APPENDIX FOR THE PAPER "MAKE SMART DECISIONS FASTER: DECIDING D2D RESOURCE ALLOCATION VIA STACKELBERG GAME GUIDED MULTI-AGENT DEEP REINFORCEMENT LEARNING"

A. Proof of Proposition 1

Given the unit reuse price α_{ij}, the transmit power $p_j^d[i]$ of each D2D link is bounded by $[p_{min}, p_{max}]$, and we firstly prove that the transmit power is continuous within the above range, as,

$$\frac{\partial U_j^f(\cdot)}{\partial p_j^d[i]} = W \cdot \frac{g_j^d[i]}{\ln 2(1+\gamma_j^d[i])(I_{ij}^d + \mathcal{N}_0)} - \beta^f \alpha_{ij} \bar{g}_{j,B}[i], \tag{24}$$

where I_{ij}^d represents the interference caused by other D2D links and cellular links reusing the same channel resources, as $I_{ij}^d = p_i^c \bar{g}_{i,j}[i] + \sum_{j',j' \neq ,j} x_{ij'} p_{j'}^d[i] \bar{g}_{j',j}^d[i]$. $\gamma_j^d[i]$ is the SINR, which can be represented with the transmit power $p_j^d[i]$ as shown in (1). Next, we prove that the utility function of the follower is strictly concave with respect to the transmit power. The second order derivative of the utility function can be expressed as,

$$\frac{\partial^2 U_j^f(\cdot)}{\partial {p_j^d}^2[i]} = \frac{-W {g_j^d[i]}^2}{\ln 2(1+\gamma_j^d[i])^2 (I_{ij}^d + \mathcal{N}_0)^2}. \tag{25}$$

Due to the fact that the second order derivative of the utility function is smaller than 0, i.e., $\frac{\partial^2 U_j^f(\cdot)}{\partial {p_j^d}^2[i]} \leqslant 0$, the utility function U_j^f of the follower is satisfied the strict concavity property with respect to the transmit power $p_j^d[i]$.

B. Proof of Proposition 2

The leaders' utility function can be represented as:

$$U^l(\alpha_{ij}, p_j^d[i]) = \beta^l \sum_{i \in \mathcal{M}} \sum_{j \in \mathcal{N}} x_{ij} B + \beta^l \sum_{i \in \mathcal{M}} \sum_{j \in \mathcal{N}} x_{ij} \alpha_{ij} C_{ij} + \sum_{i \in \mathcal{M}} W \log_2 (1 + \frac{A_i \prod_{k \in i\mathcal{K}} \alpha_{ik}}{(\sum_k x_{ik} C_{ik} + \mathcal{N}_0) \prod_{k \in i\mathcal{K}} \alpha_{ik} + B \sum_{k \in i\mathcal{K}} \nu_{ik}}) \tag{26}$$

where $A_i = p_i^c g_i^c$, $B = \frac{W}{\ln 2 \beta^f}$, and $C_{ij} = \frac{-(I_{ij}^d + \mathcal{N}_0)\bar{g}_{j,B}[i]}{g_j^d[i]}$. To facilitate the multiply operation in the utility function, the set $i\mathcal{K}$ is defined. The set $i\mathcal{K}$ contains all the D2D pairs which reuse cellular channel i, where $x_{ij} = 1$ for particular cellular channel i. Moreover, we define $\nu_{ik} = \prod_{k' \in i\mathcal{K}} \alpha_{ik'} / \alpha_{ik}$.

To obtain the optimal unit reuse price α_{ij}, we take the first order derivative of the utility function, which can be represented as:

$$\frac{\partial U^l(\alpha_{ij}, p_j^d[i])}{\partial \alpha_{ij}} = \beta^l C_{ij} + \frac{\beta^f A_i B^2}{[\alpha_{ij}(D_{ij} + A_i) + B][\alpha_{ij} D_{ij} + B]} \tag{27}$$

where $D_{ij} = \sum_k x_{ik} C_{ik} + \sum_{k/j} x_{ik} \frac{B}{\alpha_{ik}} + \mathcal{N}_0$.

Based on the different values of A_i and D_{ij}, similar to the proposed solution algorithm in [13,18], the optimal solution of (4) can be described in different situations. Therefore, the solutions are shown in the following cases:

1) $D_{ij} = 0$. The second order derivative of the leader's utility function can be calculated as:

$$\frac{\partial^2 U_j^l(\alpha_{ij}, p_j^d[i])}{\partial \alpha_{ij}^2} = \frac{-\beta^f A_i^2 B}{(A_i \alpha_{ij} + B)^2} < 0. \tag{28}$$

According to the above equation, the second order derivative is smaller than 0. Thus, the optimal unit reuse price can be uniquely obtained by setting the first order derivative, as shown in (27), equal to 0. The best response unit reuse price is:

$$\dot{\alpha}_{ij} = -\frac{B}{A_i} - \frac{\beta^f B}{\beta^l C_{ij}}. \tag{29}$$

We set the upper and lower bounds of the price, and then the best response unit reuse price α_{ij}^* can be selected in $\{\alpha_{ij}^{min}, \dot{\alpha}_{ij}, \alpha_{ij}^{max}\}$.

2) $A_i + D_{ij} = 0$. The second order derivative can be calculated as,

$$\frac{\beta^f A_i^2 B}{(-\alpha_{ij} A_i + B)^2} > 0. \tag{30}$$

Based on the above inequation, the utility function is continuous and strictly convex. Therefore, the optimal unit reuse price α_{ij}^* can be selected in $\{\alpha_{ij}^{min}, \alpha_{ij}^{max}\}$.

When $A_i + D_{ij} \neq 0$ and $D_{ij} \neq 0$. We first calculate the second order derivative of the utility function, as:

$$\frac{\partial^2 U^l(\cdot)}{\partial \alpha_{ij}^2} = \frac{-\beta^f A_i B^2((A + 2D_{ij})B + 2\alpha_{ij} D_{ij}(D_{ij} + A_i))}{[\alpha_{ij}(D_{ij} + A_i) + B]^2 [\alpha_{ij} D_{ij} + B]^2}. \tag{31}$$

Next, we focus on the denominator of the first order derivative and set $g(\alpha_{ij}) = (\alpha_{ij}(D_{ij} + A_i) + B)(\alpha_{ij} D_{ij} + B)$. The two roots of the $g(\alpha_{ij})$ are $\alpha_{ij}^1 = -\frac{B}{D_{ij}}$ and $\alpha_{ij}^2 = -\frac{B}{A_i + D_{ij}}$. We have $\alpha_{ij} < \alpha_{ij}^{max} < -\frac{B}{C_{ij}}$. Therefore, we have $C_{ij}\alpha_{ij} + B > 0$ and further get $D_{ij}\alpha_{ij} + B = \sum_{k \in i\mathcal{K}} C_{ik}\alpha_{ik} + \sum_{k \in i\mathcal{K}} B + \mathcal{N}_0 \alpha_{ij} > \mathcal{N}_0 \alpha_{ij} > 0$ and $\alpha_{ij}(D_{ij} + A_i) + B > 0$. Furthermore, we have the three following cases.

3) $D_{ij} > 0$. We have $A + 2D_{ij} > 0$ and $D_{ij}(D_{ij} + A_i) > 0$. Thus, the second order derivative is smaller than 0, as $\frac{\partial^2 U^l(\cdot)}{\partial \alpha_{ij}^2} < 0$. The utility function is concave respect to the unit reuse price α_{ij}. By setting the first order derivative equal to 0, we have the optimal solution as:

$$\dot{\alpha}_{ij} = \frac{-B(2D_{ij} + A_i) \pm \sqrt{\Delta}}{2D_{ij}(D_{ij} + A_i)}, \tag{32}$$

where $\Delta = A_i^2 B^2 (1 - \frac{4\beta^f}{\beta^l C_{ij}} (\frac{D_{ij}^2}{A_i} + D_{ij}))$. Since the smaller root is smaller than 0, we only keep the larger root, which can be represented as $\dot{\alpha}_{ij} = \frac{-B(2D_{ij} + A_i) + \sqrt{\Delta}}{2D_{ij}(D_{ij} + A_i)}$. Therefore, the optimal unit reuse price α_{ij}^* can be selected in $\{\alpha_{ij}^{min}, \dot{\alpha}_{ij}, \alpha_{ij}^{max}\}$.

4) $D_{ij} < 0$ and $A_i + D_{ij} > 0$. we have $\alpha_{ij}^2 < 0 < \alpha_{ij}^{\min} < \alpha_{ij} < \alpha_{ij}^{\max} < \alpha_{ij}^1$. Next, we talk about the cases of the second order derivative. When $\alpha_{ij} > \frac{(A_i + 2D_{ij})B}{2D_{ij}(D_{ij} + A)}$, we have $\frac{\partial^2 U^l(\cdot)}{\partial \alpha_{ij}^2} > 0$, and when $\alpha_{ij} < \frac{(A_i + 2D_{ij})B}{2D_{ij}(D_{ij} + A)}$, we have $\frac{\partial^2 U^l(\cdot)}{\partial \alpha_{ij}^2} < 0$. It should be noticed that $\alpha_{ij}^1 < \frac{(A_i + 2D_{ij})B}{2D_{ij}(D_{ij} + A)} < \alpha_{ij}^2$, and the

domain of definition $[\alpha_{ij}^{\min}, \alpha_{ij}^{\max}]$ is in the $[\alpha_{ij}^2, \alpha_{ij}^1]$. So we calculate the first order derivative on α_{ij}^2 and α_{ij}^1, respectively. We get $\lim_{\alpha_{ij}\to\alpha_{ij}^{1-}} \frac{\partial U^l}{\partial \alpha_{ij}} = +\infty$ and $\lim_{\alpha_{ij}\to\alpha_{ij}^{2+}} \frac{\partial U^l}{\partial \alpha_{ij}} = +\infty$. Therefore, the leader's utility will be increasing, decreasing and increasing sequentially in the domain of definition. Thus, the maximum point is either the smaller root of setting the first order derivative (27) equal to 0, i.e., $\dot{\alpha}_{ij} = \frac{-B(2D_{ij}+A_i)-\sqrt{\Delta}}{2D_{ij}(D_{ij}+A_i)}$ or the upper bound $\alpha_{ij}^{\max}$. In all, the best response unit reuse price can be uniquely selected in $\{\dot{\alpha}_{ij}, \alpha_{ij}^{\max}\}$.

5) $D_{ij} + A_i < 0$. We can obtain $D_{ij} < 0$ and $2D_{ij} + A_i < 0$. Further, we have $0 < \alpha_{ij}^{\min} < \alpha_{ij} < \alpha_{ij}^{\max}$. Because $D_{ij}\alpha_{ij} + B > 0$, and then we have $D_{ij}\alpha_{ij} > -B$. Therefore, we have $(A+2D_{ij})B+2\alpha_{ij}D_{ij}(D_{ij}+A_i) < (A+2D_{ij})B-2B(D_{ij}+A_i) = -A_iB < 0$. So the second order derivative of the utility function on the domain of definition between lower and upper bounds is larger than 0, as $\frac{\partial^2 U^l(\cdot)}{\partial \alpha_{ij}^2} > 0$. The utility function is strictly convex with respect to the unit reuse price. Therefore, the best response unit reuse price α_{ij}^* can be uniquely selected between the two boundaries $\{\alpha_{ij}^{\min}, \alpha_{ij}^{\max}\}$.

C. ST-Q Convergence Proof

We will prove that Q^j can be converged to the equilibrium ST-Q (Q_*^j) for agent j. Our ST-Q convergence proof is similar to the proof of Nash-Q in [23].

First of all, we have the following two assumptions and one lemma, which are similar to the convergence analysis of the Q-learning algorithm.

Assumption 1: Every state $s \in \mathcal{S}$ and every action $a^j \in \mathcal{A}$, for $j \in \mathcal{N}$, can be visited infinitely.

Assumption 2: The learning rate α follows the conditions below:

1. $0 \leq \alpha^t(s,\mathbf{a}) < 1$, $\sum_{t=0}^{\infty} \alpha^t(s,\mathbf{a}) = \infty$, $\sum_{t=0}^{\infty} [\alpha^t(s,\mathbf{a})]^2 < \infty$, the latter two hold uniformly and with probability 1.

2. $\alpha^t(s,\mathbf{a}) = 0$ if $\alpha^t(s,\mathbf{a}) \neq \alpha^t(s^t,\mathbf{a}^t)$.

Lemma 1: (Szepesvari and Littman [41], Corollary 5). Assume that α^t satisfies the Assumption 2 and the mapping $P^t : \mathbb{Q} \to \mathbb{Q}$ satisfies the following condition: there exists a number $\gamma \in [0,1]$ and a sequence γ^t converging to 0 with probability 1 such that $||P_tQ - P_tQ_*|| \leqslant \gamma||Q - Q_*|| + \lambda^t$ for all $Q \in \mathbb{Q}$ and $Q_* = E[P_tQ_*]$, then the Q-value updated by

$$Q^{t+1} = (1-\alpha^t)Q^t + \alpha^t[P_tQ^t] \tag{33}$$

convergences to Q_* with probability 1.

Definition 1: Define Q as the set of all agents value, as $Q = (Q^1, ..., Q^N)$, where $Q^j \in \mathbb{Q}^j$ for all $j \in \mathcal{N}$ and $\mathbb{Q} = \mathbb{Q}^1, , , \times\mathbb{Q}^N$. So $P_t : \mathbb{Q} \to \mathbb{Q}$ is a mapping on complete metric space $\mathbb{Q} \to \mathbb{Q}$ as $P_tQ = (P_tQ^1, ..., P_tQ^N)$, where

$$P_tQ^j(s,\mathbf{a}) = r^j(s,\mathbf{a})(t) + \gamma\pi_*(s')Q^j(s'), \tag{34}$$

where s' is the state of the next time step and $\pi_*(s')$ is the Stackelberg equilibrium at the state s', which can also be described as the Nash equilibrium for the stage game $(Q^1(s'), ..., Q^N(s'))$.

Lemma 2: For each state s, the relationship between Stackelberg Q-value Q_* and the equilibrium payoff ($V(s,\boldsymbol{\pi}_*)$) for agent j can be defined as,

$$Q_*^j(s,\mathbf{a}) = r^j(s,\mathbf{a}) + \gamma \sum_{s'\in S} p(s'|s,\mathbf{a})V^j(s',\boldsymbol{\pi}_*), \tag{35}$$

where $V^j(s',\boldsymbol{\pi}_*) = \boldsymbol{\pi}_*(s')Q_*^j(s')$. This lemma establishes the relationship between the optimal value (V) in the stochastic game and the Stackelberg equilibrium (Q_*) is the stage game. Thus, the following lemma holds.

Lemma 3: For a multiple player stochastic game, $E[P_tQ_*] = Q_*$.

Proof: According to Lemma 2, we have,

$$\begin{aligned} Q_*^j(s,\mathbf{a}) &= r^j(s,\mathbf{a}) + \gamma \sum_{s'\in S} p(s'|s,\mathbf{a})V^j(s',\boldsymbol{\pi}_*) \\ &= \sum_{s'\in S} p(s'|s,\mathbf{a})(r^j(s,\mathbf{a}) + \gamma\boldsymbol{\pi}_*(s')Q_*^j(s')) \\ &= E[P_t^jQ_*^j(s,\mathbf{a})]. \end{aligned} \tag{36}$$

Therefore, for all agent j, this will be held as $Q_* = E[P_tQ_*]$.

Assumption 3: Every stage game $(Q_t^1, ..., Q_t^N)$, for all t and all s, has a global optimal point or saddle point, and agents' payoffs in this equilibrium are used to update their Q-functions.

In our model, according to the Stackelberg game solutions, we can obtain the equilibrium for the leader and followers in each stage game, and the equilibrium can be guaranteed through proofs A and B. Thus, the assumption 3 is achievable. Further, we have the following definitions.

Definition 2:

$$\begin{aligned} ||Q - \hat{Q}|| &\equiv \max_{j,s} ||Q^j(s) - \hat{Q}^j(s)||_{(j,s)} \\ \max_{j,s,\mathbf{a}} &\equiv |Q^j(s,\mathbf{a}) - \hat{Q}^j(s,\mathbf{a})|. \end{aligned} \tag{37}$$

Lemma 4: (Hu & Wellman [23], Lemma 16). $||P_tQ - P_t\hat{Q}|| \leqslant \gamma||Q - \hat{Q}||$, for all $Q, \hat{Q} \in \mathbb{Q}$, where the assumption 3 holds.

Theorem 1: Under Assumptions $1-3$, the Q-value Q^j for all agent j updated by

$$Q^j(s,\mathbf{a})^{t+1} = (1-\alpha)Q^j(s,\mathbf{a})^t + \alpha[r^{jt} + \gamma\ ST\ V_*^j(s')], \tag{38}$$

converges to Stackelberg Q-value (ST-Q) Q_*^j, where $V_*^j(s') = \boldsymbol{\pi}_*(s')Q_*^j(s,\mathbf{a})$, which is also described in (20) and (21). $\boldsymbol{\pi}$ is the proper Stackelberg game solution for all agents.

Proof : The proof is an extended application of Lemma 1 given the following two conditions. First, P_t is a contraction operator, which is given by Lemma 4. Second, the fixed point condition $E[P_tQ_*] = Q_*$, which is also given by Lemma 3. With the following updating rules, the Q-value will converge to the ST-Q (Q_*) with probability 1.

In our model, we extend the state and action spaces to the high dimensional, and employ the neural networks to represent the value function and the policy function. However, the updating rules are still preserved, so the convergence can still be guaranteed.