Supplementary Material

Annu* Dr. Rajalakshmi Pachamuthu†

1 Proof of Theorem V-A: Packet Delivery Probability Bound

Theorem(Restated). *Consider a network of N vehicles and R available resource blocks (subchannels) per Reservation Interval (RRI). Assume that each vehicle generates one packet per RRI and that the proposed RL-based scheduling algorithm has converged to an optimal (or near-optimal) allocation. Then, if $N \leq R$, the Packet Delivery Ratio (PDR) approaches 1 (discounting channel errors). For $N > R$, the PDR is bounded as*

$$\text{PDR} \geq \frac{R}{N}\left(1 - p_e^{\max}\right), \tag{1}$$

where $p_e^{\max}$ is the maximum (worst-case) packet error probability induced by channel conditions.

Proof. We assume the following:

1. **Optimal Resource Allocation:** In steady state, the RL scheduler converges so that vehicles are allocated distinct resource blocks whenever possible. Thus, when $N \leq R$ each vehicle is assigned a unique subchannel and no collision occurs.

2. **Time-sharing under Congestion:** When $N > R$ the best possible allocation (by the pigeonhole principle) is to partition the N vehicles into $\lceil N/R \rceil$ groups such that in each RRI only one group (with approximately R vehicles) is scheduled. Hence, in steady state, each vehicle transmits once every $\frac{N}{R}$ RRIs.

3. **Channel Errors:** Independent of scheduling, each transmission is decoded successfully with probability at least $1 - p_e$, and in the worst-case $p_e \leq p_e^{\max}$.

Denote by S_i the event that vehicle i transmits successfully in a given RRI. Then, under ideal collision-free scheduling, if $N \leq R$ we have:

$$P(S_i) \geq 1 - p_e^{\max}.$$

For $N > R$, each vehicle is active in any given RRI with probability $\frac{R}{N}$. Conditioning on the vehicle actually transmitting, the success probability is at least $1 - p_e^{\max}$. By the law of total probability, the overall per-vehicle PDR is

$$P(S_i) \geq \frac{R}{N} \cdot \left(1 - p_e^{\max}\right).$$

This completes the proof. ■

2 Rigorous Proof of Theorem V-B: Collision Probability Bound

Theorem(Restated). *Assume that in steady state the vehicles implement an ϵ-greedy exploration policy in the RL-based scheduler, where ϵ_{resel} denotes the probability that an agent selects a resource randomly (i.e.,*

*Graduate Student Member, IEEE
†Senior Member, IEEE

explores) rather than using its learned (optimal) resource. Then, the steady-state collision probability P_c for a given vehicle satisfies

$$P_c \leq \frac{\epsilon_{\text{resel}}}{N} + \mathcal{O}\left(\frac{1}{T_{\text{learn}}}\right), \tag{2}$$

where T_{learn} is a measure of the convergence time of the learning algorithm, and the term $\mathcal{O}(1/T_{learn})$ accounts for transient effects during the learning phase.

Proof. In the proposed RL-based scheduling framework, we assume that once convergence is achieved the following hold:

1. **Deterministic Allocation in Steady State:** Each vehicle is assigned a unique resource block in each RRI so that if all vehicles followed their learned policy (i.e., exploited rather than explored), collisions would be eliminated.
2. **Exploration-Induced Collisions:** Collisions occur only when one or more vehicles deviate from their optimal actions due to exploration. Under an ϵ-greedy policy, each vehicle independently chooses to explore with probability ϵ_{resel}.

Fix a vehicle i that is assigned a resource r_i in the optimal allocation. For any other vehicle $j \neq i$, if vehicle j explores, then it selects a resource uniformly at random from the R available resources. The probability that j chooses r_i is therefore

$$P\big(j \text{ chooses } r_i \mid \text{exploration}\big) = \frac{1}{R}.$$

Thus, the probability that a specific vehicle $j \neq i$ causes a collision with vehicle i in a given RRI is

$$P\big(A_{j \to r_i}\big) = \epsilon_{\text{resel}} \cdot \frac{1}{R}.$$

Assuming that the exploration decisions of the $N-1$ other vehicles are independent, the probability that none of them selects resource r_i is at least

$$\prod_{j \neq i} \left(1 - \epsilon_{\text{resel}} \cdot \frac{1}{R}\right) \geq 1 - (N-1)\epsilon_{\text{resel}} \cdot \frac{1}{R},$$

where the inequality follows from the union bound (valid for small $\epsilon_{\text{resel}}/R$).

Thus, the probability that at least one other vehicle chooses resource r_i (i.e., a collision occurs) is bounded by

$$P(C_i) \leq (N-1) \cdot \frac{\epsilon_{\text{resel}}}{R}.$$

In a well-converged system, the resource allocation is nearly optimal so that the effective relation is $R \approx N$ (or, more generally, the scheduler partitions the vehicles among the R resources as evenly as possible). Under this condition, we have

$$P(C_i) \leq \frac{\epsilon_{\text{resel}}(N-1)}{N} \leq \frac{\epsilon_{\text{resel}}}{N} \cdot (1 + o(1)).$$

Finally, to account for the transient behavior during the learning phase (when the resource allocation may not be optimal), we include an additional term that decays at a rate $\mathcal{O}(1/T_{\text{learn}})$. Combining these yields

$$P_c \leq \frac{\epsilon_{\text{resel}}}{N} + \mathcal{O}\left(\frac{1}{T_{\text{learn}}}\right).$$

This completes the proof. ■

3 Convergence Analysis of Multi-Agent Reinforcement Learning

This appendix provides a rigorous justification for the convergence behavior of the proposed decentralized Q-learning algorithm when deployed across multiple vehicles in the NR V2X Mode 2 system. Each vehicle acts as an autonomous agent making resource decisions based on local observations.

3.1 Learning Setup

Let $\mathcal{N} = \{1, 2, \ldots, N\}$ denote the set of vehicles (agents), each operating in an environment with R orthogonal resource blocks. Each agent $i \in \mathcal{N}$ maintains a Q-function $Q_i(s_i, a_i)$, where s_i is the local state and $a_i \in \mathcal{A}_i$ is the resource selection action. The Q-learning update for agent i is:

$$Q_i(s_i, a_i) \leftarrow Q_i(s_i, a_i) + \alpha_t \left[r_i + \gamma \max_{a_i'} Q_i(s_i', a_i') - Q_i(s_i, a_i) \right], \tag{3}$$

where $\alpha_t \in (0, 1)$ is the learning rate, $\gamma \in [0, 1)$ is the discount factor, and r_i is the observed reward, which includes penalties for collisions and bonuses for successful transmissions.

3.2 Convergence Conditions

Let the global state space be $\mathcal{S} = \mathcal{S}_1 \times \cdots \times \mathcal{S}_N$, and action space $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_N$. The joint policy $\pi = (\pi_1, \ldots, \pi_N)$ governs the interaction among agents.

We assume the following conditions:

1. Each agent updates its Q-function asynchronously and independently.
2. The learning rate satisfies the Robbins-Monro conditions: $\sum_t \alpha_t = \infty, \quad \sum_t \alpha_t^2 < \infty$.
3. The underlying Markov Decision Process (MDP) is stationary or slowly changing with bounded rewards.
4. Exploration is maintained via an ϵ-greedy strategy with decaying $\epsilon \to 0$.

Under these conditions, the individual Q-functions converge almost surely to the optimal Q-values for a stationary environment (Watkins and Dayan, 1992). However, in our multi-agent setting, the environment is non-stationary due to mutual interference.

3.3 Potential Game Approximation

The resource allocation problem can be cast as a *congestion game*, a subclass of potential games, where each agent selects a resource and receives negative utility if the resource is shared. The game admits a potential function $\Phi : \mathcal{A} \to \mathbb{R}$ such that for any unilateral change in action by agent i,

$$\Phi(a_i', a_{-i}) - \Phi(a_i, a_{-i}) = u_i(a_i', a_{-i}) - u_i(a_i, a_{-i}), \tag{4}$$

where u_i is the utility (reward) of agent i, and a_{-i} denotes the actions of all other agents. Such games have at least one pure-strategy Nash equilibrium, and under better-reply dynamics or log-linear learning, convergence to an equilibrium is guaranteed (Monderer and Shapley, 1996).

In our RL-based scheduler, the empirical Q-values guide agents to better-performing actions. Over time, agents implicitly perform better-reply updates, and the overall learning process approximates best-response dynamics in a potential game.

3.4 Asymptotic Convergence Under Joint Policy Evolution

Let Q_i^t be the Q-function of agent i at time t, and assume all agents follow a time-varying policy π_i^t. Following Borkar and Meyn (2000), we can model the update process as a stochastic approximation:

$$Q_i^t(s, a) = Q_i^0(s, a) + \sum_{k=1}^{t} \alpha_k \left(h(Q_i^{k-1}) + M_i^k \right), \tag{5}$$

where $h(Q)$ is the expected Bellman update and M_i^k is a martingale difference noise sequence. Convergence to a locally optimal policy occurs if the induced ODE $\dot{Q}_i = h(Q_i)$ has a globally asymptotically stable equilibrium.

Theorem (Asymptotic Convergence): *Let each agent i update its Q-function using independent Q-learning with asynchronous updates, diminishing step sizes, and an ϵ-greedy exploration policy with $\epsilon \to 0$. Then, under bounded reward and finite state/action spaces, the joint policy π^t converges almost surely to a stationary point of a potential game induced by the reward structure.*

3.5 Convergence Rate and Transient Collisions

Let T_{learn} denote the number of iterations required for Q-values to converge within δ-optimality, i.e.,

$$\max_{s,a} |Q_i^t(s,a) - Q_i^*(s,a)| < \delta, \quad \forall t \geq T_{\text{learn}}.$$

Empirically, for tabular Q-learning with learning rate $\alpha_t = \frac{1}{t}$, convergence rate is sublinear: $\mathcal{O}(1/t)$. During the initial T_{learn}, exploration and uncoordinated choices may cause transient collisions. The expected collision probability during learning can be bounded as:

$$P_c(t) \leq \frac{\epsilon(t)}{N} + \mathcal{O}\left(\frac{1}{t}\right),$$

where $\epsilon(t)$ is the exploration probability at time t. As $t \to \infty$, $P_c(t) \to 0$.

3.6 Implication

The RL agents, through independent learning and bounded exploration, converge to a collision-averse resource configuration that approximates a Nash equilibrium in the induced potential game. This explains the near-zero collision probability observed in simulations beyond the learning phase and justifies the analytical bound:

$$P_c \leq \frac{\epsilon_{\text{resel}}}{N} + \mathcal{O}\left(\frac{1}{T_{\text{learn}}}\right).$$

■

4 Oracle Scheduling: Centralized Benchmark

The Oracle scheduler represents an idealized, centralized benchmark used to upper-bound performance. It assumes full global knowledge of all vehicles' SINR, location, and resource usage, and allocates resource blocks and MCS levels optimally in each Reservation Interval (RRI) to maximize throughput without collisions.

Let:

- $\mathcal{N} = \{1, \ldots, N\}$ be the set of vehicles.
- $\mathcal{R} = \{1, \ldots, R\}$ be the set of available resource blocks.
- β_i^r be the estimated SINR of vehicle i on resource r.
- θ_m be the SINR threshold required for reliable decoding with MCS level m.
- $x_{i,r} \in \{0, 1\}$ be a binary indicator if vehicle i is assigned resource r.
- $m_i \in \mathcal{M}$ be the MCS index assigned to vehicle i.

The Oracle solves the following optimization problem at each RRI:

$$\max_{x_{i,r},m_i} \quad \sum_{i=1}^{N} \log_2(1 + \mathrm{SNR}(\beta_i^{r(i)}, m_i)) \cdot x_{i,r(i)} \tag{6}$$

$$\text{s.t.} \quad \sum_{i=1}^{N} x_{i,r} \leq 1, \quad \forall r \in \mathcal{R} \quad \text{(collision-free)} \tag{7}$$

$$\beta_i^{r(i)} \geq \theta_{m_i}, \quad \forall i \in \mathcal{N} \quad \text{(decoding success)} \tag{8}$$

$$x_{i,r} \in \{0,1\}, \quad m_i \in \mathcal{M}, \quad \forall i, r. \tag{9}$$

This formulation ensures:

- No two vehicles share a resource (first constraint).
- Each vehicle is assigned the highest possible MCS that its SINR allows (second constraint).
- Maximum spectral efficiency is achieved (objective).

While this solution is not implementable in distributed Mode 2, it serves as a tight upper bound on achievable performance.