Advanced Data Analysis in Python Course Project

Name:
Nasir Khan (0075244)

Deep Reinforcement Learning based Wireless Resource Allocation in Vehicular Networks

1. Objective:

In this project, the aim is to utilize the knowledge of machine learning tools and in specific the neural networks and reinforcement learning to allocate resources optimally in a vehicular communication network.

The broader goal of the project is to build upon and utilize the knowledge gained in the course to replicate the results of a paper [1]. The paper utilize deep neural networks for efficient resource allocation. We have shown that the same results as in the original paper can be obtained with reduced number of training episodes (we use 2000 training episodes while the original paper utilize 3000 episodes for the training phase).

2. Motivation:

Since reinforcement learning is used for resource allocation which is a semi-supervised learning approach. Therefore, the agent learns from the environment (vehicular network environment) itself based on the reward/penalty. As a result, we can get a stable and faster convergence and improve learning behavior compared to purely supervised machine learning approaches which require tremendous amount of data for training and deployment (testing) phase.

Note: *We utilize python as the programming language and use Numpy, Matplotlib, Keras (tensorflow) and math libraries for implementation and generation of our results. The skills attained (Numpy and math libraries use) through different course homeworks and in-class exercises helped us a lot to implement and present the results in our project.*

3. Introduction:

Vehicle-to-everything or Internet of Vehicles (IoV) communication is a promising technology for future wireless networks that could help to facilitate future transportation systems. It also applies the mobile communication technology to realize the inter-communication and coordination among vehicle-to everything (V2X), such as vehicle-to-infrastructure (V2I) communications and vehicle-to-vehicle (V2V) communications, which can greatly improve the traffic safety and efficiency, and reduce the unfortunate incidents and congestion of the road traffic.

As high mobility of vehicles causes rapidly changing in wireless channels, a full channel state information (CSI) assumption can no longer be applied in the V2V networks. Hence, we address this challenging issue of reliable transmission by formulating the resource allocation problem as a Markov Decision Process (MDP). The conventional Q-learning cannot be applied for this problem as the state-action space is large huge, therefore, an alternative is to combine the perks of deep neural networks (DNN) and Q-learning to find the optimal Q-values for each action by utilizing a deep Q network (DQN).

We utilize deep reinforcement learning for mapping the local observations of each vehicle in the V2V network, to control transmission power and allocate spectrum to maximize the system throughput with constraints on minimum signal-to-interference-plus-noise ratio (SINR) for both V2I and the V2V links. Hence, the main purpose is using deep reinforcement learning to develop a decentralized resource allocation mechanism for V2V communications.

4. Problem Formulation:

For calculating the throughput, we should model a Vehicular environment. As shown in Fig. 1, the vehicular network includes with M cellular users (CUEs) demanding V2I links and they are orthogonally allocated spectrum bands with high capacity communication links denoted by $M = \{1, 2, .., M\}$. Additionally, K pairs of V2V users (VUEs) which need V2V links to share information for traffic safety are considered in the problem formulation denoted by $\kappa = \{1, 2, .., K\}$. In this scenario, the environment is considered to be everything outside the V2V link. It should be noted that the actions incorporated while considering the RL based formulation, such as selected spectrum, transmission power, etc., are treated as a part of the environment.

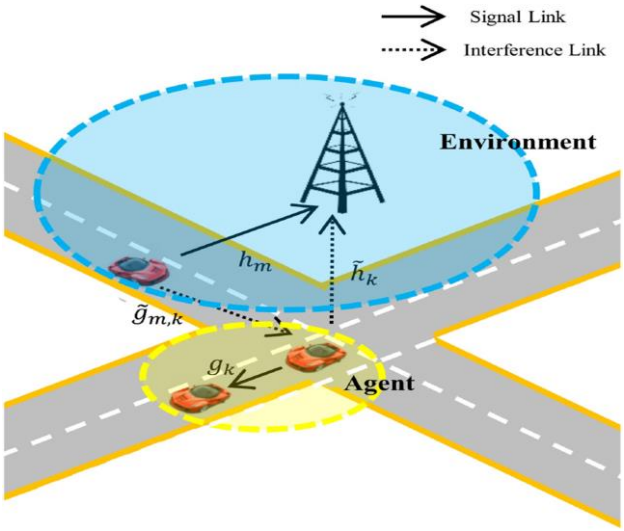


Fig. 1. An illustrative structure of vehicular communication networks.

For the V2I users, the capacity of the m^{th} CUE for calculating throughput is

$$C^{V2I}[m] = W.\log\left(1+\gamma^{V2I}[m]\right),$$

Where W is the bandwidth and $\gamma^{V2I}[m]$ is the SINR of the m^{th} CUE which can be expressed as

$$\gamma^{V2I}[m] = \frac{P_m^{V2I} h_m}{\sigma^2 + \sum_{k\in\kappa} \rho_k[m] P_k^{V2V} \tilde{h}_k},$$

Where P_m^{V2I} and P_k^{V2V} are the transmission powers of the m^{th} CUE and the k^{th} VUE, respectively, σ^2 is the noise power, h_m is the power gain of the channel corresponding to the m^{th} CUE, Hence the capacity of the CUE is m^{th}, $\tilde{h}_k$ is the interference power gain of the k^{th} VUE, $\rho_k[m]$ is the spectrum allocation indicator with $\rho_k[m] = 1$ if the k^{th} VUE reuses the spectrum of the m^{th} CUE and $\rho_k[m] = 0$ otherwise.

For V2V users, the capacity of the k^{th} VUE can be expressed as

$$C^{V2V}[m] = W.\log\left(1+\gamma^{V2V}[m]\right)$$

Where W is the bandwidth and $\gamma^{V2V}[m]$ is the SINR of the k^{th} VUE which can be expressed as

$$\gamma^{V2V} = \frac{P_k^{V2V} g_k}{\sigma^2 + G_c + G_d}$$

Where $G_c = \sum_{m \in M} \rho_k [m] P_m^{V2I} \tilde{g}_{m,k}$ is the interference power of the V2I link and,

$$G_d = \sum_{m \in M} \sum_{\substack{k' \in \kappa \\ k \neq k'}} \rho_k [m] \rho_{k'} [m] P_{k'}^{V2V} \tilde{g}_{k',k}^{V2V}$$

is the overall interference power from all V2V links. g_k is the power gain of the k^{th} VUE, $\tilde{g}_{m,k}$ is the interference power gain of the m^{th} CUE, and $\tilde{g}_{k',k}^{V2V}$ is the interference power gain of the k'^{th} VUE.

a. Problem Type:

Both the power level and channel are discrete, so it is Episodic

- The transmission power is discretized into four levels and, the dimension of the action space is 4×NRB when there are NRB resource blocks in all. The channels are discrete is a sense that the number of sub channels are equal to the number of vehicle-to-infrastructure (V2I) links.
- In the frequency and power allocation in vehicle-to-vehicle (V2V) communications, the resource blocks are considered. The resource blocks (RBs) are the spectrum assignment (frequency allocation) to every V2V links based on the transmission power. The number of resource blocks according the problem formulation are equal to the number of vehicles considered within the network.

b. State Space and/or Features:

As an agent, the V2V link observes a state, s_t, from the state space S, and accordingly takes an action, a_t, from the action space, A, selecting sub-band and transmission power based on the policy, π for each time step t. The decision policy, π, is determined by a Q-function, Q(s_t ; a_t ; ϕ), where ϕ is the parameter of the Q-function and can be obtained by deep RL. In our system, the state observed by each V2V link for characterizing the environment consists of several parts:

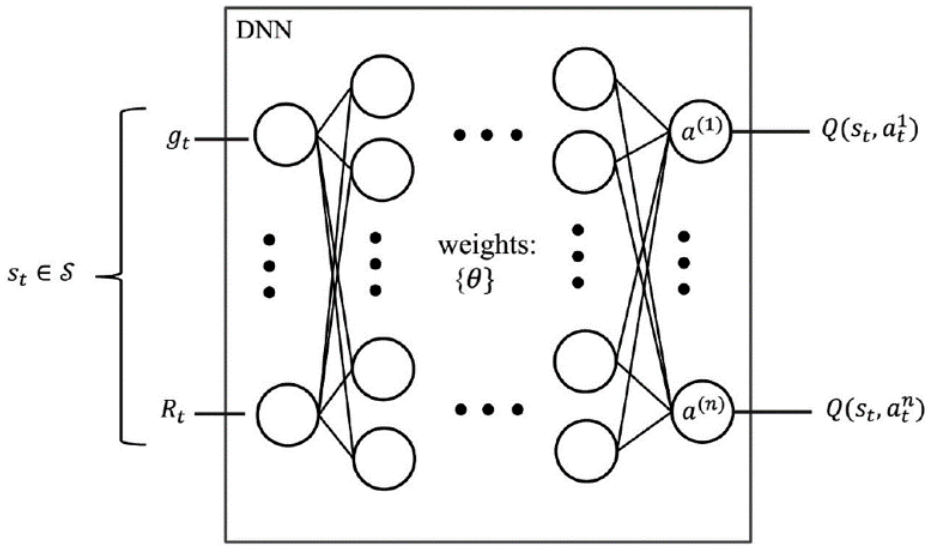
Local information of V2V and V2I links characterize environment state which includes the instant channel information of the corresponding V2V link, G_k, the interference power to the link, I_k, the remaining payload of the V2V user to transmit the packets, B_k, and the remaining time for packet delivery, T_k.

Hence, the state space can be expressed as **S = { G_k , I_k , B_k , T_k }**

c. Actions:

We consider discrete actions $a_t \in A$, which includes selecting a sub-channel and a power level for transmission, according to the current state, $s_t \in S$, based on the decision policy π for each time step. The number of power control levels, **P = [23, 10, 5, −100] dBm** where, -100 dBm effectively means no communication between the vehicle pair (i,j).

As the image below, our DQN takes the inputs as the states and outputs the Q-values. The inputs equal the number of states and the outputs (Q-values) equal the number of discrete power levels * resource blocks ($3 \times N_{RB}$). Action corresponds to one particular combination of a spectrum sub-band and power selection for each vehicle pair (i,j). We can extract the actions by taking the argmax of the Q-values (argmax of the output layer of DQN) to get the actions.



d. Reward Function:

Designed reward function must correlate with the desired objective. Reward function in our case is consists of two parts, the capacity of V2I links (C_m^{V2I}), the capacity of V2V links. Reward for the V2V link (agent) is defined as the weighted sum of V2I channel capacity and V2V transmission rate. Therefore, the reward function can be expressed as,

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

Where, λ_1 and λ_2 are the weighting coefficients treated as hyper parameters.

5. Related Work:

In this project, we consider the optimal resource allocation mechanism for vehicle-to-vehicle (V2V) communications based on deep reinforcement learning which is inspired form the work presented in [1].

The problem formulated in [1] is for a Single-Agent RL and for the scope of this project, we will consider a single-agent RL based approach to solve the problem of frequency and power allocation in V2V communications. However, as part of literature review, we have also considered the works for multi-Agent RL based approaches for resource allocation in V2V communication networks. The main challenge in multiagent RL is that simultaneous actions of all learning agents tend to make the environment observed by each agent (vehicle) highly non-stationary and compromises stability of DQN training.

In [2], a distributed resource sharing scheme based on multi-agent RL for vehicular networks with multiple V2V links is considered.

In [3], RL based scheduling mechanism is proposed for an energy-limited vehicular network. A reinforcement learning technique, i.e., protocol for energy-efficient adaptive scheduling using reinforcement learning (PEARL), is proposed for the purpose of optimizing the downlink traffic scheduling during a discharge period of the battery. The objective is to equip the base station with the required artificial intelligence to exploit the optimal scheduling policy for guaranteed performance in terms of completing service requests and operations of vehicular networks.

In [4], a reinforcement learning-based resources allocation algorithm is proposed which considers optimizing the uplink/downlink ratio by utilizing Q-learning where the base station (BS) is responsible for selecting the optimal resource control policy in each action policy interval. The goal of the agent (BS) is to maximize the expected total reward in the future by updating Q elements until convergence.

For the scope of our project and for comparison of our results, we compare the evaluation metrics with a random baseline which chooses the spectrum sub-band and transmission power for each V2V link in a random fashion at each time step.

6. Approach and Implementation Details:

For simulating the system model, we consider a single cell system with the carrier frequency of 2 GHz. We follow the simulation setup for the Manhattan case detailed in 3GPP TR 36.885. with both line-of-sight (LOS) and non-line-of-sight (NLOS) channels. We custom built our environment following the evaluation methodology for the urban case defined in Annex A of 3GPP TR 36.885 (Release 14) [5].

The Manhattan mobility model is usually used to emulate the movement pattern of mobile nodes on streets defined by maps. The map is composed of a number of horizontal and vertical streets. Each street has two lanes for each direction (north and south direction for vertical streets, east and west for horizontal streets). At an intersection of a horizontal and a vertical street, the mobile node can turn left, right or go straight. This choice is probabilistic: the probability of moving on the same street is 0.5, the probability of turning left is 0.25 and the probability of turning right is 0.25.

We considered an episodic setting with each episode spanning the V2V payload delivery time constraint. Each episode start with random initialization of the environment state. The single-agent RL based algorithm is used where we utilize a deep neural network to estimate the Q-values and updates action, i.e., spectrum sub-band and transmit power selection, based on locally acquired information. We train a single DQN while others agents' actions remain unchanged. After training, a single DQN is shared across all V2V agents.

We performed the simulation in two phases namely the Learning Phase and the Distributed Implementation Phase. In learning phase we start the environment simulator, generate vehicles and links and the environment transition due to channel evolution and actions taken by the agent. Reward id given based on capacity of the V2V link in updated vehicle location. We train the DQN based on (s,a,r,s') tuple. We use batch learning and experience replay by storing the tuples in buffer memory and update the test network every (4* time steps= 400) steps. In the implementation phase the agent observe environment state (estimate local vehicle state) and then selects the maximum action value according to trained Q-network. All V2V links then start to transmit with the power level and frequency sub-band determined by their selected action.

7. Evaluation Details, Results and Analysis/Discussion:

To evaluate the performance of our approach for jointly optimizing the frequency and power control in the considered vehicular network for the single-agent RL problem, we consider a deep RL based optimization framework to maximize the total reward.

System model and our simulation parameters are shown in Table 1 and the deep neural network related parameters are shown in Table 2:

Parameter	Value
Carrier frequency	2 GHz
Bandwidth per channel	1.5 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
Vehicle speed	36 km/h
Neighbor distance threshold	150 m
Number of lanes	3 in each direction (12 in total)

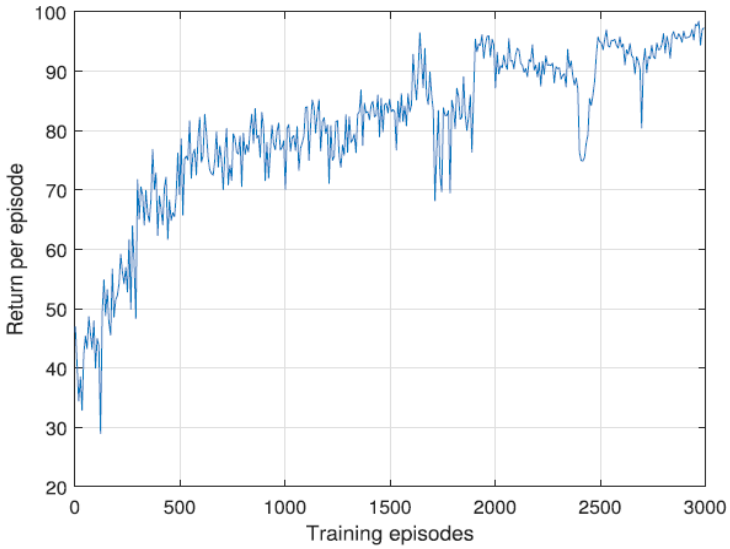
Table 1 Simulation parameters

Deep Neural Network Parameters	
Training episodes	2000
Test episodes	100
Time steps/episode	$\frac{Slow\ time}{fast\ time} = \frac{100ms}{1ms} = 100$
Hidden Layers	3 with 500, 250, and 120 neurons
Activation function	RELU
Optimizer	ADAM
Learning rate	0.001 (worked well for us!)

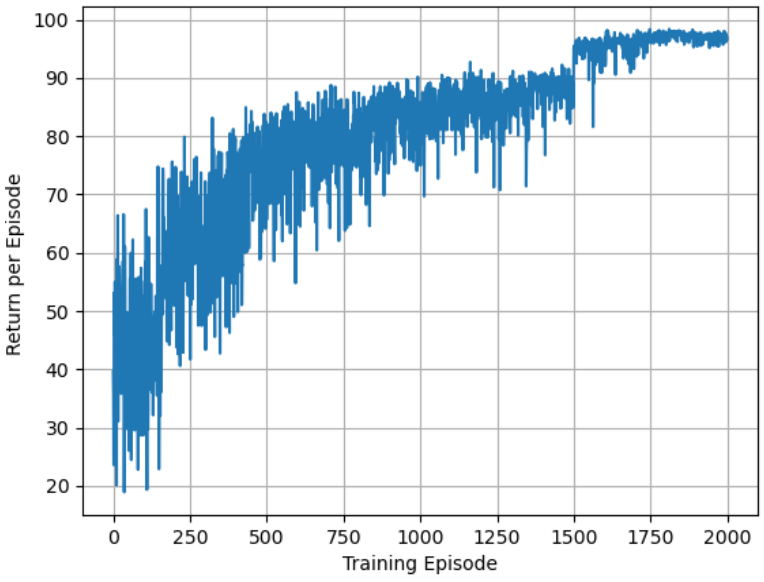
Table 2 DQN related parameters used for simulations

For generating rewards we consider 2000 episodes in training phase and fix the payload size to 2 x 8480 Bytes. Whereas, in the implementation phase for testing our approach we consider 100 episodes. We consider the cumulative rewards per training episode with increasing training iterations to validate that our agent is learning and converges to a stable behavior. We plot the returns of the agent for 2000 episodes as shown in the figure. We compare the results with the original paper results which use 3000 episodes for the training phase. From the figure we see that we get a fair convergence of the returns in around 1500 episodes. The results are shown in the figure below:

Paper Results

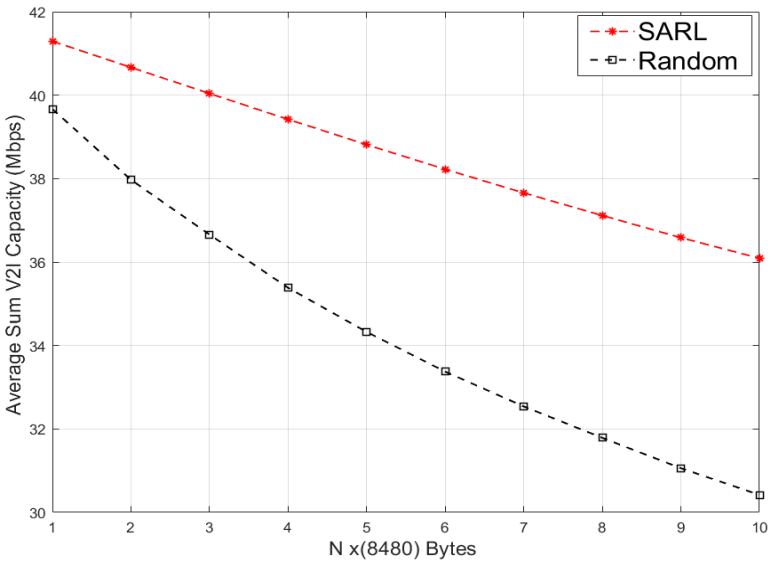


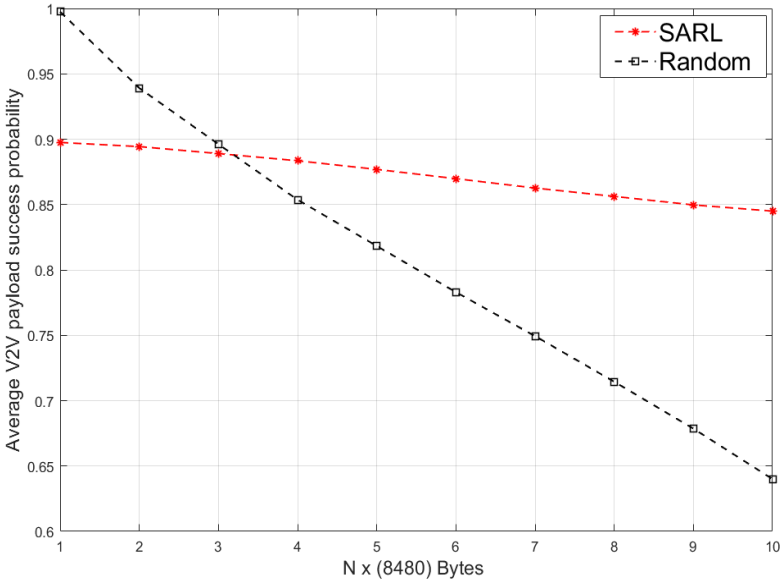
Project Results



We also compare the average sum capacity for the V2I links as a function of the payload size. The performance drops for all schemes with growing V2V payload sizes. An increase of V2V payload leads to longer V2V transmission duration and possibly higher V2V transmit power which results in stronger interference. However, compared to the random baseline our RL based strategy performs significantly well in maintaining a higher sum capacity for increased payload sizes.

Additionally, we compare the average success probability in delivering the packets within time constraint as a function of payload size. As the V2V payload size grows larger, the transmission success probabilities drop rapidly for random baseline but we get almost consistent above 85% success probability for RL based approach. The results are shown in figure below.





8. Conclusion:

We considered a single-agent RL based algorithm SARL for resource allocation in vehicular networks. We showed that with proper reward design, we can get a stable and faster convergence and improve learning behavior. We further noticed that to maximize the primary objective, our reward must correlate with the primary objective to encourage learning. Due to time limitation, we did not consider multi-agent based approach where each V2V link would have a dedicated DQN. This scenario is more practical as individuals vehicles usually have no information about neighbors and need to be trained individually for behaving as an agent.

Reference:

[1] H. Ye, G. Y. Li and B. F. Juang, "Deep Reinforcement Learning Based Resource Allocation for V2V Communications," in IEEE Transactions on Vehicular Technology, vol. 68, no. 4, pp. 3163-3173, April 2019, doi: 10.1109/TVT.2019.2897134.

[2] L. Liang, H. Ye and G. Y. Li, "Spectrum Sharing in Vehicular Networks Based on Multi-Agent Reinforcement Learning," in IEEE Journal on Selected Areas in Communications, vol. 37, no. 10, pp. 2282-2292, Oct. 2019, doi: 10.1109/JSAC.2019.2933962.

[3] R. F. Atallah, C. M. Assi and J. Y. Yu, "A Reinforcement Learning Technique for Optimizing Downlink Scheduling in an Energy-Limited Vehicular Network," in IEEE Transactions on Vehicular Technology, vol. 66, no. 6, pp. 4592-4601, June 2017, doi: 10.1109/TVT.2016.2622180.

[4] Y. Zhou, F. Tang, Y. Kawamoto and N. Kato, "Reinforcement Learning-Based Radio Resource Control in 5G Vehicular Network," in IEEE Wireless Communications Letters, vol. 9, no. 5, pp. 611- 614, May 2020, doi: 10.1109/LWC.2019.2962409.

[5] Technical Specification Group Radio Access Network; Study LTE-Based V2X Services; (Release 14), document 3GPP TR 36.885 V14.0.0, 3rd Generation Partnership Project, Jun. 2016