基于多智能体强化学习的车载网络频谱共享

摘要

本文研究了基于多智能体强化学习的车辆网络中的频谱共享问题，其中多个车辆对车辆（V2V）链路重用了车辆对基础设施（V2I）链路占用的频谱。高移动性车辆环境中的快速信道变化排除了在基站收集准确的瞬时信道状态信息以进行集中资源管理的可能性。作为回应，我们将资源共享建模为多智能体强化学习问题，然后使用适合分布式实现的基于指纹的深度 Q 网络方法来解决该问题。每个 V2V 链路都充当代理，共同与通信环境交互，接收独特的观察结果和共同的奖励，并通过使用获得的经验更新 Q 网络来学习改进频谱和功率分配。我们证明，通过适当的奖励设计和训练机制，多个V2V代理成功地学会了以分布式方式合作，以同时提高V2I链路的总容量和V2V链路的有效负载传输率。

关键词：车载网络，分布式频谱接入，频谱和功率分配，多智能体强化学习。

1 引言

车对万物（V2X）通信有望在道路安全、交通效率和无处不在的互联网接入等各个方面改变联网车辆和智能交通服务。最近，第三代合作伙伴项目 (3GPP)一直在寻求支持长期演进 (LTE) 和未来 5G 蜂窝网络中的 V2X 服务。电信和汽车行业成立了跨行业联盟，例如 5G 汽车协会 (5GAA)，以推动蜂窝 V2X 技术的开发、测试和部署。本文考虑了车辆网络中的频谱接入设计，通常包括车辆对基础设施（V2I）和车辆对车辆（V2V）连接。重点关注3GPP中讨论的基于蜂窝的V2X架构，其中分别通过蜂窝 (Uu) 和侧链路 (PC5) 无线电接口支持 V2I 和 V2V 连接。在版本15中，针对5G V2X增强功能提出并分析了一系列广泛的新用例和要求。例如，5G蜂窝V2X网络需要同时支持5G蜂窝V2X网络中的移动高数据速率娱乐和高级驾驶。娱乐应用程序需要与 BS（以及进一步的互联网）进行高带宽 V2I 连接，以实现视频流等功能。同时，先进的驾驶服务需要通过V2V通信定期在相邻车辆之间传播安全消息（例如，每秒10、20、50个数据包，具体取决于车辆的移动性），并且具有高可靠性。安全信息通常包括车辆位置、速度、航向等信息，以提高所有车辆对当地驾驶环境的"合作意识"。这项工作基于 3GPP 蜂窝 V2X 架构中定义的模式4，其中车辆拥有一组他们可以自主选择用于 V2V 通信的无线电资源。为了充分利用可用资源，建议此类侧链路 V2V 连接与 Uu (V2I) 链路共享频谱，并进行必要的干扰管理设计。

虽然有大量文献应用传统优化方法来解决类似的 V2X资源分配问题，但它们实际上很难在多个方面完全解决这些问题。一方面，车辆环境中快速变化的信道条件导致资源分配的巨大不确定性，例如，由于获取的信道状态信息（CSI）不准确而导致性能损失。另一方面，越来越多样化的服务要求正在被提出来支持新的 V2X 应用，例如同时最大化混合 V2X 流量

的吞吐量和可靠性，如前面的激励示例中所讨论的。有时很难以精确的数学方式对此类要求进行建模，更不用说寻找最佳解决方案的系统方法了。幸运的是，强化学习（RL）已被证明可以有效解决不确定性下的决策问题。特别是，深度强化学习最近在人类水平的视频游戏和AlphaGo中取得的成功，引发了人们对应用强化学习技术来解决各个领域的问题的浓厚兴趣，并且已经取得了显着的进展。它提供了一种稳健且有原则的方法来处理环境动态并在不确定性下执行顺序决策，从而代表了一种处理独特且具有挑战性的 V2X 动态的有前途的方法。此外，通过设计与最终目标相关的训练奖励，难以优化的目标问题也可以在强化学习框架中得到很好的解决。然后，学习算法可以自己找出一个聪明的策略来实现最终目标。使用 RL 进行资源分配的另一个潜在优势是分布式算法成为可能。因此，本论文提出并研究了使用多智能体强化学习工具来解决 V2X 频谱访问问题。

2 相关工作

为了解决车辆环境中的信道条件问题，[1]中开发了一种用于设备到设备(D2D)的启发式空间频谱重用方案,减轻了对完整 CSI 的要求。在[2]中，通过最大化V2I链路的吞吐量这一V2X资源分配方法，使模型适应缓慢变化的大规模信道衰落，从而减少了网络信令开销。在[3]中进一步采用了类似的策略，同时允许在V2I和V2V链路之间与对等V2V链路之间进行频谱共享。[4]中开发了一种用于 V2V 通信的邻近和 QoS 感知资源分配方案，该方案最小化所有 V2V 链路的总传输功率，同时使用基于 Lyapunov 的随机优化框架满足延迟和可靠性要求。利用[5]中的大规模衰落信道信息或[6]中的周期反馈CSI，最大化了V2I链路的总遍历容量，保证了V2V可靠性。[7]进一步发展了一种新的基于图的方法来处理通用的V2X资源分配问题。

除了传统的优化方法之外，[8]、[9]还开发了基于强化学习的方法来解决 V2X 网络中的资源分配问题。在[10]中，强化学习算法被应用于解决车载云中的资源供应问题，从而以最小的开销满足云中各种实体的动态资源需求和严格的服务质量要求。在[11]中，对软件定义车辆网络中传输延迟最小化的无线资源管理问题进行了研究，该问题被表述为无限视野部分观察马尔可夫决策过程（MDP），并通过在线求解基于等效贝尔曼方程和随机近似的分布式学习算法。在[12]中，提出了一种基于深度强化学习的方法，以联合管理具有以信息为中心的网络和移动边缘计算能力的虚拟化车辆网络中的网络、缓存和计算资源。所开发的基于深度强化学习的方法有效地解决了高度复杂的联合优化问题，并提高了虚拟网络运营商的总收入。在[13]中，使用强化学习算法对车载网络中的电池充电路边单元的下行链路调度进行了优化，以最大化放电期间完成的服务请求的数量，其中采用Q学习来获得最高的长期返回。该框架在[14]中得到了进一步扩展，其中提出了一种基于深度强化学习的方案，以使用端到端学习来学习具有高维连续输入的调度策略。[15]中针对具有异构基站的车辆网络开发了一种基于强化学习的分布式用户关联方法，所提出的方法利用 K-armed bandit 模型来学习网络负载平衡的初始关联，然后使用每个 BS 积累的历史关联模式直接更新解决方案。在[16]

中考虑了异构车辆网络中类似的切换控制问题，其中提出了一种基于模糊Q学习的方法来始终将用户连接到最佳网络，而不需要有关切换行为的先验知识。

3 本文方法

3.1 系统模型

考虑如图1所示的基于蜂窝的车辆通信网络，V2I 链路将每辆车连接到基站 (BS) 或 BS 型路边单元 (RSU)，而 V2V 链路则提供相邻车辆之间的直接通信。其中具有M个V2I和K个V2V链路，V2I 链路利用蜂窝 (Uu) 接口将 M 辆车辆连接到BS，以提供高数据速率服务，而 K 个 V2V 链路通过具有本地化 D2D 通信的侧链路 (PC5) 接口传播定期生成的安全消息。

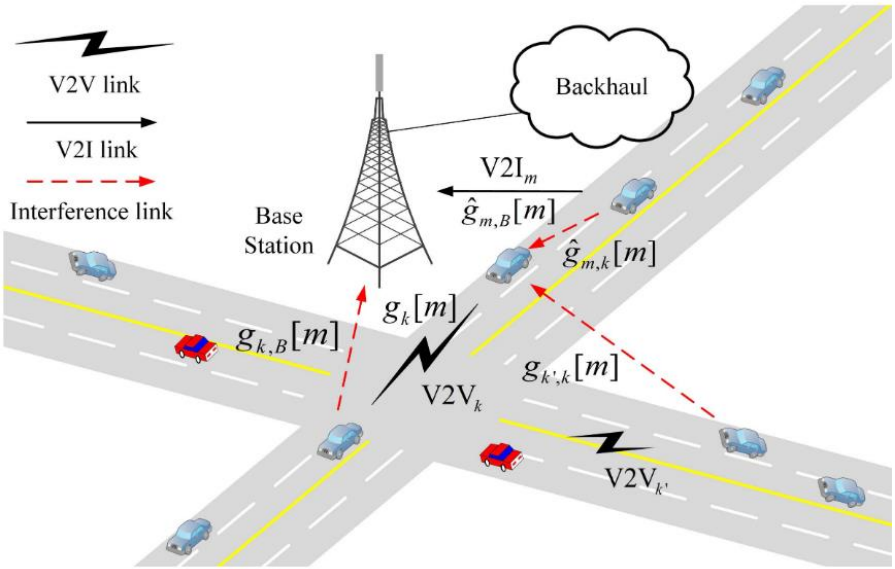


图1 车辆网络的说明性结构

假设所有收发器使用单个天线，在车辆网络中的V2I链路和V2V链路集合被分别定义为$M = \{1, \dots, M\}$，$K = \{1, \dots, K\}$。

本文重点研究蜂窝 V2X 架构中定义的模式 4，其中车辆拥有一组无线电资源，可以自主选择用于 V2V 通信。如果必要的干扰管理设计到位，此类资源池可以与蜂窝 V2I 接口的资源池重叠，以实现更好的频谱利用率。进一步假设M个V2I链路（考虑上行链路）已经被预先分配了具有固定传输功率的正交频谱子带，即，第m个V2I链路占用第m个子带。因此，主要挑战是为 V2V 链路设计有效的频谱共享方案，以便 V2I 和 V2V 链路在高移动性车辆环境的强大动态性下以最小的信令开销实现各自的目标。利用正交频分复用（OFDM）将频率选择性无线信道转换成不同子载波上的多个并行信道。几个连续的子载波被分组以形成频谱子带，并且我们假设信道衰落在一个子带内大致相同并且在不同子带之间是独立的。

在一个相干时间段内，第k个V2V链路在第m个子带(由第m个V2I链路占用)上的信道功率增益$g_k[m]$如下

$$g_k[m] = \alpha_k \hbar_k[m], \tag{1}$$

其中$h_k[m]$是与频率相关的全尺度衰落功率分量，并假设与单位均值呈指数分布。α_k捕获大规模衰落效应，包括路径损失和阴影，与频率无关。

第m个V2I链路和第k个V2V链路在第m个子带上的接收信干噪比（SINR）分别为

$$\gamma_m^c[m] = \frac{P_m^c \hat{g}_{m,B}[m]}{\sigma^2 + \sum_k \rho_k[m] P_k^d[m] g_{k,B}[m]}, \tag{2}$$

以及

$$\gamma_k^d[m] = \frac{P_k^d[m] g_k[m]}{\sigma^2 + I_k[m]}, \tag{3}$$

P_m^c和$P_k^d[m]$定义了第m个V2I发射器和第k个V2V发射器在第m个子带上的发射功率，σ^2为噪声功率，$g_{k,B}[m]$代表了第 m 个子带上第 k 个 V2V 发射机到基站的干扰信道，$\hat{g}_{m,B}[m]$代表第 m 个子带上第 m 个V2I 发射机到基站的信道，另外

$$I_k[m] = P_m^c \hat{g}_{m,k}[m] + \sum_{k' \neq k} \rho_{k'}[m] P_{k'}^d[m] g_{k',k}[m], \tag{4}$$

$I_k[m]$表示干扰功率，其中$\rho_{k'}[m]$是二进制频谱分配指标，$\rho_{k'}[m] = 1$代表第k个V2V链路使用第m个子带，否则值为0。我们假设每个V2V链路仅访问一个子带，即$\sum_m \rho_k[m] < 1$。$g_{k',k}[m]$代表第 m 个子带上第 k 个 V2V 发射机到第 k 个 V2V 接收机的干扰信道，$\hat{g}_{m,k}[m]$代表第 m 个子带上第 m 个 V2I 发射机到第 k 个 V2V 接收机的干扰信道。

然后获得第m个子带上的第m个V2I链路和第k个V2V链路的容量

$$C_m^c[m] = Wlog(1 + \gamma_m^c[m]), \tag{5}$$

以及

$$C_k^d[m] = Wlog\big(1 + \gamma_k^d[m]\big), \tag{6}$$

其中W是每个子带的带宽。

由于V2I链路目的是支持移动高数据速率娱乐服务，因此设计目标是最大化其容量，定义为$\sum_m C_m^c[m]$,方便流畅的移动宽带接入。V2V链路负责安全关键信息的可靠传播，会根据车辆的机动性以不同的频率定期生成。在数学上将这样的需求建模为在时间T内大小为B的数据包的传输速率为

$$\Pr\left\{\sum_{t=1}^{T} \sum_{m=1}^{M} \rho_k[m] C_k^d[m,t] \geq \frac{B}{\Delta T}\right\}, \ k \in \mathcal{K}, \tag{7}$$

其中B表示周期性生成的V2V有效负载的大小（以比特为单位），ΔT是信道相干时间，并且在$C_k^d[m,t]$中添加索引t以指示不同相干时隙的第k条V2V链路的容量。

因此本工作研究的资源分配问题正式表述为：设计V2V频谱分配，以$\rho_k[m]$表示，和V2V的传输功率，以$P_k^d[m]$表示，同时最大化所有V2I链路的总容量$\sum_m C_m^c[m]$和(7)中定义的V2V链路的分组传送速率。

3.2 多代理强化学习

由于车辆的高移动性妨碍了中央控制器收集完整的CSI，因此如何使V2V资源分配更加合理，协调多个V2V链路的行为是一个严峻的挑战。此外，(7)中定义的的V2V链路的数据包传送速率

涉及在时间约束T内跨多个相干时隙的顺序决策，并且由于指数复杂性而给传统优化方法带来困难。因此本文使用多智能体强化学习来开发用于V2V频谱访问的分布式算法。

在图 1 所示的资源共享场景中，多个 V2V 链路尝试访问 V2I 链路占用的有限频谱，这可以建模为多智能体RL问题。每个V2V链路都充当代理，与未知的通信环境交互以获得经验，然后指导自己的下一步决策。多个V2V智能体共同探索环境，并根据自己对环境状态的观察来调整频谱分配和功率控制策略。为了全球网络性能的利益，我们通过对所有代理使用相同的奖励，将其转变为完全合作的游戏。提出的基于多智能体强化学习的方法分为两个阶段，即学习（训练）和实施阶段。本文专注于集中学习和分布式实施的设置。意味着在学习阶段，每个 V2V 代理都可以轻松获得以系统性能为导向的奖励，然后通过更新其深度Q 网络（DQN）来调整其行为以实现最佳策略。在实施阶段，每个V2V智能体接收对环境的局部观察，然后根据其训练有素的DQN在与全尺度信道衰落相同的时间尺度上选择动作。

3.2.1 状态和观察空间

从数学上讲，该决策问题可以建模为MDP，如图2所示，在每个相关时间步长t，给定当前环境状态S_t每个V2V智能体k接受一个观测值$Z_t^{(k)}$，由观察函数O确定为$Z_t^{(k)} = O(S_t, k)$，然后选择动作$A_t^{(k)}$，多个智能体形成联合动作A_t。此后，智能体收到奖励R_{t+1}，环境以概率$p(s', r|s, a)$演化到下一个状态S_{t+1}。然后每个代理收到下一个观测值$Z_{t+1}^{(k)}$。

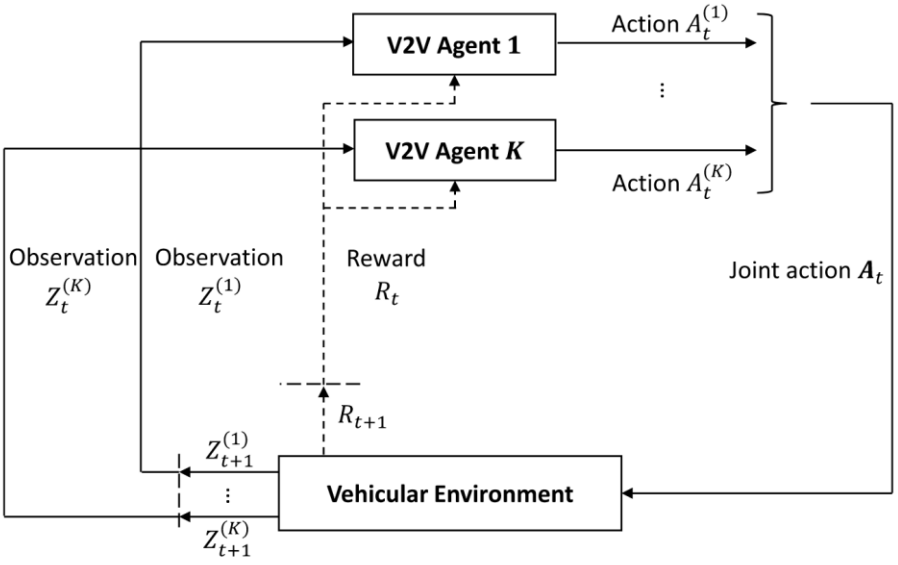


图2 车辆网络中的代理-环境交互图

代理通过观测函数了解环境，观测空间包括：剩余的V2V负载B_k，剩余时延T_k，m号子带上的接收干扰$I_k[m]$，当前agent的信道信息$G_k[m]$(除V2V到BS的干扰增益外的所有信道增益信息)，因此智能体k的观测函数总结为

$$O(S_t, k) = \{B_k, T_k, \{I_k[m]\}_{m\in M}, \{G_k[m]\}_{m\in M}\}, \tag{8}$$

其中$G_k[m] = \{g_{k',k}[m], \hat{g}_{m,k}[m], g_{k,B}[m], \hat{g}_{m,B}[m]\}$.

独立 Q 学习 [32]是解决多智能体强化学习问题最流行的方法之一，其中每个智能体根据自己的行为和观察来学习去中心化策略，并将其他智能体视为环境的一部分。然而，天真地将

DQN 与独立 Q 学习结合起来是有问题的，因为每个智能体将面临非平稳环境，而其他智能体也在学习调整其行为。随着经验重播，这个问题变得更加严重，这是 DQN 成功的关键，因为采样的经验不再反映当前的动态，从而破坏学习的稳定性。为了解决这个问题，我们采用[30]中开发的基于指纹的方法。想法是虽然一个智能体的行动价值函数随着其他智能体随着时间的推移改变其行为而变得不稳定，但它可以根据其他智能体的策略而变得静态。这意味着我们可以通过对其他代理策略的估计来扩大每个代理的观察空间，以避免非平稳性，这是hyper Q 学习的基本思想 [33]。然而，我们不希望动作值函数包含其他智能体神经网络的所有参数作为输入，因为每个智能体的策略都由高维 DQN 组成。相反，[30]中提出简单地包含一个低维指纹来跟踪其他代理的策略变化轨迹。这种方法之所以有效，是因为行动价值函数的非平稳性是由其他主体的策略随时间的变化而导致的，而不是策略本身。进一步的分析表明，每个代理的策略变化与训练迭代次数e及其探索率ϵ高度相关。因此，将它们都包含在对代理 k 的观察中，表达为

$$Z_t^{(k)} = \{O(S_t, k), e, \epsilon\} \tag{9}$$

3.2.2 动作空间

车辆链路的资源共享设计归结为V2V链路的频谱子带选择和传输功率控制。虽然频谱自然地分成 M 个不相交的子带，每个子带都由一个 V2I 链路占据。在本文中，为了便于学习和实际电路限制，将功率控制选项限制为四个级别，分别为[23,10,5,-100]dBm。选择 -100 dBm 实际上意味着 V2V 传输功率为零。因此，动作空间的维度为 4×M，每个动作对应于频谱子带和功率选择的一个特定组合。

3.2.3 奖励设计

本文的目标有两部分，一是最大化V2I容量总和，同时增加在特定时间约束T内V2V有效负载传输的成功概率。

为了实现第一个目标，简单的包括所有V2I链路的瞬时容量总和$C_m^c[m]$，为了实现第二个目标，对于每个代理k，设置奖励L_k，当负载全部交付完成前，它的值与有效的V2V传输速率相同，当能在时间t内有效的交付完所需交付负载后，奖励为常数β，该常数大于可能的最大值V2V传输速率，因此每个时隙t的V2V相关奖励设置为

$$L_k(t) = \begin{cases} \sum_{m=1}^{M} \rho_k[m] C_k^d[m,t], if\ B_k \geq 0, \\ \beta, \qquad otherwise. \end{cases} \tag{10}$$

算法的学习目标是找到一个最优策略π，该策略最大化任何初始状态s的预期回报G_t，被定义为折扣率为γ的累计折扣奖励

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}, \quad 0 \leq \gamma \leq 1. \tag{11}$$

如果折扣率γ设置成1，更大的累积奖励意味着V2V链路传输的数据量更大，直到有效载荷交付完成。因此，最大化预期的累积奖励鼓励当剩余载荷非零时，V2V链路更快的传递更多数据。

此外，学习过程还尝试获得更多的β奖励，有效的提高成功交付V2V负载几率。在实践中，β是

一个需要根据经验进行调整的超参数，在训练中，β 是一个需要根据经验进行调整的超参数。在训练中，β 被调整为大于通过运行随机资源分配的几个步骤获得的最大 V2V 传输速率，但不应该"太大"，并且理想情况下小于我们调整的最大值的两倍经验。如果出于纯粹的目标导向考虑，应该将每一步的奖励设置为0，直到V2V有效负载交付完毕后，超过该点时奖励才设置为1，然而这样的设计将阻碍学习过程，因为智能体在开始时由于只是收到0的奖励几乎学习不到任何的东西，因此我们需要将一些先验知识，即更高的V2V传输率有助于提高V2V有效负载传输率，从而提高奖励传授给每个代理。因此，提出了（10）中描述的奖励设计融合两种不同情况下的奖励数值，为此将每个时隙t的奖励设置为

$$R_{t+1} = \lambda_c \sum_m C_m^c[m,t] + \lambda_d \sum_k L_k(t), \tag{12}$$

λ_c和λ_d为平衡V2I和V2V目标的正权重。

3.2.4 学习算法

每个情境都已随机初始化的环境状态和用于传输的大小为B的完整V2V有效负载开始，并持续到T结束，小尺度信道衰落的变化会触发环境状态的改变，并导致每个单独的V2V代理改变它的动作。

我们利用深度Q学习和经验回放来训练多个V2V代理，学习频谱访问策略。Q 学习基于策略π的动作价值函数$q_\pi(s,a)$，被定义为从状态s开始，采取动作a的预期回报，并且在此后遵循策略π，表示为

$$q_\pi(s,a) = \mathbb{E}[G_t | S_t = s, A_t = a], \tag{13}$$

其中G_t在(11)定义。获得动作价值函数$q_*(s,a)$后就很容易确认最优策略。在深度Q学习中，使用以θ参数化的深度神经网络来表示动作价值函数。每个 V2V 代理都有一个专用的 DQN，将当前观测值$Z_t^{(k)}$作为输入并输出所有动作对应的价值函数。通过运行多个轮次来训练Q网络，并且在每个训练步骤中，所有V2V代理都会使用一些软策略探索状态动作空间，如ϵ-贪婪策略。由于信道的演变和所有 V2V 代理采取的操作而导致环境变化，每个代理收集并将转换元组$(Z_t^{(k)}, A_t^{(k)}, R_{t+1}, Z_{t+1}^{(k)})$存储在重播存储器中。每一轮，在重播存储器中均匀采样一小批经验并使用随机梯度下降的方法更新θ，因此也被称为经验重放，目的是最小化均方误差

$$\sum_D [R_{t+1} + \gamma \max_{a'} Q(Z_{t+1}, a'; \theta^-) - Q(Z_t, A; \theta)]^2, \tag{14}$$

其中α是目标Q网络的参数集，是从训练Q网络参数集α中定期复制，在几次更新中是固定的，算法1总结了训练过程：

算法1：多代理强化学习的资源共享算法
1：启动环境模拟器，生成车辆和连接 2：为所有代理随机生成Q-network 3：for each episode do 4：　更新车辆位置和大尺度衰落α 5：　重置$B_k = B, T_k = T$ 6：　for each step t do 7：　　for each V2V agent k do 8:　　观测$Z_t^{(k)}$ 9：　　根据ε-greedy策略根据$Z_t^{(k)}$选择$A_t^{(k)}$ 10：　end for 11：　所有agents采取动作并获得奖励R_{t+1} 12：　更新小尺度衰落信道 13：　for each V2V agent k do 14：　　观测$Z_{t+1}^{(k)}$ 15：　　在replay memory中存储$(Z_t^{(k)}, A_t^{(k)}, R_{t+1}, Z_{t+1}^{(k)})$ 16：　end for 17：　end for 18：　for each V2V agent k do 19：　　从D_k中均匀采样小批量数据 20：　　通过随机梯度下降的变体，根据(14)优化Q网络和学习目标的误差值 21：　end for 22：end for

在实现阶段的每个时间步长t，每个V2V代理估计本地信道并根据(9)中的观测变量观测环境局部观测值$Z_t^{(k)}$，其中e和ε的值为最后一个训练步骤的值。然后根据训练好的Q网络选择动作值最大的动作$A_t^{(k)}$。此后，每个V2V链路根据其选择的动作确定输出功率和频谱子带。

4 复现细节

4.1 与已有开源代码对比

本文的部分开源代码在作者的github中有展示，但下载下来并不能跑通，需要在后续对环境进行进一步配置和修改文件中的训练部分和模型部分的代码后才能运行成功。相比于源代

码，在进行复现与实验结果过程中对论文原模型数据进行了训练和比较，针对模拟结果与论文部分不吻合的问题提出了自己的改进猜想和方法。在后续的改进中，针对代码原模型中对高比特传输量的传输效率下降问题，加入了对V2V高比特传输量情况下的训练，训练后在高比特量的传输情况比训练前效果有一定提升。

4.2 实验环境搭建

本实验使用编译器为Pycharm 2023.1

所需软件包为matplotlib 3.8.2；scipy 1.11.4；tensorflow 2.15.0; numpy 1.26.2；

实验中的地图信息使用3GPP TR 36.885中规定的街区地图，但是在仿真实验中统一将地图缩小为原来的二分之一，同时初始化车辆数为4n辆(n=1,2,...)，初始每辆车默认有邻居k辆车(k≤4n-1)，即V2I通信有4n个，V2V通信有4n*k个，其余具体参数见下图：

Parameter	**Value**
Number of V2I links M	4
Number of V2V links K	4
Carrier frequency	2 GHz
Bandwidth	4 MHz
BS antenna height	25 m
BS antenna gain	8 dBi
BS receiver noise figure	5 dB
Vehicle antenna height	1.5 m
Vehicle antenna gain	3 dBi
Vehicle receiver noise figure	9 dB
Absolute vehicle speed v	36 km/h
Vehicle drop and mobility model	Urban case of A.1.2 in [3]*
V2I transmit power P^c	23 dBm
V2V transmit power P^d	[23,10,5,-100] dBm
Noise power σ^2	-114 dBm
Time constraint of V2V payload transmission T	100 ms
V2V payload size B	$[1, 2, \cdots] \times$ 1060 bytes

图3 仿真参数

Parameter	**V2I Link**	**V2V Link**
Path loss model	128.1 + 37.6$\log_{10} d$, d in km	LOS in WINNER + B1 Manhattan [36]
Shadowing distribution	Log-normal	Log-normal
Shadowing standard deviation ξ	8 dB	3 dB
Decorrelation distance	50 m	10 m
Path loss and shadowing update	A.1.4 in [3] every 100 ms	A.1.4 in [3] every 100 ms
Fast fading	Rayleigh fading	Rayleigh fading
Fast fading update	Every 1 ms	Every 1 ms

图4 V2I和V2V的链接信道模型参数

每个V2V代理的DQN由3个全连接的隐藏层组成，分别包含500、250和120个神经元。将Relu作为激活函数，RMSProp优化器用于通过学习率0.001更新网络参数。模型对每个代理的Q网络进行3000次训练，探索度ε在开始的2400个回合中从1线性退火到0.02，在之后的600个回合保持恒定，同时在训练阶段将V2V有效负载大小固定为2x1060字节。

4.3 使用说明

Environment_marl.py+main_marl_train.py+replay_memory.py用于训练多代理RL算法；

Environment_marl.py+main_sarl_train.py+replay_memory.py用于训练多代理RL算法；

Environment_marl_test.py+main_test.py+replay_memory.py用于在同一环境中测试所有模型效果；

picdraw.py+picdraw2.m用于测试结果图的绘制

4.4 创新点

针对论文原模型在高V2V负载量下的效率过低进行了进一步训练，由于时间问题，可能对模型在高V2V负载量的训练时长不够，导致训练后的模型在高V2V负载量下的表现较提升效果不大，因此未在此报告中进行展示，第五部分的内容主要是对代码进行修改后对原文的复现和对比结果。

5 实验结果分析

在后面的图中会提出四种不同的资源共享方案，方案具体如下：

(1) SARL，即基于单代理强化学习的算法[12]，其中每一时刻只有一个V2V代理更新动作，比如子带选择和功率控制，更新基于本地获取的信息和训练后的DQN，其他代理动作保持不变。所有的代理共享一个DQN网络。

(2) Random，在每个时隙内随机选取每个V2V连接的子带和传输功率。

(3) MARL，即本文提出的多代理强化学习方案

(4) Centralized maxV2V，专注于提升V2V性能而忽略V2I链路要求，在每个单独的步骤中优化V2V速率总和，穷举搜索所有V2V代理的动作空间，选择最大化V2V传输速率的动作来最大化其总和，主要是提供一个理想化的上限来对比观测出其他三种方法的实际性能情况。

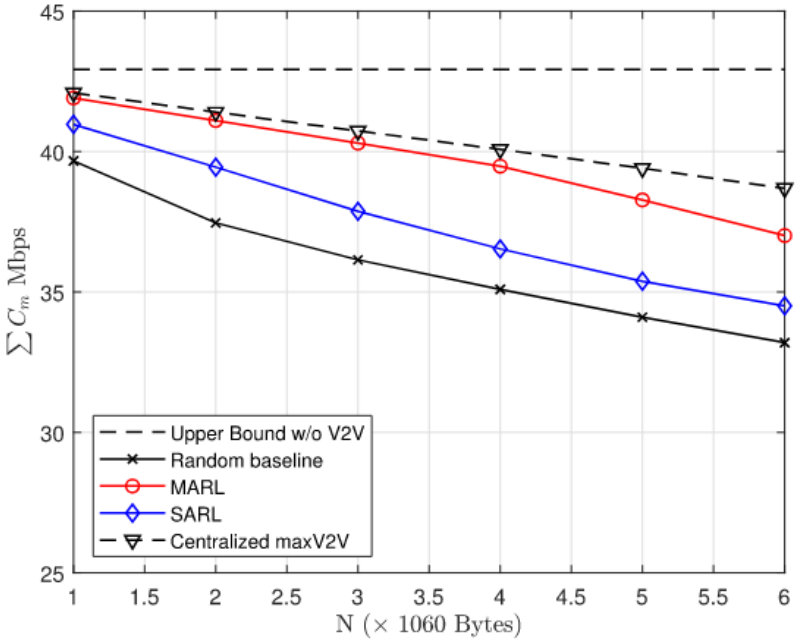


图5 V2I的最大容量随V2V负载变化图(论文)

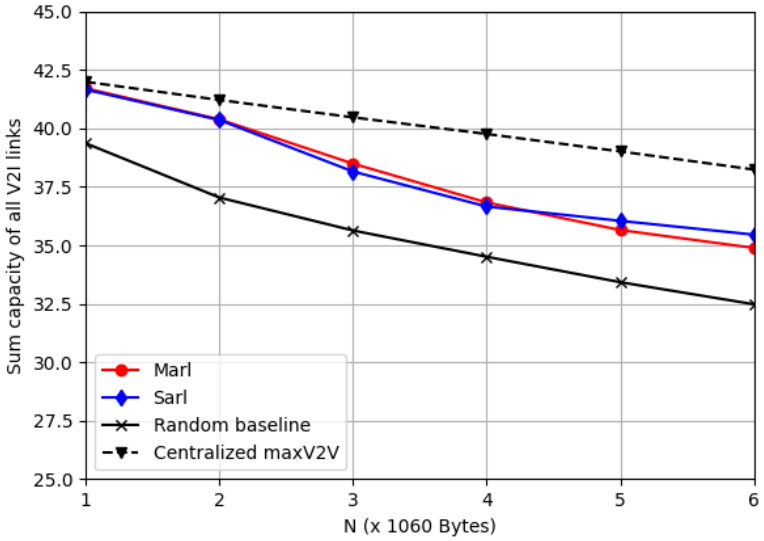


图6 V2I的最大容量随V2V负载变化图(实际)

图5图6为V2I的容量随V2V负载变化的情况，V2I容量随V2V负载增大而下降，这是因为V2V载荷增多会导致V2V的传输时间更长、发射功率更大，由此将导致对V2I链路的干扰更强，进而影响其容量。即便如此，MARL方案仍然比其他两个baseline具有更好的性能，虽然他使用2*1060进行训练，但当N增加时也便显出很好的鲁棒性。与性能上限相比，在6*1060时效果最糟。Centralized maxV2V虽然忽略V2I的需求，但是在此情况下V2I的容量性能还不错，这可能是因为集中式大大提高了V2V链路的数据传输速率，而一旦V2V传输完成，则其不会对V2I链路产生干扰。这个方案提醒我们可以进一步研究V2I和V2V链路之间的性能折中。但是我训练出来的模型的实际效果不如论文中的好，并且当V2V负载量变大时还会出现单代理算法的表现比多代理好这一情况，我的判断是V2V负载量变大时，DQN的学习效果不再适用，应该改变训练集的V2V负载大小进行训练。

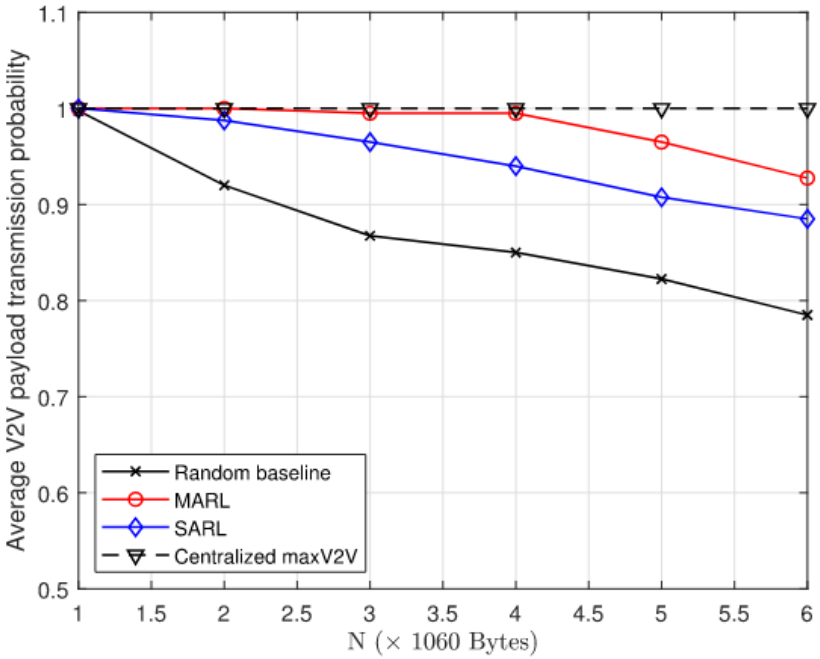


图7 不同传输负载下V2V的传输成功率(论文)

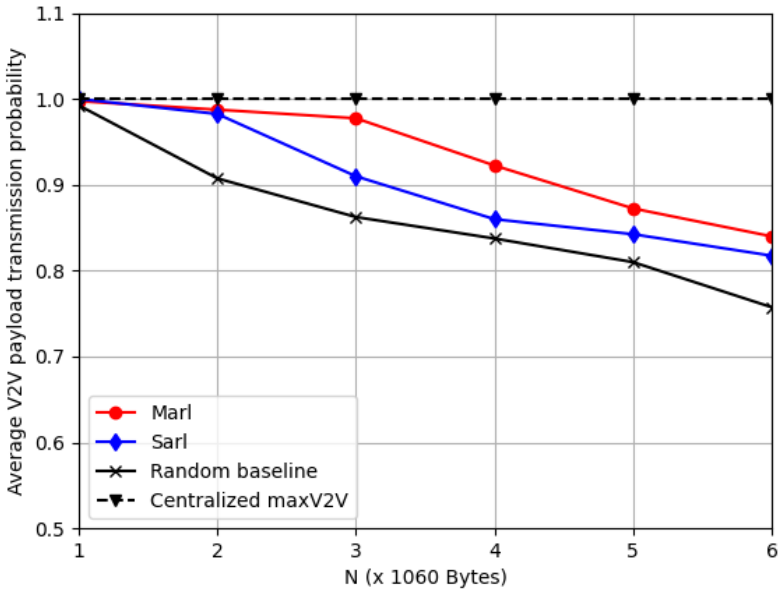


图8 不同传输负载下V2V的传输成功率(模拟)

图7图8为V2V传输成功率和传输负载的关系。论文给出图中的大致趋势是传输成功率会随着负载增大而下降，max V2V由于优先考虑V2V传输因此成功率始终为1。MARL比max V2V稍微差点但是比其他两个还是好得多。特别是在负载N为1和2时，成功率为100%；N为3和4时，性能也接近完美，在大于4后性能逐渐开始降低。我模拟出的性能趋势大致相同，即MARL的交付成功率比其他两种算法都要优秀，并且在N等于2和3时MARL的交付成功率也有较高的水平，但是当N等于4时交付成功率开始大幅度下滑，和论文的结果比差距比较大，结合图6的情况来看，推测是由于机器性能的不足或者训练出的模型与论文相比鲁棒性不足，导致负载增大时的性能下降较大。

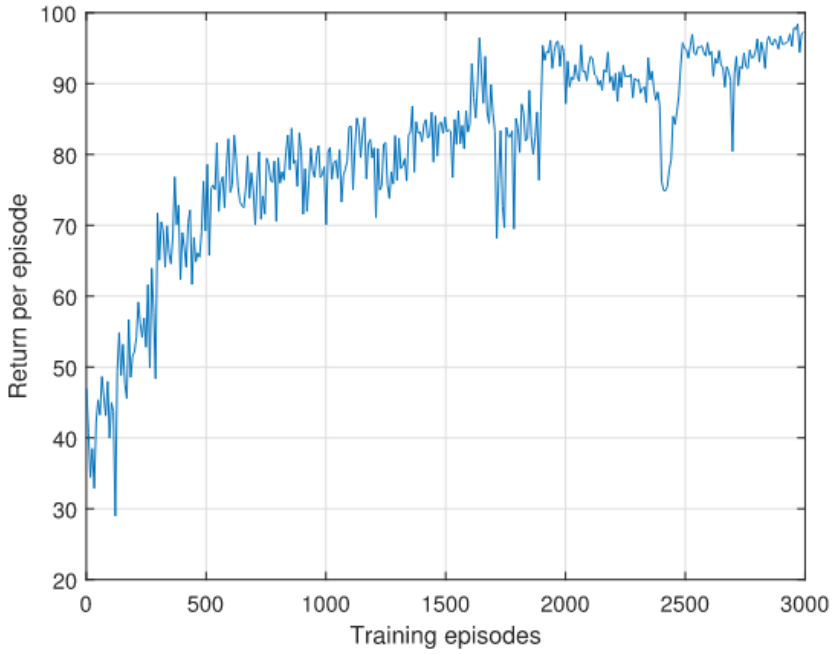


图9 回报随迭代次数的改变(论文)

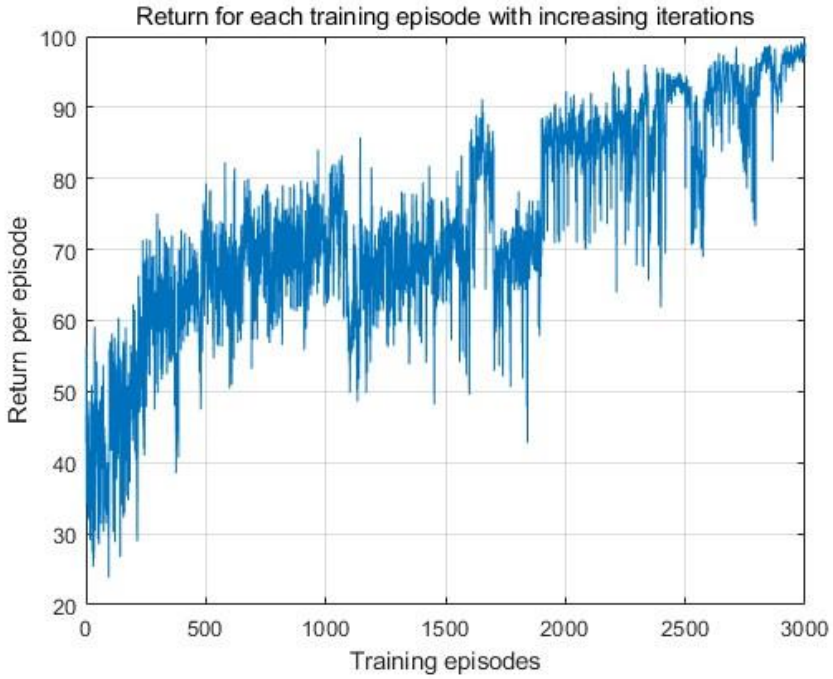


图10 回报随迭代次数的改变(模拟)

图9和图10呈现了随着迭代进行累计奖励的变化情况，从这个图可以看出本算法的收敛性。首先可以看到随着训练进行，累计奖励数值不断增加，当训练到2000时，尽管因为信道衰落仍存在一些波动，但数值开始收敛，在这一情况上模拟出的结果和论文中的结果是类似的，但奖励值上的波动会更大。

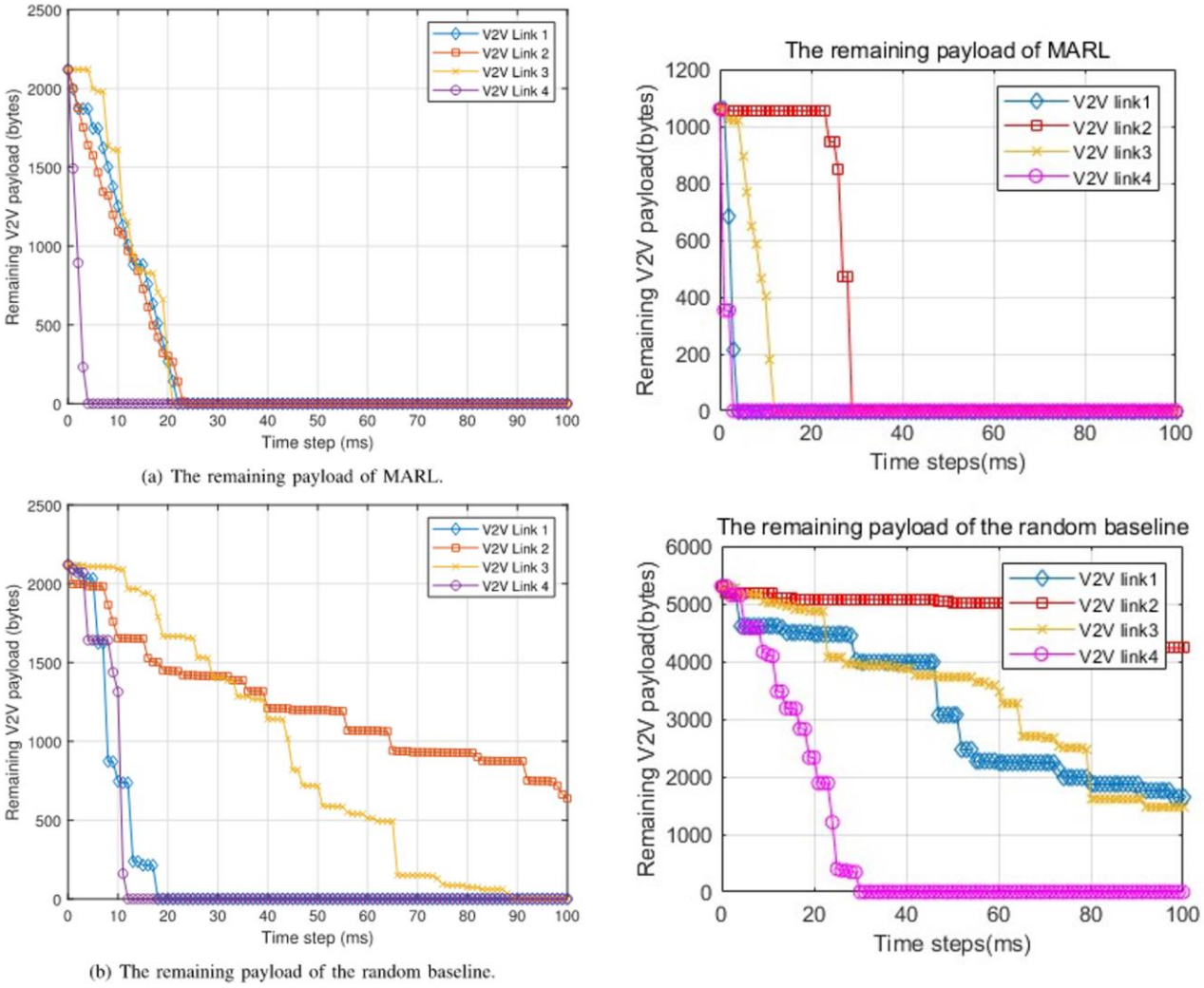
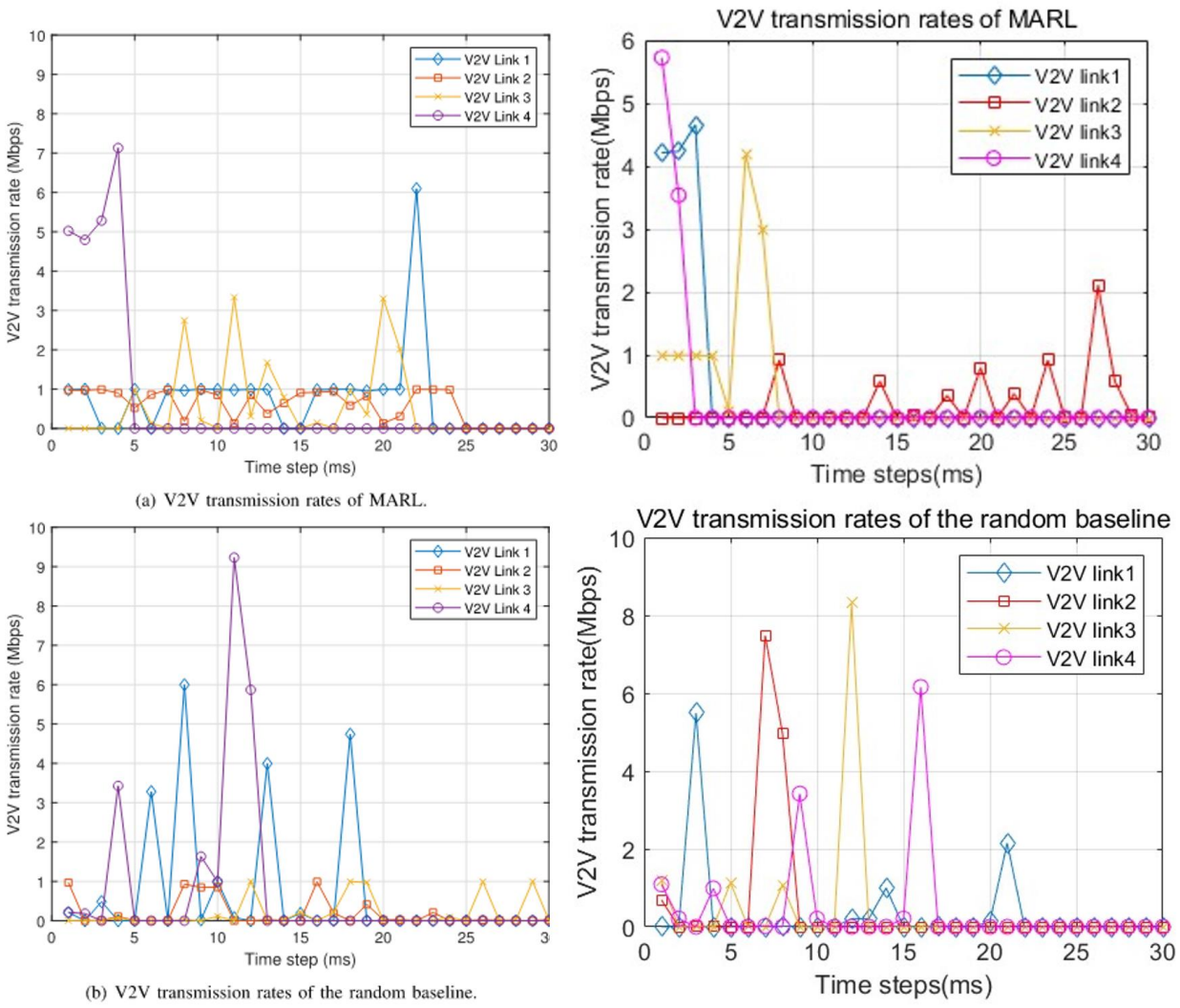


图11 在时间约束 T = 100 ms 内，初始有效负载大小B=2120字节下，MARL和Random方案的剩余 V2V 有效负载的变化(左论文右模拟)。

图11显示出规定时间和初始约束下两种方案的剩余V2V有效负载变化，其中MARL下四个link能够合理的分配自己的传输时间，通过差分传输来实现V2V所有负载的有效传输，相反随机算法中四个link都会抢着传输负载，导致最后的传输失败，这一特点不论是论文还是我的模拟结果中都体现了这一特点。



(a) V2V transmission rates of MARL.

(b) V2V transmission rates of the random baseline.

图11 在时间约束 T = 100 ms 内，初始有效负载大小B=2120字节下，MARL和Random方案的资源分配方案(左论文右模拟)。

图11显示了在时间约束内V2V链路瞬时速率的变化，在多代理RL（a图）中，链路4以很快的速率完成了传输，而链路1首先低速以确保其他链路的高速率，而观察链路2 3可见，他们貌似在进行轮流传输，以便载荷可以快速传递。而随机baseline（b图）不能为易被干扰的链路提供保护，所以其传输失败的可能性很高。MARL算法在模拟中的大致趋势与论文一致，都是通过在不同时隙中交叉调度已实现链路传输的高效，而random算法中就由于随机问题无法实现有效调度，链路的争抢调度也带来干扰的增加，这可能导致random算法中V2V链路的传输失败问题。

6 总结与展望

本报告中提出了一种基于多智能体强化学习的分布式资源共享方案，用于具有多个V2V链路的车辆网络，重用V2I链路的频谱，用于具有多个V2V链路的车辆网络，重用V2I链路的频谱，当与具有经验回放的DQN相结合时，基于指纹的Q学习可以用于解决多智能体强化学习问题中独立Q学习的非平稳问题。报告中提出的基于多智能体强化学习的方法分为集中式训练阶

段和分布式实施阶段，通过这样的机制，能够有效的鼓励V2V链路之间的合作，以提高系统级性能。未来的工作将包括对基于单智能体和多智能体强化学习算法的鲁棒性进行深入分析和比较，以便更好的了解何时应该更新Q网络以及如何有效执行此类更新，同时，也可以将提出的基于多智能体强化学习的资源分配方法扩展到车辆通信的多输入多输出和毫米波场景。

参考文献

[1] M.Botsov, M.Klügel, W.Kellerer, and P.Fertl, "Location dependent resource allocation for mobile device-to-device communications," in *Proc.IEEE WCNC*, Apr.2014, pp.1679 – 1684.

[2] W.Sun, E. G.Ström, F.Brännström, K.Sou, and Y.Sui, "Radio resource management for D2D-based V2V communication," *IEEE Trans.Veh.Technol.*, vol.65, no.8, pp.6636 – 6650, Aug.2016.

[3] W.Sun, D.Yuan, E.G.Ström, and F.Brännström, "Cluster-based radio resource management for D2D-supported safety-critical V2X communications," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2756 – 2769, Apr. 2016.

[4] M.I.Ashraf, C.-F.Liu, M.Bennis, W.Saad, and C.S.Hong, "Dynamic resource allocation for optimized latency and reliability in vehicular networks," *IEEE Access*, vol. 6, pp.63843 – 63858, Oct. 2018.

[5] L.Liang, G.Y.Li, and W.Xu, "Resource allocation for D2D-enabled vehicular communications," *IEEE Trans. Commun.*, vol. 65, no. 7, pp. 3186 – 3197, Jul. 2017.

[6] L.Liang, J.Kim, S.C.Jha, K.Sivanesan, and G.Y.Li, "Spectrum and power allocation for vehicular communications with delayed CSI feedback," *IEEE Wireless Commun. Lett.*, vol. 6, no. 4, pp. 458 – 461, Aug.2017.

[7] L.Liang, S.Xie, G.Y.Li, Z.Ding, and X.Yu, "Graph-based resource sharing in vehicular communication," *IEEE Trans.Wireless Commun.*, vol. 17, no. 7, pp. 4579 – 4592, Jul.2018.

[8] H.Ye ,L.Liang, G.Y.Li, J.Kim, L.Lu, and M.Wu, "Machine learning for vehicular networks: Recent advances and application examples," I*EEE Veh. Technol. Mag.*, vol. 13, no. 2, pp. 94 – 101, Jun. 2018.

[9] L.Liang, H.Ye, and G.Y.Li, "Toward intelligent vehicular networks: A machine learning framework," *IEEE Internet Things J.*, vol. 6, no. 1, pp. 124 – 135, Feb. 2019.

[10] M.A.Salahuddin, A.Al-Fuqaha, and M.Guizani, "Reinforcement learning for resource provisioning in the vehicular cloud," *IEEE Wireless Commun.*, vol. 23, no. 4, pp. 128 – 135, Aug.2016.

[11] Q.Zheng, K.Zheng, H.Zhang, and V.C.M.Leung, "Delay-optimal virtualized radio resource scheduling in software-defined vehicular networks via stochastic learning," *IEEE Trans. Veh. Technol.*, vol. 65, no. 10, pp. 7857 – 7867, Oct.2016.

[12] Y.He, N.Zhao, and H.Yin, "Integrated networking, caching, and computing for connected vehicles: A deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 67, no. 1, pp. 44 – 55, Jan. 2018.

[13] R.F.Atallah, C.M.Assi, and J.Y.Yu, "A reinforcement learning technique for optimizing downlink scheduling in an energy-limited vehicular network," *IEEE Trans. Veh. Technol.*, vol. 66, no. 6, pp. 4592 – 4601, Jun.2017.

[14] R.Atallah, C.Assi, and M.Khabbaz, "Deep reinforcement learningbased scheduling for roadside communication networks," *in Proc. 15th Int. Symp. Modeling Optim. Mobile, Ad Hoc, Wireless Netw. (WiOpt)*, May 2017, pp. 1 – 8.

[15] Z.Li, C.Wang, and C.-J.Jiang, "User association for load balancing in vehicular networks: An online reinforcement learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 8, pp. 2217 – 2228, Aug. 2017.

[16] Y.Xu, L.Li, B.-H.Soong, and C. Li, "Fuzzy Q-learning based vertical handoff control

for vehicular heterogeneous wireless network,” in *Proc. IEEE ICC*, Jun. 2014, pp. 5653 – 5658.

[17]L. Liang, H. Ye and G. Y. Li, "Multi-Agent Reinforcement Learning for Spectrum Sharing in Vehicular Networks," *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 1-5, 2019