Age of Information Aware Radio Resource Management in Vehicular Networks: A Proactive Deep Reinforcement Learning Perspective

Xianfu Chen*, Member, IEEE, Celimuqge Wu*, Senior Member, IEEE, Tao Chen, Senior Member, IEEE, Honggang Zhang*, Senior Member, IEEE, Zhi Liu*, Senior Member, IEEE, Yan Zhang*, Fellow, IEEE, and Mehdi Bennis*, Senior Member, IEEE

Abstract—In this paper, we investigate the problem of age of information (AoI)-aware radio resource management for expected long-term performance optimization in a Manhattan grid vehicle-to-vehicle network. With the observation of global network state at each scheduling slot, the roadside unit (RSU) allocates the frequency bands and schedules packet transmissions for all vehicle user equipment-pairs (VUE-pairs). We model the stochastic decision-making procedure as a discrete-time single-agent Markov decision process (MDP). The technical challenges in solving the optimal control policy originate from high spatial mobility and temporally varying traffic information arrivals of the VUE-pairs. To make the problem solving tractable, we first decompose the original MDP into a series of per-VUE-pair MDPs. Then we propose a proactive algorithm based on long short-term memory and deep reinforcement learning techniques to address the partial observability and the curse of high dimensionality in local network state space faced by each VUE-pair. With the proposed algorithm, the RSU makes the optimal frequency band allocation and packet scheduling decision at each scheduling slot in a decentralized way in accordance with the partial observations of the global network state at the VUE-pairs. Numerical experiments validate the theoretical analysis and demonstrate the significant performance improvements from the proposed algorithm.

Index Terms—Vehicular communications, multi-user resource scheduling, Markov decision process, long short-term memory, deep reinforcement learning, Q-function decomposition.

I. INTRODUCTION

VEHICLE-TO-VEHICLE (V2V) communication is envisioned as one of the key enablers for spearheading the next generation intelligent transportation systems [1]–[6]. Yet this ad hoc type of vehicular communication requires intense coordinations among the vehicles in close proximity [1], [7], and the high vehicle mobility makes designing efficient radio resource management (RRM) schemes an extremely challenging problem [8]–[10]. There is a huge body of literature on RRM in V2V communications. In [11], Sun et al. proposed a separate resource block and power allocation algorithm for RRM in device-to-device based V2V communications. In [12], Yao et al. derived a loss differentiation rate adaptation scheme to meet the stringent delay and reliability requirements for V2 safety communications. In [13], Egea-Lopez et al. designed a fair adaptive beaconing rate algorithm for beacon rate control in inter-vehicular communications. In [14], Guo et al. developed a joint spectrum and power allocation algorithm based on the effective capacity theory with the aim of maximizing the sum ergodic capacity of vehicle-to-infrastructure links while guaranteeing the latency violation probability for V2 links. However, while interesting, most efforts have focused on the instantaneous performance optimization ignoring the network dynamics, such as the temporal and spatial variations in communication quality and traffic information.

A Markov decision process (MDP) framework has been widely applied to model the problem of long-term RRM in time-varying vehicular networks [15]. In [16], Liu et al. formulated a latency and reliability [17] constrained transmit power minimization problem, in which the Lyapunov stochastic optimization was leveraged to handle the network dynamics. The Lyapunov optimization can only construct an approximately optimal solution. In [18], we studied the non-cooperative RRM
in vehicular communications from an oblivious game-theoretic perspective and put forward an online reinforcement learning algorithm (RL) to approach the optimal control policy. In [19], Zheng et al. proposed a stochastic learning scheme for delay-optimal virtualized radio resource scheduling in a software-defined vehicular network. For a more practical vehicular network, the explosion in the state space makes the techniques in [18], [19] infeasible. In recent years, deep neural networks have been revisited to deal with the curse of state space explosion [20]–[22]. For example, in [23], Ye et al. devised a decentralized RRM mechanism based on deep RL (DRL) for V2V communication systems. In [24], Liang et al. proposed a fingerprint-based deep Q-network (DQN) method to solve the RRM problem in vehicular networks. However, these works do not take into account the vehicle mobility, which not only affects the channel qualities, but also provides the possibility of frequency sharing among different groups of VUE-pairs [16].

In this paper, we investigate a Manhattan grid V2V network, where the traffic information changes across the time horizon while the channel quality state depends on the geographical locations of vehicle user equipment (VUE)-transmitter (vTx) and VUE-receiver (vRx) of a VUE-pair. For such a V2V communication network, the traffic information is time-critical, and hence the fresh traffic updates are of particularly high importance [25], [26]. An emerging metric for capturing the freshness of information is the age of information (AoI) [26]. By definition, AoI is the amount of time elapsed since the most recent information update (at the destination, namely, the vRx of a VUE-pair) was generated (at the source, namely, the corresponding vTx) [27], [28]. It should be noted that optimizing AoI is totally different from queuing delay minimization as in our prior work [29]. The queuing delay can unnecessarily increase the age of an information update [30]. Motivated by the AoI research and the importance of traffic information freshness, this paper is primarily concerned with the design of an AoI-aware RRM algorithm, which achieves the optimal expected long-term performance for all VUE-pairs. The following briefly summarizes the main technical contributions of this work.

- We formulate the AoI-aware RRM problem in a Manhattan grid V2V network as a single-agent MDP with an infinite horizon discounted criterion, where the roadside unit (RSU) makes decisions regarding frequency band allocation and packet scheduling over time in order to optimize the expected long-term performance for all VUE-pairs.
- As the number of VUE-pairs in the network increases, the space of frequency band allocation and packet scheduling decision makings grows exponentially. We linearly decompose the MDP, leading to a much simplified decision making procedure. The linear decomposition approach allows the VUE-pairs to locally compute the per-VUE-pair Q-functions.
- With the consideration of vehicle mobility, the local network state space of all VUE-pairs is extremely huge. While the assumption of partial observability at each VUE-pair avoids the overwhelming overhead of information coordination among all VUE-pairs, we propose a proactive algorithm by exploring the recent advances in long short-term memory (LSTM) [31], [32] and DRL [33], using which VUE-pairs behave optimally with only partial network state observations and without any a priori statistical knowledge of the network dynamics. The proposed algorithm includes a centralized offline training at the RSU and a decentralized online testing at the VUE-pairs.
- Numerical experiments using TensorFlow [34] are carried out to verify the theoretical studies in this paper, showing that our proposed algorithm outperforms four state-of-the-art baseline algorithms.

The remainder of this paper is structured as follows. In next section, we introduce the considered model of a Manhattan grid V2V communication network and the assumptions used throughout the paper. In Section III, we formally formulate the problem of AoI-aware long-term RRM as a single-agent MDP at the RSU and discuss the technical challenges for a general solution. In Section IV, we linearly decompose the MDP and propose a proactive algorithm to solve an optimal control policy. In Section V, we present numerical experiments under various parameter settings to compare the performance of the proposed algorithm against other state-of-the-art baseline algorithms. Finally, we draw the conclusions in Section VI.

To ease readability, we list in Table I the major notations of this paper.

II. SYSTEM DESCRIPTIONS

In this section, we elaborate on the network and AoI models. Furthermore, we assume that the RSU clusters the VUE-pairs into multiple groups apart from each other.

A. Network and Channel Models

This work considers the Manhattan grid V2V communications, an illustrative example of which is depicted in Fig. 1. It has been validated that for a well defined road segment,
the vehicle density remains stable [35]. Hereinafter, we concentrate on the roads that are covered by a single RSU without loss of generality. The coverage area can be represented by a two-dimensional Euclidean space \( \mathcal{Y} \). A set \( \mathcal{K} = \{1, \cdots, K\} \) of VUE-pairs coexist in the network and share a set \( B = \{1, \cdots, B\} \) of orthogonal frequency bands. The time horizon is discretized into scheduling slots, with each being of equal time duration \( \tau \) and indexed by a positive integer \( j \in \mathbb{N}_+ \). For each VUE-pair, the vTx always follows at a fixed distance of \( \ell \) to the vRx, which moves according to a Manhattan mobility model [18], [29]. We let \( y^j_{k,(vTx)} = (y^1_{k,(vTx)}, y^2_{k,(vTx)}) \) and \( y^j_{k,(vRx)} = (y^1_{k,(vRx)}, y^2_{k,(vRx)}) \) denote, respectively, the Euclidean coordinates of geographical locations of the vTx and the vRx of a VUE-pair \( k \in \mathcal{K} \) within the duration of each scheduling slot \( j \).

Depending on whether the vTx and the vRx of a VUE-pair are in the same lane or in the perpendicular lanes, the channel model during each scheduling slot belongs to three categories, namely: 1) line-of-sight (LOS) – both the vTx and the vRx are in the same lane; 2) weak-line-of-sight (WLOS) – the vTx and the vRx are in perpendicular lanes and at least one of them is near the intersection within a distance of \( \ell/2; \) and otherwise, 3) none-line-of-sight (NLOS). More specifically, the channel state \( H^j_k = \psi \cdot h(y^j_{k,(vTx)}, y^j_{k,(vRx)}) \in \mathcal{H} \) over the frequency bands experienced by a VUE-pair \( k \in \mathcal{K} \) during each slot \( j \) includes a fast fading component \( \psi \) and a path loss \( h(y^j_{k,(vTx)}, y^j_{k,(vRx)}) \) that applies (1), which is shown at the bottom of this page, for an urban area using 5.9 GHz carrier frequency [16], [36], where we assume that \( \mathcal{H} \) is a finite space of the channel states,\(^2\) \( |·| \) is the absolute value norm of a one-dimensional vector, \( \|·\|_1 \) and \( \|·\|_2 \) are the 1-norm and the 2-norm of a vector, \( \eta \) is the path loss exponent, while both \( \varphi \) and \( \rho \) are the path loss coefficients satisfying \( \rho < \varphi \cdot \left( \ell/2 \right)^{0.5} \).

B. AoI Evolution

vTxs update time-critical traffic information to the vRx over the frequency bands on a per-scheduling slot basis. At the beginning of each scheduling slot, the RSU allocates the limited number of frequency bands to the VUE-pairs in the network. We assume that during a scheduling slot, a VUE-pair can be assigned at most one frequency band. Let \( f^j_k = (f^j_{k,b} : b \in B) \) denote the frequency allocation vector for each VUE-pair \( k \in \mathcal{K} \) at a slot \( j \) with

\[
    f^j_{k,b} = \begin{cases} 
        1, & \text{if frequency band } b \text{ is allocated to } \text{VUE-pair } k \text{ during scheduling slot } j; \\
        0, & \text{otherwise.}
    \end{cases}
\]

Let \( F^j_k = \sum_{b \in B} f^j_{k,b} \) be the frequency allocation indicator. We then have

\[
    F^j_k \in \{1, 0\}, \quad \forall k \in \mathcal{K}. \tag{3}
\]

\(^1\)The Rayleigh distribution of the fast fading component is averaged out in this work by applying the previous result in [18].

\(^2\)The assumption is reasonable due to the discrete-time Manhattan mobility model adopted in this paper.

\[
    h\left(y^j_{k,(vTx)}, y^j_{k,(vRx)}\right) = \begin{cases} 
        \varphi \cdot \left\| y^j_{k,(vTx)} - y^j_{k,(vRx)} \right\|^{-\eta}_2, & \text{for LOS} \\
        \varphi \cdot \left\| y^j_{k,(vTx)} - y^j_{k,(vRx)} \right\|^{-\eta}_1, & \text{for WLOS} \\
        \rho \cdot \left( \left| y^1_{k,(vTx)} - y^1_{k,(vRx)} \right|^2 + \left| y^2_{k,(vRx)} - y^2_{k,(vRx)} \right|^2 \right)^{-\eta}, & \text{for NLOS}
    \end{cases}
\]
A traffic information update arrives at the vTx of VUE-pair $k$ only at the beginning of each slot. We assume that the data packets pertaining to the update arrivals are independently distributed over VUE-pairs and identically distributed across scheduling slots with the mean rate of $\lambda$. For each VUE-pair, the packet arrivals have to be delivered by the end of the scheduling slot and otherwise, the packets will be dropped. Let $X_{j}^k$ denote the number of packet arrivals of VUE-pair $k$ at each scheduling slot $j$. Hence $E[X_{j}^k] = \lambda$, where the notation $E[\cdot]$ means the expectation of a random variable.

With the frequency allocation $f_{k,b}^j$ during a scheduling slot $j$ by the RSU, the power consumed by the vTx of a VUE-pair $k \in \mathcal{K}$ for the reliable transmission of scheduled $R_{k,j}$ packets to the corresponding vRx can be computed as

$$ P_{k,b}^j = \sum_{b \in B} f_{k,b}^j \cdot \frac{C_{k,b}^j + W \cdot \sigma^2}{H_{k}^j} \cdot \left( \frac{2^{\frac{\mu n_j^k}{H_{k}^j}}} - 1 \right), \quad (4) $$

where $C_{k,b}^j$ is the received aggregate interference at the vRx of VUE-pair $k$ over a band $b \in B$ from other VUE-pairs due to frequency sharing, $W$ is the bandwidth of the frequency bands, $\sigma^2$ is the power spectral density of additive white Gaussian background noise, and $\mu$ is the bit size of packet arrivals. We denote by $P_{k,\text{max}}^j$ the maximum transmit power for all vTxs, that is, $P_{k,b}^j \leq P_{k,\text{max}}^j \forall k \in \mathcal{K}$ and $\forall j$. Thus, $0 \leq R_{k,j}^i \leq \min\{X_{j}^k, R_{k,\text{max}}^j\}$, where

$$ R_{k,\text{max}}^j = \left( \left\lfloor \frac{1}{\tau} \sum_{b \in B} f_{k,b}^j \cdot W \cdot \log_2 \left( 1 + \frac{H_{k}^j \cdot P_{k,\text{max}}^j}{C_{k,b}^j + W \cdot \sigma^2} \right) \right\rfloor \frac{1}{\mu} \right), \quad (5) $$

with $\lfloor \cdot \rfloor$ meaning the floor function. The number of dropped data packets can be tracked by

$$ L_{k,j}^i = X_{j}^k - F_{k,j}^i \cdot R_{k,j}^i, \quad (6) $$

due to the deadline expiry. The V2V communications rely on fresh data packets. To this end, the data freshness is quantified by the AoI [26]. We designate $A_{k,j}^i$ as the AoI of VUE-pair $k$ up to the beginning of scheduling slot $j$, or equivalently, the end of scheduling slot $j - 1$, that is, the time elapsed since the most recently successful packet transmission. The AoI of VUE-pair $k$ evolves according to

$$ A_{k,j}^{i+1} = \begin{cases} A_{k,j}^i + \tau, & \text{if } F_{k,j}^i \cdot R_{k,j}^i > 0; \\ A_{k,j}^i, & \text{otherwise}. \end{cases} \quad (7) $$

Note that by default, we decrease the AoI of a VUE-pair in the network to be $\tau$ at the subsequent scheduling slot if there are any packets delivered during a current slot.

### C. VUE-Pair Clustering

To alleviate the mutual interference among the VUE-pairs during packet transmissions and improve the efficiency of frequency band utilization, the RSU clusters the VUE-pairs into a set $\mathcal{G} = \{1, \cdots, G\}$ of disjoint groups based on their geographical locations, where $G > 1$. The $B$ frequency bands are then exclusively allocated to the VUE-pairs in a same group and reused by the vTxs in different groups. With a

### Algorithm 1 Spectral Clustering for VUE-Pair Grouping at the Beginning of Each Scheduling Slot

1. Calculate $D^j$ and a diagonal matrix $\Omega^j = \text{diag}\{\omega_{k}^j : k \in \mathcal{K}\}$, where the $k$-th diagonal entry of $\Omega^j$ is given by $\omega_{k}^j = \sum_{k' \in \mathcal{K}} d_{k,k'}^j$.
2. Let $\Phi^j = \{\phi_{g}^j : g \in \mathcal{G}\}$, where $\phi_{g}^j$ is the eigenvector of the $g$-th smallest eigenvalue of $I - (\Omega^j)^{-1/2} \cdot D^j \cdot (\Omega^j)^{-1/2}$ with $I$ being the identity matrix.
3. Use the approach as in [38] to cluster $K$ normalized row vectors of matrix $\Phi^j$ into $G$ groups.

VUE-pair grouping pattern $\{K_{g}^j : g \in \mathcal{G}\}$, the frequency band allocation during each scheduling slot $j$ satisfies an additional constraint, namely,

$$ \sum_{k \in K_{g}^j} f_{k,b}^j \leq 1, \quad \forall b \in B, \forall g \in \mathcal{G}, \quad (8) $$

where $K_{g}^j$ denotes the set of VUE-pairs in a group $g \in \mathcal{G}$ during a scheduling slot $j$ and $\cup_{g \in \mathcal{G}} K_{g}^j = \mathcal{K}$. In this work, VUE-pair grouping is done by spectral clustering [37]. Denote by $y_k^j = \begin{pmatrix} y_k^j(1) \end{pmatrix}$ the $g$-th smallest eigenvector of $y_k^j = (y_k^j(1), y_k^j(2), \cdots, y_k^j(n_j^k))$ and $y_k^j$ as the AoI of VUE-pair $k \in \mathcal{K}$ at scheduling slot $j$. We choose a distance-based Gaussian similarity matrix $D^j = [d_{k,k'}^j]_{k,k' \in \mathcal{K}}$ to represent the geographical proximity information with each element $d_{k,k'}^j$ being given by

$$ d_{k,k'}^j = \begin{cases} \exp \left( \frac{||y_k^j - y_{k'}^j||_2^2}{\theta^2} \right), & \text{if } ||y_k^j - y_{k'}^j||_2 \leq \zeta; \\ 0, & \text{otherwise}, \end{cases} \quad (9) $$

where $\zeta$ controls the neighbourhood size and $\theta$ balances the impact of neighbourhood size. Accordingly, the VUE-pairs are clustered into groups based on $D^j$, the procedure of which is described by Algorithm 1. Now we can rewrite the transmit power consumption in (4) as

$$ P_{k,b}^j = F_{k,j}^i \cdot C + W \cdot \sigma^2 \cdot \left( \frac{2^{\frac{\mu n_j^k}{H_{k}^j}}} - 1 \right), \quad (10) $$

by approximating the received interference from other groups at the vRx of each VUE-pair $k$ over each frequency band $b \in B$ as a constant $C_{k,b}^j = C, \forall j$.

### III. Problem Formulation

In this section, we formulate the problem of AoI-aware RRM in the Manhattan grid V2V network as a discrete-time single-agent MDP with an infinite horizon discounted criterion and discuss the general solution.

#### A. AoI-Aware RRM

During each scheduling slot $j$, the local state of a VUE-pair $k \in \mathcal{K}$ can be characterized by $S_{k,j}^i = (y_k^j(1), y_k^j(2), \cdots, y_k^j(n_j^k), H_{k}^j, X_{j}^k, A_{k,j}^i) \in \mathcal{S} \equiv \mathcal{Y} \times \mathcal{H} \times \mathcal{A} \times \mathcal{X}$, which includes the
information of the geographical location \( (y^j_{k, (vTx)}, y^j_{k, (vRx)}) \), the channel state \( H^j_{k,} \), the packet arrivals \( X^j_{k,} \) and the AoI \( A^j_{k,} \).

Herein, \( X \) and \( A \) are the state sets of packet arrivals and AoI. Let \( \text{card}(\cdot) \) denote the cardinality of a set. We assume that \( \text{card}(X) \) and \( \text{card}(A) \) are finite, but can be arbitrarily large. \( S^j = (S^j, S^j_k) \in S^K \) can be used to represent the global network state with \(-k\) denoting all the other VUE-pairs in \( K \) without the presence of a VUE-pair \( k \). Let \( \pi = (\pi(F), \pi(R)) \) be a stationary control policy employed by the RSU, where \( \pi(F) \) and \( \pi(R) \) are, respectively, the frequency band allocation policy and the packet scheduling policy. When deploying \( \pi \), the RSU observes \( S^j \) at the beginning of a scheduling slot \( j \) and accordingly, makes frequency band allocation as well as packet scheduling decisions for the VUE-pairs, that is, \( \pi(S^j) = \left( \pi(F)(S^j), \pi(R)(S^j) \right) = (F^j, R^j) \), where \( F^j = (F^j_k : k \in K) \) and \( R^j = (R^j_k : k \in K) \). We next need an immediate utility function \( ^3 \) for a VUE-pair \( k \) at each slot \( j \) as below

\[
U_k(S^j, F^j_k, R^j_k) = \exp\left(-P^j_k\right) + \vartheta \cdot \exp\left(-L^j_k\right) + \xi \cdot \exp\left(-A^j_k\right), \quad (11)
\]

to assess the QoE, which is defined as a satisfaction measurement of the QoS parameters including the transmit power consumption, the packet drops as well as the AoI. Herein, \( \vartheta \) and \( \xi \) are the non-negative weighting constants.

From the assumptions on vehicle mobility, packet arrivals as well as AoI evolutions, it can be easily verified that the randomness lying in a sequence of the global network states over the time horizon \( \{S^j : j \in \mathbb{N}_+\} \) is Markovian with the following controlled state transition probability \( \mathbb{P} \)

\[
\mathbb{P}(S^{j+1} | S^j, \pi(S^j)) = \prod_{k \in K} \mathbb{P}\left( (y^{j+1}_{k, (vTx)}, y^{j+1}_{k, (vRx)}) | (y^j_{k, (vTx)}, y^j_{k, (vRx)}) \right) \cdot \mathbb{P}\left( H^{j+1}_{k,} | (y^{j+1}_{k,}, y^{j+1}_{k,}) \right) \cdot \mathbb{P}\left( X^{j+1}_{k,} \right) \cdot \mathbb{P}\left( A^{j+1}_{k,} | A^j_{k,}, F^j_k, R^j_k \right), \quad (12)
\]

where \( \mathbb{P}(\cdot) \) denotes the probability of an event occurrence. Given a stationary control policy \( \pi \) by the RSU and an initial global network state \( S^1 = S \in S^K \), we express the expected long-term discounted utility function \( V_k(S, \pi) \) of each VUE-pair \( k \in K \) as

\[
V_k(S, \pi) = (1 - \gamma) \cdot \mathbb{E}_\pi \left[ \sum_{j=1}^{\infty} (\gamma)^{j-1} \cdot U_k(S^j, F^j_k, R^j_k) | S^1 = S \right], \quad (13)
\]

where \( \gamma \in [0, 1) \) is the discount factor, \( (\gamma)^{k-1} \) denotes the discount factor to the \((k-1)\)-th power, and the expectation \( \mathbb{E}_\pi[\cdot] \) is taken over different decision makings under different global network states following a control policy \( \pi \) across the scheduling slots. When \( \gamma \) approaches 1, (13) approximates the expected long-term un-discounted utility as well \( [41] \), i.e., \( V_k(S, \pi) = \mathbb{E}_\pi[\lim_{J \rightarrow \infty} (1/J) \cdot \sum_{j=1}^{J} U_k(S^j, F^j_k, R^j_k)] | S^1 = S \). We define (13) as the optimization objective for each VUE-pair \( k \) in the network \( [42] \). Eventually, the AoI-aware RRM problem, which the RSU aims to solve, can be formally formulated as a single-agent MDP, namely, \( \forall S \in S^K \),

\[
\max \quad V(S, \pi) = \sum_{k \in K} V_k(S, \pi) = (1 - \gamma) \cdot \mathbb{E}_\pi \left[ \sum_{j=1}^{\infty} (\gamma)^{j-1} \cdot U(S^j, \pi(S^j)) | S^1 = S \right], \quad (14)
\]

s.t. constraints (3) and (8), where we define \( U(S^j, \pi(S^j)) = \sum_{k \in K} U_k(S^j, F^j_k, R^j_k) \) as the immediate utility from the viewpoint of the RSU, which is accumulated across all VUE-pairs in the network at a scheduling slot \( j \). \( V(S, \pi) \) is also termed as the state value function of a global network state \( S \) under a stationary control policy \( \pi \) \( [43] \).

B. General Solution

The stochastic optimization problem formulated as in (14) is a typical infinite-horizon discrete-time single-agent MDP with a discounted criterion. Denote the optimal stationary control policy by \( \pi^* = (\pi^*_F, \pi^*_R) \), which can be obtained through solving the Bellman’s equation \( [44] \): \( \forall S \in S^K \),

\[
V(S) = \max_{\pi(S)} \left\{ (1 - \gamma) \cdot U(S, \pi(S)) + \gamma \cdot \mathbb{E}_\pi \left[ \sum_{S' \in S^K} \mathbb{P}(S'|S, \pi(S)) \cdot V(S') \right] \right\}, \quad (15)
\]

where \( V(S) = V(S, \pi^*) \) is the optimal state value function and \( S^* = (S^*_k, S^*_k) \in S^K \) is the resulting global network state at a subsequent slot. A conventional dynamic programming solution to (15) based on the value or policy iteration \( [43] \) requires the complete knowledge of network dynamics (12), which is practically challenging. We then define the right-hand side of (15) as the Q-function

\[
Q(S, F, R) = (1 - \gamma) \cdot U(S, F, R) + \gamma \cdot \mathbb{E}_\pi \left[ \sum_{S' \in S^K} \mathbb{P}(S'|S, F, R) \cdot V(S') \right], \quad (16)
\]

where \( F = (F_k : k \in K) \) and \( R = (R_k : k \in K) \) are the decisionmakings under the current global network state \( S \). In turn, the optimal state value function \( V(S) \) can be straightforwardly obtained from

\[
V(S) = \max_{F, R} Q(S, F, R). \quad (17)
\]

By substituting (17) back into (16), we arrive at

\[
Q(S, F, R) = (1 - \gamma) \cdot U(S, F, R) + \gamma \cdot \mathbb{E}_\pi \left[ \sum_{S' \in S^K} \mathbb{P}(S'|S, F', R') \cdot \max_{F', R'} Q(S', F', R') \right]. \quad (18)
\]
where $F' = (F'_k : k \in K)$ and $R' = (R'_k : k \in K)$ denote the frequency allocation and packet scheduling decision makings under the subsequent global network state $S'$. In a decentralized decision-making procedure as will be later discussed in Section IV, the Q-learning algorithm [45], which involves the off-policy updates, is not applicable to finding the optimal per-VUE-pair Q-functions. Using the state-action-reward-state-action (SARSA) algorithm [43], [46], the RSU learns the Q-function in an iterative on-policy manner without a priori statistics of the network dynamics. With observations of the global network state $S = S_j$, the decision making $(F, R) = (F_j, R_j)$, the realized utility $U(S, F, R)$ at a current scheduling slot $j$ and the decision making $(F', R') = (F_{j+1}', R_{j+1}')$ under the global network state $S' = S_{j+1}'$ at the next scheduling slot $j + 1$, the RSU updates the Q-function according to the rule

$$Q^{j+1}(S, F, R) = (1 - \alpha^j) \cdot Q^j(S, F, R) + \alpha^j \cdot ((1 - \gamma) \cdot U(S, F, R) + \gamma \cdot Q^j(S', F', R')) , \quad (19)$$

where $\alpha^j \in [0, 1)$ is the learning rate. It has been proven that if: i) the global network state transition probability under the optimal stationary control policy $\pi^*$ is time-invariant; ii) $\sum_{j=1}^{\infty} \alpha^j$ is infinite and $\sum_{j=1}^{\infty} (\alpha^j)^2$ is finite; and iii) all global network state decision-making pairs are visited infinitely often, which can be satisfied by a $\epsilon$-greedy strategy [43], the SARSA learning process converges and finds $\pi^*$ [47]. However, two main technical challenges arise as follows:

1) under the system model being investigated in this paper, the global network state space $S^K$ faced by the RSU is extremely huge; and

2) the number $((1 + B) \cdot (1 + R_{(max)}))^K$ of decision makings at the RSU grows exponentially as $K$ increases, where for the purpose of analysis convenience, we use $R_{(max)}$ as the maximum number of packet departures at the vTxs, i.e., $R_k' \leq R_{(max)}$, $\forall k \in K$ and $\forall j$.

IV. A PROACTIVE DRL ALGORITHM

This section discusses how an optimal control policy can be approached for the RSU. An overview of our solution is shown in Fig. 2. We first decompose the single-agent MDP as in (14) into a series of per-VUE-pair MDPs, which can be solved through the decentralized SARSA algorithm. To address the partial observability and the curse of high dimensionality in local network state space faced by each VUE-pair, we then derive a proactive algorithm based on the LSTM and DRL techniques. The proactive algorithm is composed of a centralized offline training at the RSU and a decentralized online testing at the VUE-pairs.

A. Linear Q-Function Decomposition

The centralized decision making of frequency band allocation and packet scheduling at a slot from the RSU is executed

\[
\begin{align*}
Q(S, F, R) = \sum_{k \in K} Q_k(S, F_k, R_k),
\end{align*}
\]

where $Q_k(S, F_k, R_k)$ is defined to be the per-VUE-pair Q-function for each VUE-pair $k \in K$ that satisfies

\[
Q_k(S, F_k, R_k) = (1 - \gamma) \cdot U_k(S, F_k, R_k) + \gamma \cdot \sum_{S' \in S^K} P(S'|S, (F_k, F_{-k}), (R_k, R_{-k})) \cdot Q_k(S', F'_k, R'_k).
\]

We emphasize that the decision makings across the scheduling slots are performed by each VUE-pair in accordance with the optimal control policy implemented by the RSU. In other words, $(F'_k, R'_k)$ in (21), $\forall k \in K$, under the global network state $S'$ follows $\pi^*(S')$, i.e.,

\[
\pi^*(S') = \arg \max_{F''', R'''} \sum_{k \in K} Q_k(S', F''', R''').
\]

which maximizes the sum of per-VUE-pair Q-functions of all VUE-pairs in the network, where the joint $F''' = (F''_k : k \in K)$ and $R''' = (R''_k : k \in K)$ are the possible decision makings under $S'$. Particularly, we highlight two key advantages of the linear decomposition approach in the following.

1) Simplified decision makings: The linear decomposition motivates the RSU to let the VUE-pairs submit their decisions to the RSU from the local per-VUE-pair Q-function values of the frequency band allocation and packet scheduling decisions with the observation of a global network state, based on which the RSU allocates frequency bands and the VUE-pairs then schedule packet transmissions. This dramatically reduces the total number of decision makings at the RSU from $((1 + B) \cdot (1 + R_{(max)}))^K$ to $K \cdot (1 + B) \cdot (1 + R_{(max)})$.

2) Guaranteed optimal performance: Theorem 1 proves that the linear Q-function decomposition approach guarantees the optimal control policy $\pi^*$.

\textbf{Theorem 1:} The linear Q-function decomposition approach as in (20) asserts the optimal expected long-term performance for all VUE-pairs.

\textbf{Proof:} For the Q-function of a centralized decision making $(F, R)$ under a global network state $S$ in (16), we have (23), shown at the bottom of the next page, which completes the proof. □
Therefore, instead of learning the Q-function at the RSU, the SARSA updating rule (19) of the RSU is slightly adapted for each VUE-pair $k \in K$ to

$$Q_k^{j+1}(S_k, F_k, R_k) = (1 - \alpha^j) \cdot Q_k^j(S_k, F_k, R_k) + \alpha^j \cdot \left( (1 - \gamma) \cdot U_k(S_k, F_k, R_k) + \gamma \cdot Q_k^j(S_k', F_k', R_k') \right).$$  \hspace{1cm} (24)

Theorem 2 ensures the convergence of the decentralized learning process for all VUE-pairs.

**Theorem 2:** The sequence $\{(Q_k^j(S_k, F_k, R_k) : \forall k \in K) : j \in \mathbb{N}_+\}$ by (24) surely converges to the per-VUE-pair Q-functions $(Q_k(S_k, F_k, R_k) : \forall k \in K)$ if and only if for each VUE-pair $k \in K$, the $(S_k, F_k, R_k)$-pairs are visited for an infinite number of times.

**Proof:** Since the per-VUE-pair Q-functions are learned simultaneously, we consider the monolithic updates during the decentralized learning process. That is, the iterative rule in (24) can be then encapsulated as (25), which is shown at the bottom of this page. From both sides of (25), subtracting the sum of per-VUE-pair Q-functions leads to

$$\sum_{k \in K} Q_k^{j+1}(S_k, F_k, R_k) - \sum_{k \in K} Q_k(S_k, F_k, R_k) = (1 - \alpha^j) \cdot \left( \sum_{k \in K} Q_k^j(S_k, F_k, R_k) - \sum_{k \in K} Q_k(S_k, F_k, R_k) \right) + \alpha^j \cdot T^j(S, F, R), \hspace{1cm} (26)$$

where

$$T^j(S, F, R) = (1 - \gamma) \cdot \sum_{k \in K} U_k(S_k, F_k, R_k) + \gamma \cdot \max_{F_k', R_k'} \sum_{k \in K} Q_k^j(S_k', F_k', R_k') - \sum_{k \in K} Q_k(S_k, F_k, R_k))$$

and

$$\Delta^j = \left\{ (S_k^c, (F_k^c, R_k^c), U_k(S_k^c, F_k^c, R_k^c) : k \in K) : c \leq j \right\}, \left( Q_k^j(S_k, F_k, R_k) : \forall k \in K \right\).$$  \hspace{1cm} (27)

We let $\Delta^j$, which is given by (28), shown at the bottom of this page, denote the history for the first $j$ scheduling slots during the decentralized learning process. The per-VUE-pair Q-functions are $\Delta^j$-measurable, thus both $(\sum_{k \in K} Q_k^j(S_k, F_k, R_k) - \sum_{k \in K} Q_k(S_k, F_k, R_k))$ and $T^j(S, F, R)$ are $\Delta^j$-measurable. We then attain (29), shown at the bottom of this page, where $\| \cdot \|_\infty$ is the maximum norm of a vector and (a) is due to the convergence property of the standard Q-learning [45]. We are now left with verifying that $\| E[\gamma \cdot \sum_{k \in K} Q_k^j(S_k', F_k', R_k') - \max_{F_k'', R_k''} \sum_{k \in K} Q_k^j(S_k', F_k', R_k')]\|_\infty$ converges to zero, which establishes following: i) an $\epsilon$-greedy policy is deployed to trade-off exploration and exploitation when making frequency band allocation and packet scheduling decisions; ii) the per-VUE-pair Q-function values are upper bounded; and iii) both the global network state and the decision making spaces are finite. All conditions in [47, Lemma 1] are met. Thus the convergence of the decentralized learning is ensured. \hfill \Box

\begin{align*}
Q(S, F, R) &= (1 - \gamma) \cdot E_{\pi^*} \left[ \sum_{j=1}^{\infty} (\gamma)^{j-1} \cdot U(S_j, F_j, R_j) \right] \big| S_1 = S, F_1 = F, R_1 = R \\
&= (1 - \gamma) \cdot E_{\pi^*} \left[ \sum_{j=1}^{\infty} (\gamma)^{j-1} \cdot \sum_{k \in K} U_k(S_j, F_j, R_j) \right] \big| S_1 = S, F_1 = F, R_1 = R \\
&= \sum_{k \in K} (1 - \gamma) \cdot E_{\pi^*} \left[ \sum_{j=1}^{\infty} (\gamma)^{j-1} \cdot U_k(S_j, F_j, R_j) \right] \big| S_1 = S, F_1 = F, R_1 = R_k = R_k \\
&= \sum_{k \in K} Q_k^j(S_k, F_k, R_k) = (1 - \alpha^j) \cdot \sum_{k \in K} Q_k^j(S_k, F_k, R_k) + \alpha^j \cdot \left( (1 - \gamma) \cdot \sum_{k \in K} U_k(S_k, F_k, R_k) + \gamma \cdot \sum_{k \in K} Q_k^j(S_k', F_k', R_k') \right) \hspace{1cm} (25) \\
\Delta^j &= \left\{ (S_k^c, (F_k^c, R_k^c), U_k(S_k^c, F_k^c, R_k^c) : k \in K) : c \leq j \right\}, \left( Q_k^j(S_k, F_k, R_k) : \forall k \in K \right\). \hspace{1cm} (28)
\end{align*}
B. Proactive DRL for Optimal Control Policy

The linear Q-function decomposition developed in previous section brings the immediate advantage of simplified decision makings at the RSU by letting the VUE-pairs to locally learn the per-VUE-pair Q-functions. However, during the decentralized learning process as in (24), obtaining a global network state requires exhaustive information coordination between all the VUE-pairs. Note that a specific decision is made by the RSU under a specific global network state. Therefore, it is rational for each VUE-pair \( k \in K \) in a stationary network to acquire a partial observation \( O_k^j \) of the private network state \( S_k^j \) at other VUE-pairs during each scheduling slot \( j \). In this work, the partial observation \( O_k^j \) at the beginning of a current scheduling slot \( j \) includes the decision making of frequency band allocation as well as packet scheduling from the previous scheduling slot \( j-1 \). That is, \( O_k^j = (F_k^{j-1}, R_k^{j-1}) \in \mathcal{O} \), where \( \mathcal{O} = \{0, 1\} \times \{0, 1, \cdots , R_{(max)}\} \) represents a VUE-pair the set of partial observations of private network states at other VUE-pairs. It’s worth noting that \( \text{card}(\mathcal{O}) \ll \text{card}(S^K) \).

With the local observation \( O_k \in \mathcal{O} \) of \( S_k \) at a current scheduling slot, the per-VUE-pair Q-function (21) of a VUE-pair \( k \in K \) can be abstracted as

\[
Q_k(S_k, F_k, R_k) \approx Q_k(S_k, O_k, F_k, R_k), \forall(F_k, R_k),
\]

which does not need to exchange the state information with other VUE-pairs in the network. Then VUE-pair \( k \) learns the Q-function \( Q_k(S_k, O_k, F_k, R_k), \forall(S_k, O_k, F_k, R_k) \), only with local available information. We can easily find that the local partial observation space \( S \times \mathcal{O} \) is still huge. The tabular nature in representing Q-function values makes the decentralized SARSA learning in Subsection IV-A impractical. Inspired in representing Q-function values makes the decentralized SARSA learning in Subsection IV-A impractical. Inspired by the widespread success of a deep neural network [22], we adopt a double DQN to model the Q-function of a VUE-pair [33]. Simply using an independent DQN for each VUE-pair raises a new set of technical challenges:

1) the possibly asynchronous training of independent DQNs at the VUE-pairs constrains the overall network performance;
2) in practice, the limited computation capability at a VUE-pair adds another constraint on the feasibility of training a DQN locally; and
3) the accuracy of (30), which is based on partial observations, can be, in general, arbitrarily bad.

From the assumptions made throughout this paper and the definition of an identical utility function structure as in (11), there exists homogeneity in the VUE-pairs during the stochastic frequency band allocation and packet scheduling process. Training a common DQN at the RSU has the potential to be a promising alternative. Moreover, in order to overcome the partial observability of the global network states, we propose to modify the DQN architecture by leveraging recent advances in recurrent neural networks. That is, we replace the first fully-connected layer of the DQN with a recurrent LSTM layer [31], [32], resulting in a deep recurrent Q-network (DRQN) [48], [49]. Specifically, we let \( Q_k(S_k, F_k, R_k) \approx Q(\mathbf{n}_k, F_k, R_k; \theta), \forall k \in K \), where \( \mathbf{n}_k = (S_k^{j-n+1}, O_k^{j-n+1}) : n = N, N-1, \cdots , 1 \) consists of the \( N \) most recent local partial observations of the global network states up to a current scheduling slot \( j \) at VUE-pair \( k \) and is taken as the input to the LSTM layer for a more precise and proactive prediction of the current global network state \( S \), while \( \theta \) contains a vector of parameters associated with the DRQN. Considering the computation constraint of the VUE-pairs, the RSU trains the DRQN offline, the procedure of which is illustrated in Fig. 3. Instead of finding the per-VUE-pair Q-function, the parameters of the DRQN are learned.

To implement the DRQN offline training, the RSU maintains a replay memory \( M^j \) with the most recent \( M \) experiences \{\( m^j, m^{j-1}, \cdots , m^1 \)\} at the beginning of each scheduling slot \( j \), where an experience \( m^{j-m} \) (\( \forall m \in \{1, \cdots , M\} \)) is given as in (31), shown at the bottom of the next page. Meanwhile, a pool \( N^j = \{\mathbf{n}_k : k \in K\} \) of \( N \) latest partial observations is kept to predict the global network state \( S^j \) for control policy evaluation at slot \( j \). Both \( M^j \) and \( N^j \) from all VUE-pairs in the network are refreshed over the scheduling slots. The RSU first randomly samples a mini-batch \( M^j = \{M^{j_1}, \cdots , M^{j_M}\} \) of size \( M \) from \( M^j \), where each \( M^{j_m} \not\subseteq M^j \) (\( \forall m \in \{1, \cdots , M\} \)) is given by (32), shown at the bottom of the next page. Then the set \( \theta^j \) of parameters at scheduling slot \( j \) is updated by minimizing the accumulative loss function, which is defined as in (33), shown at the bottom of the next page. Where \( \theta^j \) is the set of parameters of the target DRQN at a certain previous scheduling slot before slot \( j \). The gradient is calculated as (34), shown at the bottom of the next page. Algorithm 2 summarizes the DRQN offline training.

For the online testing as in Fig. 3, each VUE-pair \( k \in K \) in the network takes \( \mathbf{n}_k \) at the beginning of each scheduling slot \( j \) as an input to the trained DRQN with parameters \( \theta \), which outputs the Q-function values \( Q(\mathbf{n}_k, F_k, R_k; \theta), \forall(F_k, R_k) \). After receiving \( (Q(\mathbf{n}_k, F_k, R_k; \theta) : (F_k, R_k)) : k \in K \), the RSU makes the decisions

\[
(F^j, R^j) = \arg \max_{F, R} \sum_{k \in K} Q(\mathbf{n}_k, F_k, R_k; \theta),
\]

as an input to the trained DRQN with parameters \( \theta \), which outputs the Q-function values \( Q(\mathbf{n}_k, F_k, R_k; \theta), \forall(F_k, R_k) \). After receiving \( (Q(\mathbf{n}_k, F_k, R_k; \theta) : (F_k, R_k)) : k \in K \), the RSU makes the decisions

\[
(F^j, R^j) = \arg \max_{F, R} \sum_{k \in K} Q(\mathbf{n}_k, F_k, R_k; \theta),
\]
that correspond to the frequency band allocation $F^j$ and the packet scheduling $R^j$ for all VUE-pairs while satisfying constraints (3) and (8).

V. NUMERICAL EXPERIMENTS

In this section, we proceed to carry out numerical experiments based on TensorFlow [34] to validate the theoretical studies for the problem of AoI-aware RRM in V2V communications and evaluate the performance achieved from our proposed proactive algorithm. During the algorithm implementations, the DRQN is first offline trained at the RSU using Algorithm 2, after which the online frequency band allocation and packet scheduling decision makings are made according to (35) across the scheduling slots.

A. General Setups

We use a Manhattan grid model as depicted in Fig. 1, which is with nine intersections in a 250 × 250 m² area [16], [18]. In the model, one road consists of two lanes, each of which is in one direction and is of width 4 m. The average vehicle speed is 60 km/h, and the data packets of an update arrival at the vTx of a VUE-pair follow a Poisson arrival process. For the DRQN, we design two fully connected layers after LSTM layer, each of the three layers containing 32 neurons.\(^6\) ReLU is selected as the activation function [50] and Adam as the optimizer [53]. Table II lists other parameter values.

For the performance comparisons, we consider the following four baseline algorithms as well.

1) Channel-Aware: At the beginning of a scheduling slot, the RSU allocates the frequency bands to the groups of VUE-pairs based on the channel state information.

\(\text{2) Packet-Aware: According to the packet arrival status at the beginning of a slot, the RSU allocates the frequency bands to the VUE-pairs of different groups.}\)

\(\text{3) AoI-Aware: The RSU allocates the frequency bands at each scheduling slot according to the AoI of VUE-pairs from different groups.}\)

\(\text{4) Random: With this baseline, the RSU first randomly allocates the frequency bands to the groups of VUE-pairs, and then schedules a random number of packets for the vTx of each VUE-pair.}\)

During the implementation of first three baselines, after the centralized frequency band allocation at the RSU at the beginning of each scheduling slot \(j\), the vTx of each VUE-pair \(k \in K\) transmits a maximum feasible number \(\min\{X^j_k, R^{j,(max)}_k\}\) of packets [27].

B. Experiment Results

1) Efficiency of Offline Training: The aim of first experiment is to examine the offline training efficiency of a common DRQN centralized at the RSU, for which the metric can be the convergence speed of the loss function (33) during the training procedure as in Algorithm 2. In the experiment,

\[\begin{align*}
\text{Table II} \\
\text{PARAMETER VALUES IN EXPERIMENTS} \\
\begin{array}{|l|l|l|}
\hline
\text{Parameter} & \text{Value} & \text{Value} \\
\hline
\psi & -68.5 \text{ dB} & \rho & -54.5 \text{ dB} \\
\eta & 1.61 & \ell_0 & 15 \text{ m} \\
\zeta & 150 \text{ m} & \theta & 30 \text{ m} \\
G & 10 & W & 800 \text{ kHz} \\
\sigma^2 & -174 \text{ dBm/Hz} & \tau & 3 \text{ ms} \\
\vartheta & 2 & \xi & 0.9 \\
\gamma & 0.9 & \mu & 2 \text{ kb} \\
F_{(max)} & 2 \text{ W} & N & 10 \\
M & 5000 & M & 200 \\
\hline
\end{array}
\end{align*}\]

\[\text{2) Packet-Aware: According to the packet arrival status at the beginning of a slot, the RSU allocates the frequency bands to the VUE-pairs of different groups.}\]

\[\text{3) AoI-Aware: The RSU allocates the frequency bands at each scheduling slot according to the AoI of VUE-pairs from different groups.}\]

\[\text{4) Random: With this baseline, the RSU first randomly allocates the frequency bands to the groups of VUE-pairs, and then schedules a random number of packets for the vTx of each VUE-pair.}\]
between the vTxA and the vRx of each VUE-pair is fixed to as B
we demonstrate the average performance per scheduling slot in
costly and needs to be done from the RSU side.

convergence speed also confirms that the offline training is
we select the number of VUE-pairs and the packet arrival rate
as \( K = 56 \) and \( \lambda = 5 \) packets/slot, respectively. The distance
between the vTxA and the vRx of each VUE-pair is fixed to be \( \ell = 50 \) m and the number of frequency bands is chosen as \( B = 5 \). We plot the variations in LOSS(\( \theta^j \)) across the
scheduling slots in Fig. 4, which tells that the offline training
procedure converges within \( 1.2 \times 10^4 \) scheduling slots. The slow
convergence speed also confirms that the offline training is
costly and needs to be done from the RSU side.

2) Performance With Changing Number of Bands: Next,
we demonstrate the average performance per scheduling slot in
terms of the average transmit power consumption, the average
packet drops, the average AoI and the average utility by
changing the number \( B \) of frequency bands. The average
performance per VUE-pair across the time horizon has been
commonly used, for example, in works [16], [21]. Other
parameters values are the same as in the previous experiment.
The results are depicted in Figs. 5, 6, 7 and 8. Fig. 5 illustrates the average transmit power consumption per VUE-pair per
scheduling slot, Fig. 6 illustrates the average packet drops per
VUE-pair per scheduling slot, Fig. 7 illustrates the average AoI per
VUE-pair across the time horizon versus \( B: K = 56, \ell = 50 \) m and \( \lambda = 5 \) packets/slot.
Fig. 8. Average utility per VUE-pair across the time horizon versus $B$: $K = 56, \ell = 50$ m and $\lambda = 5$ packets/slot.

Fig. 9. Average transmit power per VUE-pair across the time horizon versus $\ell$: $K = 68, B = 5$ and $\lambda = 4$ packets/slot.

Fig. 10. Average packet drops per VUE-pair across the time horizon versus $\ell$: $K = 68, B = 5$ and $\lambda = 4$ packets/slot.

Fig. 11. Average AoI per VUE-pair across the time horizon versus $\ell$: $K = 68, B = 5$ and $\lambda = 4$ packets/slot.

Fig. 12. Average utility per VUE-pair across the time horizon versus $\ell$: $K = 68, B = 5$ and $\lambda = 4$ packets/slot.

3) Performance Under Various VUE-Pair Distance: We then exhibit the average transmit power consumption, the average packet drops, the average AoI and the average utility per VUE-pair under different values of $\ell$ in Figs. 9, 10, 11 and 12. We configure the parameters in this experiment as: $K = 68, B = 5$ and $\lambda = 4$ packets/slot. We can see from the curves in Fig. 12 that as the VUE-pair distance increases, the VUE-pairs receive a smaller utility on average. When the distance between the vTx and the vRx of a VUE-pair increases, the channel quality becomes worse. Thus transmitting the
same number of packets requires more power consumption (as shown in Fig. 9) and limited by the transmit power at the vTxs, the frequency bands become unavailable for data transmissions more frequently across scheduling slots, resulting in increased average packet drops (as shown in Fig. 10) and average AoI (as shown in Fig. 11). Since the Packet-Aware and the AoI-Aware baseline algorithms do not account for the channel qualities, the increase of VUE-pair distance increases the possibility of allocating a frequency band to a VUE-pair with a bad channel state, under which it is impossible to deliver data packets even with the maximum transmit power. This is the reason why the realized average transmit power consumption of both algorithms first increases and then decreases.

4) Performance of Different Number of VUE-Pairs: The last experiment compares the average performance from the proposed proactive algorithm to the other four baseline algorithms versus different numbers of VUE-pairs. We assume a Manhattan grid V2V communication network in which the VUE-pair distance and the average packet arrival rate are \( \ell = 65 \text{ m} \) and \( \lambda = 6 \text{ packets/slot} \). Figs. 13, 14, 15 and 16 draw the curves of average transmit power consumption, average packet drops, average AoI and average utility per VUE-pair over the scheduling slots.

It is straightforward that more VUE-pairs means less opportunities to be allocated a frequency band for the transmission of scheduled data packets, which indicates for all algorithms, smaller average transmit power consumption (as shown in Fig. 13), more average packet drops (as shown in Fig. 14), larger average AoI (as shown in Fig. 15), and hence with the current weight value settings, worse utility performance (as shown in Fig. 16). Note that in last three experiments, the Queue-Aware and the Random baseline algorithms achieve almost the same average AoI performance. The reason behind this observation is that with the Queue-Aware baseline, the RSU allocate as with the Random baseline the frequency bands according to the randomly generated variables, which are the packet arrivals. All these experiments clearly illustrate that the proposed proactive algorithm is able to ensure better average utility performance for the VUE-pairs than the other four baseline algorithms.

VI. CONCLUSION

In this paper, we investigate the AoI-aware RRM problem for an expected long-term performance optimization in a Manhattan grid V2V communication network. The RSU allocates frequency bands and schedules data packet transmissions for all VUE-pairs according to the observations of global network states across the discrete time horizon. The decision-making process falls into the realm of a single-agent MDP. The daunting technical challenges in solving an optimal control policy prompt us to first decompose the MDP at the RSU...
into the per-VUE-pair MDPs with much simplified decision makings. Furthermore, to overcome the partial observability and the curse of high dimensionality in local network state space of a VUE-pair, we resort to the LSTM technique and the DRQN, and propose a proactive algorithm based on the DRQN. The proposed algorithm enables frequency band allocation and packet scheduling decisions with only local partial network state observations from the VUE-pairs but without a priori statistics knowledge of network dynamics. Numerical experiments show significant gains in average utility performance from the proposed proactive algorithm compared with four other state-of-the-art baselines.

REFERENCES


