEC	48572.10	107606.8	193994.8	46082.15	101895.9	248808.3
MADRL	AWT	2632.77	4918.49	7250.07	1639.37	4739.16	9197.08
	AQL	23.44	45.04	66.99	14.91	44.92	86.36
	AEC	66381.33	133672.9	220888.3	41723.10	135803.8	279932.2
CODRL	AWT	1900.32	4055.32	6255.32	1100.95	3016.90	6578.94
	AQL	20.48	39.72	64.72	13.89	30.99	70.07
	AEC	47562.89	114800.84	204800.84	34085.64	92452.98	239135.66

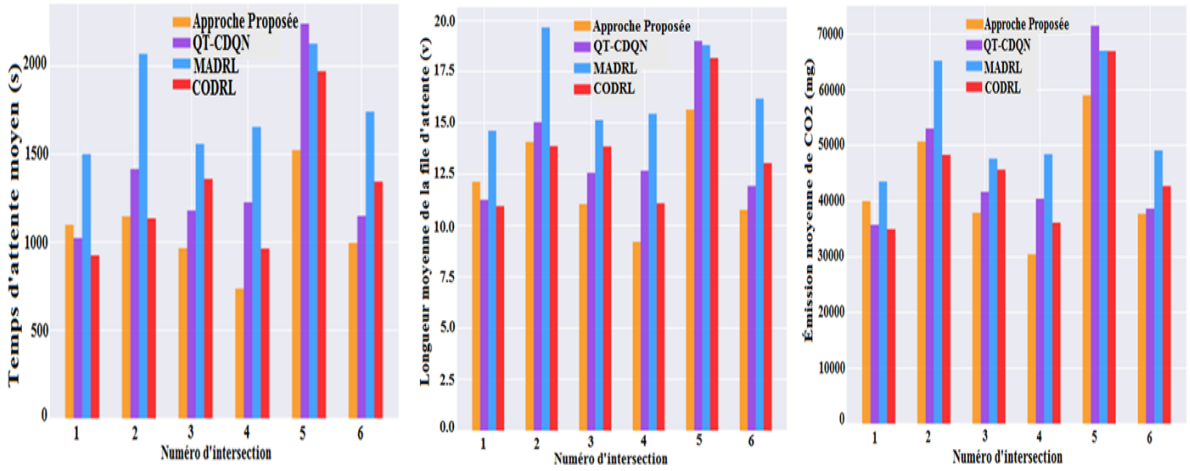
Tableau 5.2 Efficacité de l'approche proposée.

Les figures 5.6(a) à (c) et 5.7(a) à (c) illustrent les comparaisons de performances entre toutes les intersections, évaluées individuellement pour les métriques *AWT, AQL* et *AEC*, en variant entre les deux scénarios. Dans le premier scénario, notre approche surpasse de manière significative les autres méthodes, enregistrant les meilleures performances à chaque intersection pour toutes les métriques. Dans le deuxième scénario, elle maintient les meilleures performances à chaque intersection, à l'exception de la première, où le *QT-CDQN* surpasse dans toutes les métriques. Cependant, il est important de noter que les différences de performances sont minimes en termes de toutes les métriques.



(a) (b) (c)

Figure 5.6 Comparaisons des performances de chaque intersection pour les mesures AWT, AQL et AEC dans le premier scénario.



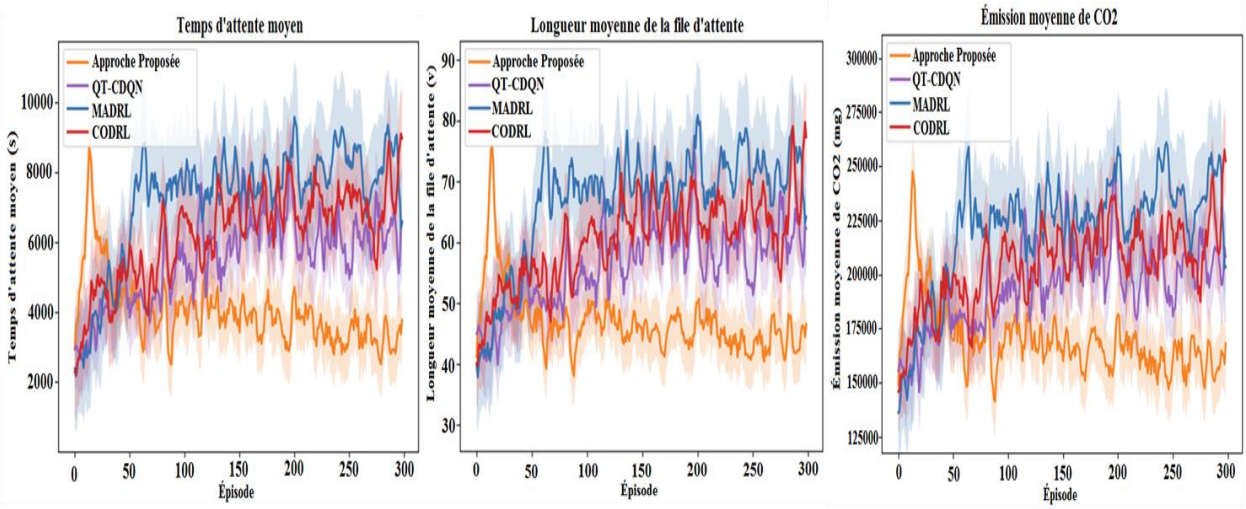
(a) (b) (c)

Figure 5.7 Comparaisons des performances de chaque intersection pour les mesures AWT, AQL et AEC dans le deuxième scénario.

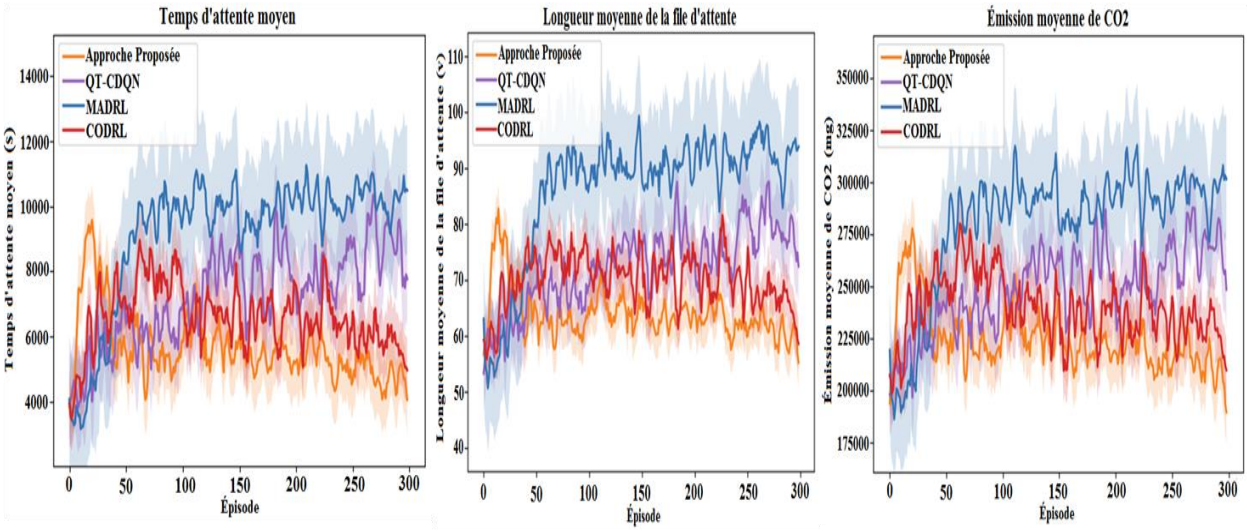
Les courbes présentées dans la figure 5.8 résultent de l'évaluation des performances sur 300 épisodes, chacun appliqué aux réseaux entraînés. Ces épisodes servent de base pour la comparaison des métriques *AWT*, *AQL* et *AEC*, offrant ainsi un aperçu global des performances de chaque approche tout au long du processus d'entraînement.

La figure 5.8 démontre les comparaisons de performances globales des métriques *AWT*, *AQL* et *AEC* sur les deux scénarios. Ces courbes sont obtenues en diffusant 300 épisodes sur les différents réseaux entraînés. De plus, comme indiqué dans le tableau

5.3 qui résume les valeurs moyennes et maximales de toutes les métriques, notre approche surpasse ensemble des approches comparées. Dans les deux scénarios, les valeurs maximales d'*AWT*, *AQL* et *AEC* sont observées respectivement aux épisodes 0−50, 150−200 et 250−300, pour chaque approche, à savoir notre approche, *QT-CDQN*, *MADRL* et *CODRL*. Ainsi, les différentes courbes démontrent qu'à partir de l'épisode 50, notre approche maintient une stabilité supérieure aux autres méthodes, tout en préservant ses performances optimales.



Scénario 1 : (a) **Scénario 1** : (b) **Scénario 1** : (c)



Scénario 2 : (a) **Scénario 2** : (b) **Scénario 2** : (c)

Figure 5.8 Graphiques des comparaisons de performances des trois métriques pour tous les agents des deux scénarios. (a), (b) et (c) représentent respectivement les changements des mesures AWT, AQL et AEC.

Scénarios	Métriques	Moyenne				Max			
		Approche proposée	**QT-CDQN**	**MADRL**	**CODRL**	**Approche proposée**	**QT-CDQN**	**MADRL**	**CODRL**
1st scénario	AWT	**1252.51**	1724.89	2288.73	1935.37	**1428,18**	1816,77	2539,83	2005.82
	AQL	**14.48**	17.39	20.86	18.59	**16.31**	18.35	21.88	19.71
	AEC	**52531.09**	60080.9	68995.44	62498.93	**57746,45**	61771,63	74721,44	64818.89
2nd scénario	AWT	**1078.07**	1372.61	1774.54	1283,15	**1524,10**	2240,12	2127,48	1969.79
	AQL	**12.14**	13.73	16.64	12.55	**15,65**	18,98	19,65	18.15
	AEC	**42629.55**	46834.17	53479.19	45788,30	**58995,92**	71490,64	66987,07	66951.38

Tableau 5.3 Aperçu de la valeur moyenne et maximale des métriques de trois approches pour les deux scénarios.

5.6 Conclusion

Dans ce chapitre, nous avons abordé le problème de la circulation *ATSC* dans le contexte des intersections multiples, et nous avons proposé une nouvelle approche collaborative basée sur le *MARL*. Notre démarche consiste à modéliser le problème en utilisant un ensemble d'agents *DRL*, chacun étant responsable de la gestion d'une intersection du réseau routier. Dans le but d'améliorer les performances globales du système de circulation, l'approche proposée favorise la coopération entre les agents, leur permettant ainsi de partager leurs décisions et observations mutuelles. En conséquence, l'ensemble des agents agit de manière synergique, formant ainsi un groupe cohérent plutôt qu'une simple collection d'individus. Plus précisément, nous estimons la valeur *Q globale* en agrégeant de multiples valeurs locales d'agents, telles que les valeurs Q, les *états* et les *actions*. Pendant le processus d'apprentissage, chaque agent gère une intersection individuellement en utilisant la méthode *DQN*, tout en prenant en considération les *états*, *actions* et *récompenses* récentes reçues de ses voisins, particulièrement lorsque cela impacte sa propre fonction de perte. Nos résultats expérimentaux, basés sur différentes configurations de paramètres, démontrent que notre approche surpasse les méthodes *QT-CDQN, MADRL* et *CODRL*, en termes des trois métriques : l'*AWT,* l'*AQL* et l'*AEC*.

L'approche proposée offre des avantages notables, allant de l'efficacité et de l'évolutivité pour le contrôle de multiples intersections à la gestion efficace des situations de congestion. Elle permet également une coopération fluide entre les différents agents pour prendre les décisions nécessaires. Cette coopération joue un rôle crucial dans l'amélioration globale de la qualité du trafic à toutes les intersections du réseau, bénéficiant ainsi à tous les agents contrôleurs potentiellement touchés par la congestion.

Conclusion générale et perspectives

A notre époque, la congestion du trafic est devenue un problème important dans le monde en particulier dans les grandes villes et métropoles. Cela conduit non seulement à augmenter le temps de trajet et à réduire les conditions de sécurité, mais aussi à exacerber le bruit et la pollution. La capacité des réseaux routiers est souvent insuffisante pour répondre aux exigences des véhicules, ce qui rend très difficile le contrôle de la fluidité du trafic. Les systèmes de transport présentent des coûts importants pour l'environnement, la santé humaine et la mobilité lorsqu'ils sont régis par des systèmes qui agissent de manière sous-optimale. Un meilleur contrôle générant des décisions optimales dans les systèmes de transport réduisent les coûts et profitent à la société dans son ensemble. Récemment, une littérature considérable s'est développée autour du thème du contrôle des feux de circulation et a montré son pouvoir d'améliorer l'efficacité de la régulation de la circulation. Une politique *TSC* efficace augmente considérablement la capacité des intersections en fluidifiant la circulation et en réduisant le temps d'attente des véhicules.

Après avoir exposé les divers apports formulés dans le cadre de cette thèse doctorale, cette section finale propose une synthèse et une conclusion des principales contributions de cette recherche, tout en abordant les perspectives pour de futures recherches connexes.

1. Contribution

Dans cette thèse, nous avons exposé un aperçu des progrès récents dans le domaine de *RL* appliqué au contrôle des feux de signalisation. Notre étude a porté sur les problématiques liées à la régulation des feux de circulation tant pour une intersection isolée que pour un réseau routier comprenant plusieurs intersections. Nous avons préconisé l'utilisation de techniques qui englobent à la fois la conceptualisation théorique et l'efficacité pratique de *RL* dans le domaine du contrôle des feux de

circulation, adaptées à divers scénarios de circulation. Nous avons également amélioré l'efficacité des méthodes traditionnelles de contrôle de feux de signalisation basées sur le *RL* dans un contexte de circulation plus réaliste. Dans ce qui suit, nous récapitulons les principales contributions de cette thèse :

— Proposition d'une approche adaptative qui s'appuie sur un modèle basé sur les technologies *IoT* pour collecter des données en temps réel sur les statistiques de trafic. Ces données constituent une base essentielle pour mesurer à la fois le temps d'attente et le nombre de véhicules dans toutes les phases afin de gérer dynamiquement le contrôle des feux de circulation.
— Proposition d'une nouvelle approche basée sur le *DRL*. Ainsi, le contrôleur du réseau de trafic dans une intersection isolée est modélisé comme un agent intelligent qui perçoit le codage d'état discret des informations de trafic comme les entrées du réseau. Notre contribution réside dans l'utilisation d'un Double Deep Q-Network (*DDQN*).
— Proposition d'une nouvelle approche coopérative basée sur *DRL* pour contrôler plusieurs intersections. Le problème est modélisé comme un système d'apprentissage par renforcement multi-agents (*MARL*), tandis que chaque agent est formé pour sélectionner la meilleure action pour contrôler une intersection en obtenant des informations sur l'état de ses voies locales.

2. Travaux futurs

Bien que les approches proposées donnent des résultats intéressants. Cette section présente néanmoins de nombreuses limites et défis dans les travaux futurs à savoir :

— La prise en compte des différentes catégories de véhicules qui peuvent avoir un impact sur la qualité de la solution de contrôle des feux de circulation comme les vélos, les bus, les véhicules d'urgence (les ambulances, les camions de pompiers et les voitures de police), etc. Ces catégories de véhicule sont une priorité plus élevée que les véhicules privées pour obtenir des signaux verts. En effet, en incluant des caractéristiques supplémentaires dans la représentation de l'état du trafic, la dimension de l'état du trac peut être augmentée de manière exponentielle, ce qui entraîne un coût d'apprentissage plus importants. Par conséquent, améliorer la conception de l'état avec moins de perte d'informations reste un défi dans les recherches futures.
— De plus, une étude d'ablation est prévue pour évaluer l'impact de différents paramètres sur l'efficacité des méthodes que nous avons développées. Ainsi, l'amélioration des aspects susmentionnés contribuera à la création d'un modèle de trafic réel plus précis et permettra d'acquérir des stratégies de régulation de la signalisation routière plus transférables.

— L’extension de la sphère des tests pour couvrir des réseaux routiers de plus grande envergure, comme ceux présents dans les zones urbaines, peut potentiellement améliorer la qualité de nos résultats, particulièrement dans des scénarios de circulation plus réalistes. En raison du nombre important de véhicules circulant et les différentes intersections qui composent un réseau routier, il est essentiel de prendre en compte d'autres circonstances, telles que les conditions météorologiques inhabituelles, susceptibles de provoquer des congestions particulièrement critiques à certaines intersections. Les méthodes de régulation des feux de signalisation que nous proposons doivent être en mesure d'évaluer efficacement et précisément ces conditions de circulation. Dans cette optique, nous proposons une autre perspective à cette thèse d’aborder le problème de la gestion des feux de circulation dans les réseaux à grande échelle tout en considérant les circonstances météorologiques.

Références

Abdelghaffar, H.M., Hao Yang, and Rakha, H.A. (2016). Isolated traffic signal control using a game theoretic framework. In 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), (Rio de Janeiro, Brazil: IEEE), pp. 1496–1501.

Abdulhai, B., Pringle, R., and Karakoulas, G.J. (2003). Reinforcement Learning for True Adaptive Traffic Signal Control. Journal of Transportation Engineering *129*.

AbuAli, N., and Abou-zeid, H. (2016). Driver Behavior Modeling: Developments and Future Directions. International Journal of Vehicular Technology *2016*, 1–12. https://doi.org/10.1155/2016/6952791.

Adams, W.F. (1936). Road traffic considered as a random series. (INCLUDES PLATES). Journal of the Institution of Civil Engineers *4*, 121–130. https://doi.org/10.1680/ijoti.1936.14802.

Alghamdi, T., Mostafi, S., Abdelkader, G., and Elgazzar, K. (2022). A Comparative Study on Traffic Modeling Techniques for Predicting and Simulating Traffic Behavior. Future Internet *14*, 294. https://doi.org/10.3390/fi14100294.

Apicella, A., Donnarumma, F., Isgrò, F., and Prevete, R. (2021). A survey on modern trainable activation functions. Neural Networks *138*, 14–32. https://doi.org/10.1016/j.neunet.2021.01.026.

Araghi, S., Khosravi, A., Johnstone, M., and Creighton, D. (2013). Q-learning method for controlling traffic signal phase time in a single intersection. In 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013), (The Hague, Netherlands: IEEE), pp. 1261–1265.

Ardekani, S., and Herman, R. (1987). Urban Network-Wide Traffic Variables and Their Relations. Transportation Science *21*, 1–16. https://doi.org/10.1287/trsc.21.1.1.

Arulkumaran, K., Deisenroth, M.P., Brundage, M., and Bharath, A.A. (2017). Deep Reinforcement Learning: A Brief Survey. IEEE Signal Process. Mag. *34*, 26–38. https://doi.org/10.1109/MSP.2017.2743240.

Aslani, M., Mesgari, M.S., Seipel, S., and Wiering, M. (2019). Developing adaptive traffic signal control by actor–critic and direct exploration methods. Proceedings of the Institution of Civil Engineers - Transport *172*, 289–298. https://doi.org/10.1680/jtran.17.00085.

Bakker, B., Whiteson, S., Kester, L., and Groen, F.C.A. (2010). Traffic Light Control by Multiagent Reinforcement Learning Systems. In Interactive Collaborative Information Systems, R. Babuška, and F.C.A. Groen, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 475–510.

Barceló, J. (2010). Models, Traffic Models, Simulation, and Traffic Simulation. In Fundamentals of Traffic Simulation, J. Barceló, ed. (New York, NY: Springer New York), pp. 1–62.

Bellman, R. (1957). A Markovian decision process. Journal of Mathematics and Mechanics 679–684.

Bingham, E. (2001). Reinforcement learning in neurofuzzy traffic signal control. European Journal of Operational Research *131*, 232–241. https://doi.org/10.1016/S0377-2217(00)00123-5.

Boukerche, A., Zhong, D., and Sun, P. (2022). A Novel Reinforcement Learning-Based Cooperative Traffic Signal System Through Max-Pressure Control. IEEE Trans. Veh. Technol. *71*, 1187–1198. https://doi.org/10.1109/TVT.2021.3069921.

Bouriachi, F., Zatla, H., Tolbi, B., Becha, K., and Ghermoul, A. (2021). Traffic Signal Control Model on Isolated Intersection Using Reinforcement Learning: A Case Study on Algiers City, Algeria. RIA *35*, 417–424. https://doi.org/10.18280/ria.350508.

Buisson, C., and Lesort, J.-B. (2010). Comprendre le trafic routier: méthodes et calculs (Lyon: Éd. du Certu).

Buşoniu, L., Babuška, R., and De Schutter, B. (2010). Multi-agent Reinforcement Learning: An Overview. In Innovations in Multi-Agent Systems and Applications - 1, D. Srinivasan, and L.C. Jain, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 183–221.

Camponogara, E., and Kraus, W. (2003). Distributed Learning Agents in Urban Traffic Control. In Progress in Artificial Intelligence, F.M. Pires, and S. Abreu, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 324–335.

Chakraborty, P.S. (2014). Real Time Optimized Traffic Management Algorithm. IJCSIT *6*, 119–136. https://doi.org/10.5121/ijcsit.2014.6408.

Chancey, K. (1991). Combined application of the hierarchical decision process with time series analysis: a telecommunications industry forecasting application. In Technology Management : The New International Language, (Portland, OR, USA: IEEE), p. 677.

Chandler, R.E., Herman, R., and Montroll, E.W. (1958). Traffic Dynamics: Studies in Car Following. Operations Research *6*, 165–184. https://doi.org/10.1287/opre.6.2.165.

Chin, Y., Bolong, N., Yang, S.S., and Teo, K.T.K. (2011). Q-Learning Based Traffic Optimization in Management of Signal Timing Plan. International Journal of Simulation: Systems, Science & Technology https://doi.org/10.5013/IJSSST.a.12.03.05.

Chisalita, L., and Shahmehri, N. (2002). A peer-to-peer approach to vehicular communication for the support of traffic safety applications. In Proceedings. The IEEE 5th International Conference on Intelligent Transportation Systems, (Singapore: IEEE), pp. 336–341.

Chu, T., Wang, J., Codeca, L., and Li, Z. (2020a). Multi-Agent Deep Reinforcement Learning for Large-Scale Traffic Signal Control. IEEE Trans. Intell. Transport. Syst. *21*, 1086–1095. https://doi.org/10.1109/TITS.2019.2901791.

Chu, T., Chinchali, S., and Katti, S. (2020b). Multi-agent Reinforcement Learning for Networked System Control. arXiv:2004.01339 [Cs, Stat].

Chu, T., Wang, J., Codeca, L., and Li, Z. (2020c). Multi-Agent Deep Reinforcement Learning for Large-Scale Traffic Signal Control. IEEE Trans. Intell. Transport. Syst. *21*, 1086–1095. https://doi.org/10.1109/TITS.2019.2901791.

«CNBC» (2022). Consumer News and Business Channel. Retrieved September 20, 2022, from https://www.cnbc.com/2019/02/11/americas-87-billion-traffic-jam-ranks-boston-and-dc-as-worst-in-us.html.

Cui, H., and Zhang, Z. (2021). A Cooperative Multi-Agent Reinforcement Learning Method Based on Coordination Degree. IEEE Access *9*, 123805–123814. https://doi.org/10.1109/ACCESS.2021.3110255.

Dalla Pozza, N., Buffoni, L., Martina, S., and Caruso, F. (2022). Quantum reinforcement learning: the maze problem. Quantum Mach. Intell. *4*, 11. https://doi.org/10.1007/s42484-022-00068-y.

Daston, L. (2020). Thomas S. Kuhn, The Structure of Scientific Revolutions (1962). Public Culture *32*, 405–413. https://doi.org/10.1215/08992363-8090152.

Deka, A., and Sycara, K. (2021). Natural Emergence of Heterogeneous Strategies in Artificially Intelligent Competitive Teams. In Advances in Swarm Intelligence, Y. Tan, and Y. Shi, eds. (Cham: Springer International Publishing), pp. 13–25.

El-Tantawy, S., Abdulhai, B., and Abdelgawad, H. (2013). Multiagent Reinforcement Learning for Integrated Network of Adaptive Traffic Signal Controllers (MARLIN-

ATSC): Methodology and Large-Scale Application on Downtown Toronto. IEEE Trans. Intell. Transport. Syst. *14*, 1140–1150. https://doi.org/10.1109/TITS.2013.2255286.

Eom, M., and Kim, B.-I. (2020). The traffic signal control problem for intersections: a review. Eur. Transp. Res. Rev. *12*, 50. https://doi.org/10.1186/s12544-020-00440-8.

Faye, S. (2014). Contrôle et gestion du trafic routier urbain par un réseau de capteurs sans fil. Thèse de doctorat. Paris, ENST.

Faye, S., Chaudet, C., and Demeure, I. (2012). A distributed algorithm for adaptive traffic lights control. In 2012 15th International IEEE Conference on Intelligent Transportation Systems, (Anchorage, AK, USA: IEEE), pp. 1572–1577.

Gao, J., Shen, Y., Liu, J., Ito, M., and Shiratori, N. (2017). Adaptive Traffic Signal Control: Deep Reinforcement Learning Algorithm with Experience Replay and Target Network. arXiv:1705.02755 [Cs].

Gartner, N.H. (1983). OPAC : A demand-responsive strategy for traffic signal control. Transportation Research Board 906 75–81.

Gartner, N.H., Little, J.D.C., and Gabbay, H. (1975). Optimization of Traffic Signal Settings by Mixed-Integer Linear Programming: Part I: The Network Coordination Problem. Transportation Science *9*, 321–343. https://doi.org/10.1287/trsc.9.4.321.

Ge, H., Song, Y., Wu, C., Ren, J., and Tan, G. (2019). Cooperative Deep Q-Learning With Q-Value Transfer for Multi-Intersection Signal Control. IEEE Access *7*, 40797–40809. https://doi.org/10.1109/ACCESS.2019.2907618.

Genders, W., and Razavi, S. (2016). Using a Deep Reinforcement Learning Agent for Traffic Signal Control. arXiv:1611.01142 [Cs].

Genders, W., and Razavi, S. (2018). Evaluating reinforcement learning state representations for adaptive traffic signal control. Procedia Computer Science *130*, 26–33. https://doi.org/10.1016/j.procs.2018.04.008.

Gipps, P.G. (1986). A model for the structure of lane-changing decisions. Transportation Research Part B: Methodological *20*, 403–414. https://doi.org/10.1016/0191-2615(86)90012-3.

Glorot, X., Bordes, A., and Bengio, Y. (2011). Deep Sparse Rectifier Neural Networks. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. In: JMLR Workshop and Conference Proceedings 315–323.

Gradinescu, V., Gorgorin, C., Diaconescu, R., Cristea, V., and Iftode, L. (2007). Adaptive Traffic Lights Using Car-to-Car Communication. In 2007 IEEE 65th Vehicular Technology Conference - VTC2007-Spring, (Dublin, Ireland: IEEE), pp. 21–25.

Greenshields, B., Bibbins, J., Channing, W., and Miller, H. (1935). A study of traffic capacity. In Highway Research Board Proceedings (Vol. 1935). National Research Council (USA), Highway Research Board.

Gu, J., Fang, Y., Sheng, Z., and Wen, P. (2020). Double Deep Q-Network with a Dual-Agent for Traffic Signal Control. Applied Sciences *10*, 1622. https://doi.org/10.3390/app10051622.

Guo, J., Kong, Y., Li, Z., Huang, W., Cao, J., and Wei, Y. (2019). A model and genetic algorithm for area-wide intersection signal optimization under user equilibrium traffic. Mathematics and Computers in Simulation *155*, 92–104. https://doi.org/10.1016/j.matcom.2017.12.003.

Haddad, T.A., Hedjazi, D., and Aouag, S. (2021). An IoT-Based Adaptive Traffic Light Control Algorithm for Isolated Intersection. In Advances in Computing Systems and Applications, M.R. Senouci, M.E.Y. Boudaren, F. Sebbak, and M. Mataoui, eds. (Cham: Springer International Publishing), pp. 107–117.

Haddad, T.A., Hedjazi, D., and Aouag, S. (2022a). A deep reinforcement learning-based cooperative approach for multi-intersection traffic signal control. Engineering Applications of Artificial Intelligence *114*, 105019. https://doi.org/10.1016/j.engappai.2022.105019.

Haddad, T.A., Hedjazi, D., and Aouag, S. (2022b). A New Deep Reinforcement Learning-Based Adaptive Traffic Light Control Approach for Isolated Intersection. In 2022 5th International Symposium on Informatics and Its Applications (ISIA), (M'sila, Algeria: IEEE), pp. 1–6.

Hall, F.L. (1997). Traffic stream Theory. US Federal Highway Administration *36*.

Hartanti, D., Aziza, R.N., and Siswipraptini, P.C. (2019). Optimization of smart traffic lights to prevent traffic congestion using fuzzy logic. TELKOMNIKA *17*, 320. https://doi.org/10.12928/telkomnika.v17i1.10129.

Hawi, R., Okeyo, G., and Kimwele, M. (2017). Smart traffic light control using fuzzy logic and wireless sensor network. In 2017 Computing Conference, (London: IEEE), pp. 450–460.

Head, K.L., Mirchandani, B., and Sheppard, D. (1992). Hierarchical Framework for Real-Time Traffic Control. Transportation Research Record.

Heredia, P.C., and Mou, S. (2019). Distributed Multi-Agent Reinforcement Learning by Actor-Critic Method. IFAC-PapersOnLine *52*, 363–368. https://doi.org/10.1016/j.ifacol.2019.12.182.

Hidas, P. (2005). Modelling vehicle interactions in microscopic simulation of merging and weaving. Transportation Research Part C: Emerging Technologies *13*, 37–62. https://doi.org/10.1016/j.trc.2004.12.003.

Huo, Y., Tao, Q., and Hu, J. (2020). Cooperative Control for Multi-Intersection Traffic Signal Based on Deep Reinforcement Learning and Imitation Learning. IEEE Access *8*, 199573–199585. https://doi.org/10.1109/ACCESS.2020.3034419.

Hussain, A., Wang, T., and Jiahua, C. (2020). Optimizing Traffic Lights with Multi-agent Deep Reinforcement Learning and V2X communication. arXiv:2002.09853 [Cs].

Johri, M., Goel, A., and Tiwari, A.K. (2012). Dynamic traffic signal control algorithm in intelligent transportation system through wireless sensor networks. International Journal of Engineering & Science Research *2*, 9.

Joo, H., and Lim, Y. (2021). Traffic Signal Time Optimization Based on Deep Q-Network. Applied Sciences *11*, 9850. https://doi.org/10.3390/app11219850.

Joo, H., Ahmed, S.H., and Lim, Y. (2020). Traffic signal control for smart cities using reinforcement learning. Computer Communications *154*, 324–330. https://doi.org/10.1016/j.comcom.2020.03.005.

Kato, S., Tsugawa, S., Tokuda, K., Matsui, T., and Fujii, H. (2002). Vehicle control algorithms for cooperative driving with automated vehicles and intervehicle communications. IEEE Trans. Intell. Transport. Syst. *3*, 155–161. https://doi.org/10.1109/TITS.2002.802929.

Kekuda, A., Anirudh, R., and Krishnan, M. (2021). Reinforcement Learning based Intelligent Traffic Signal Control using n-step SARSA. In 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), (Coimbatore, India: IEEE), pp. 379–384.

Kesting, A. (2008). Microscopic modeling of human and automated driving: Towards traffic-adaptive cruise control. PhD Thesis, Faculty of Traffic Sciences. Technische Universität Dresden (Germany).

Kim, D., and Jeong, O. (2019). Cooperative Traffic Signal Control with Traffic Flow Prediction in Multi-Intersection. Sensors *20*, 137. https://doi.org/10.3390/s20010137.

Kingma, D.P., and Ba, J. (2014). Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [Cs] http://arxiv.org/abs/1412.6980.

Kita, H. (1999). A merging–giveway interaction model of cars in a merging section: a game theoretic analysis. Transportation Research Part A: Policy and Practice *33*, 305–312. https://doi.org/10.1016/S0965-8564(98)00039-1.

Klein, L.A., Mills, M.K., and Gibson, D.R. (2006). Traffic detector handbook. Turner-Fairbank Highway Research Center. *No. FHWA-HRT-06-108*.

Kolat, M., Kővári, B., Bécsi, T., and Aradi, S. (2023). Multi-Agent Reinforcement Learning for Traffic Signal Control: A Cooperative Approach. Sustainability *15*, 3479. https://doi.org/10.3390/su15043479.

Konda, V.R., and Tsitsiklis, J.N. (1999). Actor-Critic Algorithms. Advances in Neural Information Processing Systems *12*.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2017). ImageNet classification with deep convolutional neural networks. Commun. ACM *60*, 84–90. https://doi.org/10.1145/3065386.

Kuyer, L., Whiteson, S., Bakker, B., and Vlassis, N. (2008). Multiagent Reinforcement Learning for Urban Traffic Control Using Coordination Graphs. In Machine Learning and Knowledge Discovery in Databases, W. Daelemans, B. Goethals, and K. Morik, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 656–671.

Lee, D., and Lee, S.J. (2020). Motion predictive control for DPS using predicted drifted ship position based on deep learning and replay buffer. International Journal of Naval Architecture and Ocean Engineering *12*, 768–783. https://doi.org/10.1016/j.ijnaoe.2020.09.004.

Lee, D., He, N., Kamalaruban, P., and Cevher, V. (2020). Optimization for Reinforcement Learning: From a single agent to cooperative agents. IEEE Signal Process. Mag. *37*, 123–135. https://doi.org/10.1109/MSP.2020.2976000.

Li, Z., Yu, H., Zhang, G., Dong, S., and Xu, C.-Z. (2021). Network-wide traffic signal control optimization using a multi-agent deep reinforcement learning. Transportation Research Part C: Emerging Technologies *125*, 103059. https://doi.org/10.1016/j.trc.2021.103059.

Liang, X., Du, X., Wang, G., and Han, Z. (2019). A Deep Reinforcement Learning Network for Traffic Light Cycle Control. IEEE Trans. Veh. Technol. *68*, 1243–1253. https://doi.org/10.1109/TVT.2018.2890726.

Liao, Y., and Cheng, X. (2009). Study on Traffic Signal Control Based on Q-Learning. In 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery, (Tianjin: IEEE), pp. 581–585.

Lieberman, E.B., Lai, J., and Ellington, R.E. (1983). SIGOP III User's Manual: Technical Report. FHWA, Washington, DC.

Lighthill, M.J., and Whitham (1955). On kinematic waves II. A theory of traffic flow on long crowded roads. Proc. R. Soc. Lond. A *229*, 317–345. https://doi.org/10.1098/rspa.1955.0089.

Liu, B., and Ding, Z. (2022). A distributed deep reinforcement learning method for traffic light control. Neurocomputing *490*, 390–399. https://doi.org/10.1016/j.neucom.2021.11.106.

Liu, J., Zhang, H., Fu, Z., and Wang, Y. (2021). Learning scalable multi-agent coordination by spatial differentiation for traffic signal control. Engineering Applications of Artificial Intelligence *100*, 104165. https://doi.org/10.1016/j.engappai.2021.104165.

Liu, M., Deng, J., Xu, M., Zhang, X., and Wang, W. (2017). Cooperative Deep Reinforcement Learning for Traffic Signal Control. In: Proc. 23rd ACM SIGKDD Conf. Knowl. Discovery Data Mining (KDD), Halifax, NS, Canada 8.

Long-Ji, L. (1992). Reinforcement learning for robots using neural networks. Technical report, Carnegie-Mellon Univ Pittsburgh PA School of Computer Science.

Lu, K., Tian, X., Jiang, S., Lin, Y., and Zhang, W. (2023). Optimization Model of Regional Green Wave Coordination Control for the Coordinated Path Set. IEEE Trans. Intell. Transport. Syst. *24*, 7000–7011. https://doi.org/10.1109/TITS.2023.3263847.

Lu Shoufeng, Ximin, L., and Shiqiang, D. (2008). Q-Learning for Adaptive Traffic Signal Control Based on Delay Minimization Strategy. In 2008 IEEE International Conference on Networking, Sensing and Control, (Sanya, China: IEEE), pp. 687–691.

Ma, X., Yang, Y., Li, C., Lu, Y., Zhao, Q., and Jun, Y. (2021). Modeling the Interaction between Agents in Cooperative Multi-Agent Reinforcement Learning. arXiv:2102.06042 [Cs].

Maerivoet, S., and De Moor, B. (2005). Traffic Flow Theory. arXiv Preprint Physics/0507126.

Mannion, P., Duggan, J., and Howley, E. (2016). An Experimental Review of Reinforcement Learning Algorithms for Adaptive Traffic Signal Control. In Autonomic Road Transport Support Systems, T.L. McCluskey, A. Kotsialos, J.P.

Müller, F. Klügl, O. Rana, and R. Schumann, eds. (Cham: Springer International Publishing), pp. 47–66.

Mekky, A. (2007). The Cost of Congestion in the Greater Toronto Area. In 2007 Annual Conference and Exhibition of the Transportation Association of Canada: Transportation-An Economic Enabler (Les Transports: Un Levier Economique) Transportation Association of Canada (TAC).

Melo, F.S., and Ribeiro, M.I. (2007). Q-Learning with Linear Function Approximation. In Learning Theory, N.H. Bshouty, and C. Gentile, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 308–322.

Miikkulainen, R., Liang, J., Meyerson, E., Rawal, A., Fink, D., Francon, O., Raju, B., Shahrzad, H., Navruzyan, A., Duffy, N., et al. (2017). Evolving Deep Neural Networks. arXiv:1703.00548 [Cs].

Miletić, M., Ivanjko, E., Gregurić, M., and Kušić, K. (2022). A review of reinforcement learning applications in adaptive traffic signal control. IET Intelligent Trans Sys *16*, 1269–1285. https://doi.org/10.1049/itr2.12208.

Miller, A.J. (1963). Settings for Fixed-Cycle Traffic Signals. Journal of the Operational Research Society *14*, 373–386. https://doi.org/10.1057/jors.1963.61.

Mimbela, L.-E.Y., and Klein, L.A. (2007). Summary of Vehicle Detection and Surveillance Technologies used in Intelligent Transportation Systems. United States. Joint Program Office for Intelligent Transportation Systems.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. Nature *518*, 529–533. https://doi.org/10.1038/nature14236.

Montana, D.J., and Czerwinski, S. (1996). Evolving Control Laws for a Network of Traffic Signals. Genetic Programming 333–338.

Mousavi, S.S., Schukat, M., and Howley, E. (2017). Traffic light control using deep policy-gradient and value-function-based reinforcement learning. IET Intelligent Transport Systems *11*, 417–423. https://doi.org/10.1049/iet-its.2017.0153.

Nam Bui, K.-H., and Jung, J.J. (2018). Cooperative game-theoretic approach to traffic flow optimization for multiple intersections. Computers & Electrical Engineering *71*, 1012–1024. https://doi.org/10.1016/j.compeleceng.2017.10.016.

Ouessai, A. (2020). Analyse du trafic routier dans un contexte de sécurité routière. Thèse de Doctorat, Université d'Oran Des Sciences et Technologies.

Ozan, C., Baskan, O., Haldenbilen, S., and Ceylan, H. (2015). A modified reinforcement learning algorithm for solving coordinated signalized networks. Transportation Research Part C: Emerging Technologies *54*, 40–55. https://doi.org/10.1016/j.trc.2015.03.010.

Papageorgiou, G., Damianou, P., Pitsillides, A., Aphamis, T., Charalambous, D., and Ioannou, P. (2009). Modelling and Simulation of Transportation Systems: a Scenario Planning Approach. Automatika: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije, 50(1-2), 39-50.

Papageorgiou, M., Kiakaki, C., Dinopoulou, V., Kotsialos, A., and Yibing Wang (2003). Review of road traffic control strategies. Proc. IEEE *91*, 2043–2067. https://doi.org/10.1109/JPROC.2003.819610.

Perronnet, F. (2015). Régulation coopérative des intersections: protocoles et politiques. Doctoral Dissertation, Belfort-Montbéliard.

Pipes, L.A. (1953). An Operational Analysis of Traffic Dynamics. Journal of Applied Physics *24*, 274–281. https://doi.org/10.1063/1.1721265.

«Processus de décision markovien» (2023). Wikipédia. Retrieved April 05, 2023, from https://fr.wikipedia.org/wiki/Processus_de_d%C3%A9cision_markovien#cite_note-1.

Qu, Z., Pan, Z., Chen, Y., Wang, X., and Li, H. (2020). A Distributed Control Method for Urban Networks Using Multi-Agent Reinforcement Learning Based on Regional Mixed Strategy Nash-Equilibrium. IEEE Access *8*, 19750–19766. https://doi.org/10.1109/ACCESS.2020.2968937.

Radović, N., and Erceg, M. (2021). Hardware implementation of the upper confidence-bound algorithm for reinforcement learning. Computers & Electrical Engineering *96*, 107537. https://doi.org/10.1016/j.compeleceng.2021.107537.

Reuschel, A. (1950). Fahrzeugbewegungen in der Kolonne. Osterreichisches Ingenieur Archiv *4*, 193–215.

Richards, P.I. (1956). Shock Waves on the Highway. Operations Research *4*, 42–51. https://doi.org/10.1287/opre.4.1.42.

Rida, N., and Hasbi, A. (2019). Dynamic Traffic Lights Control for Isolated Intersection Based Wireless Sensor Network. In Innovations in Smart Cities Applications Edition 2, M. Ben Ahmed, A.A. Boudhir, and A. Younes, eds. (Cham: Springer International Publishing), pp. 1036–1044.

Robertson, D.I. (1968). TRANSYT : traffic network study tool. Fourth International Symposium on the Theory of TrafficFlow, Karlsruhe, Germany.

Robertson, D.I., and Bretherton, R.D. (1991). Optimizing networks of traffic signals in real time-the SCOOT method. IEEE Trans. Veh. Technol. *40*, 11–15. https://doi.org/10.1109/25.69966.

Rothrock, C.A., and Keefer, L.E. (1957). Measurement of Urban Traffic Congestion. Highway Research Board pp 1-13.

Salkham, A., Cunningham, R., Garg, A., and Cahill, V. (2008). A Collaborative Reinforcement Learning Approach to Urban Traffic Control Optimization. In 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, (Sydney, Australia: IEEE), pp. 560–566.

Sammoud, B. (2015). Contribution à la modélisation et à la commande des feux de signalisation par réseaux de Petri hybrides. Doctoral dissertation, Université de Technologie de Belfort-Montbeliard; Université de Tunis El Manar.

Savrasovs, M. (2011). Urban Transport Corridor Mesoscopic Simulation. In ECMS 2011 Proceedings Edited by: T. Burczynski, J. Kolodziej, A. Byrski, M. Carvalho, (ECMS), pp. 587–593.

Schneider, J., Wong, W.-K., Moore, A., and Riedmiller, M. (1999). Distributed Value Functions. In: Proceedings of International Conference on Machine Learning. 8.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal Policy Optimization Algorithms. arXiv preprint arXiv:1707.06347.

Shamsi, M., Rasouli Kenari, A., and Aghamohammadi, R. (2022). Reinforcement learning for traffic light control with emphasis on emergency vehicles. J Supercomput *78*, 4911–4937. https://doi.org/10.1007/s11227-021-04068-w.

Shani, G., Heckerman, D., and Brafman, R.I. (2005). An MDP-Based Recommender System. Journal of Machine Learning Research, 6(9).

Sims, A.G., and Dobinson, K.W. (1980). The Sydney coordinated adaptive traffic (SCAT) system philosophy and benefits. IEEE Trans. Veh. Technol. *29*, 130–137. https://doi.org/10.1109/T-VT.1980.23833.

Sun, D.J., and Kondyli, A. (2010). Modeling Vehicle Interactions during Lane-Changing Behavior on Arterial Streets: Modeling vehicle interactions during lane-changing behavior on arterial streets. Computer-Aided Civil and Infrastructure Engineering *25*, 557–571. https://doi.org/10.1111/j.1467-8667.2010.00679.x.

Sutton, R.S., and Barto, A.G. (2018). Reinforcement learning: An introduction. MIT Press.

Tan, K.L., Sharma, A., and Sarkar, S. (2020). Robust Deep Reinforcement Learning for Traffic Signal Control. J. Big Data Anal. Transp. *2*, 263–274. https://doi.org/10.1007/s42421-020-00029-6.

Tan, T., Bao, F., Deng, Y., Jin, A., Dai, Q., and Wang, J. (2019). Cooperative Deep Reinforcement Learning for Large-Scale Traffic Grid Signal Control. IEEE Trans. Cybern. 1–14. https://doi.org/10.1109/TCYB.2019.2904742.

Teknomo, K., Takeyama, Y., and Inamura, H. (2016). Review on Microscopic Pedestrian Simulation Model. arXiv preprint arXiv:1609.01808.

«The Economist» (2022). The cost of traffic jams. Retrieved September 20, 2022, from https://www.economist.com/the-economist-explains/2014/11/03/the-cost-of-traffic-jams.

«Théorie» (2023). Wikipédia. Retrieved Mai 14, 2023, from https://fr.wikipedia.org/wiki/Th%C3%A9orie.

Thorpe, T.L., and Anderson, C.W. (1996). Traffic Light Control Using SARSA with Three State Representations. Technical Report, Citeseer.

Touhbi, S., Babram, M.A., Nguyen-Huu, T., Marilleau, N., Hbid, M.L., Cambier, C., and Stinckwich, S. (2017). Adaptive Traffic Signal Control : Exploring Reward Definition For Reinforcement Learning. Procedia Computer Science *109*, 513–520. https://doi.org/10.1016/j.procs.2017.05.327.

Treiber, M., and Kesting, A. (2013). Traffic Flow Dynamics: Data, Models and Simulation (Berlin, Heidelberg: Springer Berlin Heidelberg).

Vidali, A., Crociani, L., Vizzari, G., and Bandini, S. (2019). A Deep Reinforcement Learning Approach to Adaptive Traffic Lights Management. In WOA 9.

Wallace, C.E., Courage, K.G., Reaves, D.P., Shoene, G.W., Euler, G.W., and Wilbur, A. (1988). TRANSYT-7F user's manual: Technical report. Prepared for FHWA by the Transportation Research Center, University of Florida, Gainesville, FL.

Wan, C., and Hwang, M. (2018). Value-based deep reinforcement learning for adaptive isolated intersection signal control. IET Intelligent Transport Systems *12*, 1005–1010. https://doi.org/10.1049/iet-its.2018.5170.

Wang, B., He, Z., Sheng, J., and Liu, Y. (2023a). Multi-agent deep reinforcement learning with actor-attention-critic for traffic light control. Proceedings of the Institution of

Mechanical Engineers, Part D: Journal of Automobile Engineering 095440702311679. https://doi.org/10.1177/09544070231167986.

Wang, D., Wang, X., Chen, L., Yao, S., Jing, M., Li, H., Li, L., Bao, S., Wang, F.-Y., and Lin, Y. (2023b). TransWorldNG: Traffic Simulation via Foundation Model. *a*rXiv preprint arXiv:2305.15743.

Wang, T., Cao, J., and Hussain, A. (2021a). Adaptive Traffic Signal Control for large-scale scenario with Cooperative Group-based Multi-agent reinforcement learning. Transportation Research Part C: Emerging Technologies *125*, 103046. https://doi.org/10.1016/j.trc.2021.103046.

Wang, X., Ke, L., Qiao, Z., and Chai, X. (2021b). Large-Scale Traffic Signal Control Using a Novel Multiagent Reinforcement Learning. IEEE Trans. Cybern. *51*, 174–187. https://doi.org/10.1109/TCYB.2020.3015811.

Wang, Y., Liu, Y., Chen, W., Ma, Z.-M., and Liu, T.-Y. (2020). Target transfer Q-learning and its convergence analysis. Neurocomputing *392*, 11–22. https://doi.org/10.1016/j.neucom.2020.02.117.

Wang, Z., Schaul, T., Hessel, M., van Hasselt, H., Lanctot, M., and de Freitas, N. (2016). Dueling Network Architectures for Deep Reinforcement Learning. In International Conference on Machine Learning.

Wardrop, J.G. (1952). ROAD PAPER. SOME THEORETICAL ASPECTS OF ROAD TRAFFIC RESEARCH. Proceedings of the Institution of Civil Engineers *1*, 325–362. https://doi.org/10.1680/ipeds.1952.11259.

Watkins, C.J.C.H. (1989). Learning from Delayed Rewards. Ph.D. Thesis. University of Cambridge, Cambridge, England.

Watkins, C.J.C.H., and Dayan, P. (1992). Q-learning. Mach Learn *8*, 279–292. https://doi.org/10.1007/BF00992698.

Wei, H., Zheng, G., Yao, H., and Li, Z. (2018). IntelliLight: A Reinforcement Learning Approach for Intelligent Traffic Light Control. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, (London United Kingdom: ACM), pp. 2496–2505.

Wei, H., Li, Z., Xu, N., Zhang, H., Zheng, G., Zang, X., Chen, C., Zhang, W., Zhu, Y., and Xu, K. (2019). CoLight: Learning Network-level Cooperation for Traffic Signal Control. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management - CIKM '19, (Beijing, China: ACM Press), pp. 1913–1922.

Whitham, G.B. (2011). Linear and nonlinear waves. John Wiley & Sons.

Wu, J. (2011). Utilisation de la conduite coopérative pour la régulation de trafic dans une intersection. Doctoral Dissertation, Université de Technologie de Belfort-Montbeliard.

Wu, Q., Wu, J., Shen, J., Du, B., Telikani, A., Fahmideh, M., and Liang, C. (2022). Distributed agent-based deep reinforcement learning for large scale traffic signal control. Knowledge-Based Systems *241*, 108304. https://doi.org/10.1016/j.knosys.2022.108304.

Xiao-Feng Chen and Zhong-Ke Shi (2002). Real-coded genetic algorithm for signal timing optimization of a single intersection. In Proceedings. International Conference on Machine Learning and Cybernetics, (Beijing, China: IEEE), pp. 1245–1248.

Xu, L.-H., Xia, X.-H., and Luo, Q. (2013). The Study of Reinforcement Learning for Traffic Self-Adaptive Control under Multiagent Markov Game Environment. Mathematical Problems in Engineering *2013*, 1–10. https://doi.org/10.1155/2013/962869.

Yan, F. (2012). Contribution à la modélisation et à la régulation du trafic aux intersections: Intégration des communications Véhicule-Infrastructure. Doctoral Dissertation, Université de Technologie de Belfort-Montbeliard.

Yan, X.T., and Shang, Z.L. (2022). Urban intelligent traffic signal coordination control system based on machine learning. Advances in Transportation Studies *4*.

Yang, S., Yang, B., Wong, H.-S., and Kang, Z. (2019). Cooperative traffic signal control using Multi-step return and Off-policy Asynchronous Advantage Actor-Critic Graph algorithm. Knowledge-Based Systems *183*, 104855. https://doi.org/10.1016/j.knosys.2019.07.026.

Yousef, K.M., Al-Karaki, J.N., and Shatnawi, A.M. (2010). Intelligent Traffic Light Flow Control System Using Wireless Sensors Networks. J. Inf. Sci. Eng. *26*, 753–768.

Zeng, J., Hu, J., and Zhang, Y. (2018). Adaptive Traffic Signal Control with Deep Recurrent Q-learning. In 2018 IEEE Intelligent Vehicles Symposium (IV), (Changshu: IEEE), pp. 1215–1220.

Zhang, H., and Zhang, S. (2020). Multi-Agent Reinforcement Learning. In Deep Reinforcement Learning, H. Dong, Z. Ding, and S. Zhang, eds. (Singapore: Springer Singapore), pp. 335–346.

Zhang, C., Tian, Y., Zhang, Z., Xue, W., Xie, X., Yang, T., Ge, X., and Chen, R. (2022). Neighborhood Cooperative Multiagent Reinforcement Learning for Adaptive Traffic Signal Control in Epidemic Regions. IEEE Trans. Intell. Transport. Syst. 1–12. https://doi.org/10.1109/TITS.2022.3173490.

Zhong, D. (2021). Reinforcement Learning-based Traffic Signal Control for Signalized Intersections. Doctoral Dissertation, Université d'Ottawa/University of Ottawa.

Zhou, P., Chen, X., Liu, Z., Braud, T., Hui, P., and Kangasharju, J. (2021). DRLE: Decentralized Reinforcement Learning at the Edge for Traffic Light Control in the IoV. IEEE Trans. Intell. Transport. Syst. *22*, 2262–2273. https://doi.org/10.1109/TITS.2020.3035841.

Annexe 1. Notre Production Scientifique

Publications dans des Revues Internationales:

Haddad, T. A., Hedjazi, D., & Aouag, S. (2022). **A deep reinforcement learning-based cooperative approach for multi-intersection traffic signal control**. Engineering Applications of Artificial Intelligence, 114, 105019.

Publication dans les Actes de Conférences Internationales :

Haddad, T. A., Hedjazi, D., & Aouag, S. (2022, November). **A New Deep Reinforcement Learning-Based Adaptive Traffic Light Control Approach for Isolated Intersection**. In 2022 5th International Symposium on Informatics and its Applications (ISIA) (pp. 1-6). IEEE.

Haddad, T. A., Hedjazi, D., & Aouag, S. (2021). **An IoT-based adaptive traffic light control algorithm for isolated intersection**. In International Conference on Computing Systems and Applications (pp. 107-117). Springer, Cham.