Pricing QoE With Reinforcement Learning For Intelligent Wireless Multimedia Communications

Shuan He, Wei Wang
Department of Computer Science, San Diego State University
she@sdsu.edu, wwang@sdsu.edu

***Abstract*—With the unification of modern wireless services of Mixed Reality and Tactile Internet, providing Quality of Experience (QoE) with ultra low latency become critical challenges in edge resource allocation. In this paper we propose a reinforcement learning-based economic pricing model for wireless multimedia QoE, leveraging economic theories and machine intelligence. In the proposed approach, the QoE pricing model considers the User Equipment (UE)'s perceived QoE, the amount of purchased data, the wireless channel conditions, and the user's subjective multimedia content preference. In addition, the QoE gain of UE, cost of three entities in wireless networks - Content Provider (CP), Wireless Carrier (WC), and UE, are integrated in the economic concept of social utility. The social utility would be affected by all system factors such as unit data price, multimedia quality requirement, and wireless channel conditions. The proposed reinforcement learning method improves the social utility performance by maximizing the accumulated utility through obtaining the optimal factors set up. At last, through numerical simulations we show the impacts of different system parameters on UE's QoE gain and the improvement of social utility performance by using the proposed reinforcement learning approach.**

***Index Terms*– Reinforcement Learning, Smart Media Pricing, Quality of Experience**

I. Introduction

How to improve the Quality of Experience (QoE) and to reduce latency of User Equipment (UE) remains the obstacles for emerging wireless Tactile Internet with Mixed Reality multimedia traffic. The increased 4K or 8K picture resolution and low delivery latency lead to new challenges to rethink the economic modeling the relationships among User Equipment (UE), Content Provider (CP), and Wireless Carrier (WC) [1-3]. The CP and WC would have to bear financial cost to provide the Mixed Reality multimedia service through wireless Tactile Internet devices, and the Mixed Reality headset UE would have to purchase wireless resources to obtain desirable QoE. As illustrated in Fig. 1, the economic cost of CP to keep certain level of data quality (i.e., data distortion reduction and data length), the cost of WC to ensure the desirable packet error rate (i.e., through control power), and the cost of UE to purchase the data are collectively considered in the social utility model.

By combining the QoE and the economic social utility together, we are able to address the aforementioned challenge through properly analyzing the impacts of all system factors (i.e., channel conditions, data quality, and incurred cost of entity) on the QoE gain and social utility gain. While taking all factors into consideration simultaneously would be handful.

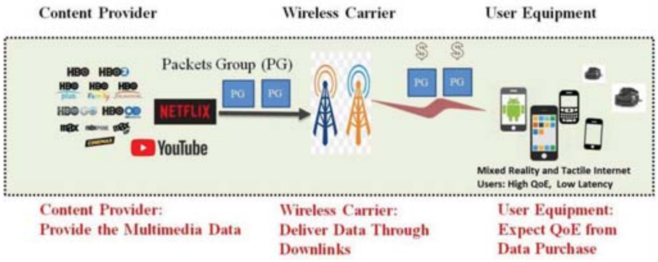
Content Provider
Wireless Carrier
User Equipment
Packets Group (PG)
PG
PG
PG
PG
$
$
NETFLIX
YouTube
Mixed Reality and Tactile Internet Users: High QoE, Low Latency
Content Provider: Provide the Multimedia Data
Wireless Carrier: Deliver Data Through Downlinks
User Equipment: Expect QoE from Data Purchase

Fig. 1. The wireless multimedia quality pricing scenario we studied in this paper. Three entities (i.e., content provider, wireless carrier, and user equipment) are participating in the pricing design.

Therefore, a machine learning based method is introduced in our work, where all system factors are included and manifested in terms of state. The three entities (CP-WC-UE) are treated as the agent in the learning procedure and it wanders among states (the agent's transition from one state to another state is accomplished by taking action). In addition, the social utility gain at each state serves as the reward of agent by taking the action. Our goal of learning is to obtain the optimal action policy for the agent, so that the maximization of accumulated reward is achievable.

The rest of this manuscript is organized as follows. A short review of related work about machine learning method and QoE improvement issue is given in Section II. The detailed explanations of our studied multimedia service system model is shown in Section III. The reinforcement Q-learning method is designed in Section IV. We carry out the numerical simulations to evaluate the system performance in Section V. The conclusion of this paper is drawn in Section VI. The key notations and nomenclature used in this paper are summarized in Table I.

II. Related Work

To solve the UE's QoE performance issue in the wireless multimedia service, lots of research has been conducted in the literature, from the objective Quality of Service (QoS) model developing to the cognitive QoE model perfection [4-6]. Authors in [4] provided a multi-polling controlled access scheme to guarantee the important video delivering latency and reduce the transmission overhead, for the purpose of improving the QoS/QoE in the wireless video service. In their work, the Multiple Input Multiple Output (MIMO) feature is

978-1-7281-5089-5/20/$31.00 ©2020 IEEE

TABLE I
THE KEY NOTATIONS USED IN THIS PAPER.

Symbol	Comments
p_k	The packet error rate of packet k.
π_i	Positional dependency set of packet i.
q_i	Multimedia quality contribution of packet i.
l_i	The length of the $i-th$ frame.
y_i	The per-bit price of multimedia packet i.
N, M	UE's requested data, N PGs and M packets per PG.
$a_1 \sim a_4$	System parameters to describe UE's QoE.
α, β, γ	The price coefficients in the social utility definition.
φ	The coefficient of UE's preference on the multimedia content.
s, a, r	The state, action, and reward in the Q-learning procedure.
d_f	The discount factor of the old reward in Q-learning.
l_r	The learning rate in the Epsilon greedy algorithm.

exploited as well in their proposed cross-layer quality adjustment strategy. The MIMO technology is also utilized in [5] where authors provided an optimal multimedia relay scheme to improve the UE's QoE gain. The multimedia content distribution and devices' antenna selection are jointly considered in their research. In addition, authors in [6] proposed a context-aware wireless multimedia relay solution to improve the QoE of mobile user. The game theory is utilized by the authors and a Stackelberg game model is provided in their work to analyze the relationships among mobile user, relay device and base station. In the aforementioned work, the QoE model is constructed only based on the objective factors such as data rates and multimedia quality. In our work, the UE's multimedia content preference will be considered in the QoE model as well.

Furthermore, plenty of research has been published on modeling the business relationships among CP, WC and UE, in fixed pricing type and dynamic pricing type [7-11]. Authors in [8] classified the static-based pricing and dynamic-based pricing schemes between wireless service provider and UE, where factors such as network capacity, available bandwidth, frequency spectrum, and network hops are involved in the price calculation. A smart media pricing framework is provided in [9] to deal with the price and data usage between service provider and client device. By seeking the game equilibrium, both service provider and client could reach the maximum utility gain. To extend the two players situation to three players in the wireless service, authors in [10] adopted a generalized best response game solution to improve the UE's QoE and the profit of CP, WC simultaneously. Authors in [11] proposed a new polymatroid theoretic framework to maximize the three-party achievable profit through proper bandwidth allocation. Similar to [10] and [11], we focus on the three-party wireless service with specific consideration of pricing the QoE. The social utility concept is proposed in this paper to unite three entities together and our goal is to maximize the social utility gain.

One challenge to unite three entities together in terms of social utility is: the factors would not only affect individual entity but also impact the social utility in a complex fashion. To handle such challenge that taking various factors into consideration simultaneously, a machine learning based method is proposed in our work. As one of the promising artificial intelligence tools, lots of research has been carried out in the literature by using the machine learning approaches in the wireless network service [12-14]. Authors in [12] investigated the machine learning algorithm's motivation refining, problem formulation, and methodology in their work. The learning efficiency is important to achieve QoS, facing complex challenges. Authors in [13] presented the application of machine learning techniques to improve the congestion control of TCP in wired and wireless networks. The reinforcement learning is adopted in [14] to develop a novel decentralized resource allocation mechanism for vehicle-to-vehicle communications. All the factors in the environment are modeled as states in their work. Inspired by aforementioned work, a reinforcement learning based Epsilon-greedy Q-learning method is proposed in this paper to address the social utility maximization problem for wireless multimedia. Through the learning procedure of the three agents, i.e., CP-WC-UE, the system would reach the maximum social utility with the optimal policy output.

III. SYSTEM MODEL

As shown in Fig. 1, the economic pricing model of wireless multimedia QoE is illustrated at the network edge. The UEs are typically Tactile Internet Mixture Reality users with picky QoE requirements. The CP will provide the source multimedia data to WC based on UEs' content purchasing requests. The WC is responsible for transmitting data to UE through wireless channels and provide the QoE. While for a individual UE, it purchases the source data from CP and obtain the QoE gain form the WC service.

A. QoE Model of Wireless Multimedia

It is realistic to assume UEs request multimedia data in the form of Packets Group (PG), where each PG contains multiple packets. Let p_k denote the packet error rate of the $k-th$ packet, l denote the packet length, q_i denote the multimedia quality contribution of packet i, π_i denote the positional dependency set of packet i, i.e., decoding ancestor set including itself. Then, the multimedia service quality-throughput contribution of a certain PG could be quantitatively approximated as $\sum_i q_i * l_i * \prod_{k \in \pi_i}(1-p_k)$, and the financial cost of acquiring this PG is estimated as $\sum_i y_i(q_i, \pi_i) * l_i$. The term y_i here denotes the per-bit price of multimedia packet i, which is related to the packet's quality contribution q_i.

In this paper, the personalized QoE model is adopted in a way similar to [15], and we added additional packet level multimedia distortion to the QoE model. We assume each UE consumes totally N PGs in the multimedia service, and thus the QoE expectation will be approximated as:

$$QoE = \frac{a_1}{1 + e^{-a_2 \sum_{i=1}^{N} \sum_{j=1}^{M_i} q_{i,j} * l_{i,j} * \prod_{k \in \pi_{i,j}} (1-p_k) + a_3 * \varphi + a_4}}. \tag{1}$$

where a_1, a_2, a_3, and a_4 are system parameters to configure the QoE model. Term φ denotes the UE's personal preference of a certain multimedia content. For example, some UEs could be a big fan of a football match video, and some UEs may not be interested in football at all. Thus these UEs could have very different values of φ even on the same data content. The financial cost of UE to purchase those PGs is calculated as follows, assuming per-bit price y_i for each packet is determined by the CP.

$$C_{ue} = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{i,j} * l_{i,j} \tag{2}$$

B. Social Utility Definition

Let C_{cp} and C_{wc} denote the cost function at CP side and WC side, respectively. The social utility U_{social} is defined as the UE's utility, which is related to UE's obtained QoE, subtracted by the cost of all entities:

$$U_{social} = U(QoE) - C_{ue}\left(\sum y * l\right) - C_{cp}\left(\sum q * l\right) - C_{wc}\left(\sum \sum \prod_{k\in\pi_{i,j}} (1-p_k)\right). \tag{3}$$

$$U(QoE) = \frac{\alpha * a_1}{1 + e^{-a_2 \sum_{i=1}^{N} \sum_{j=1}^{M_i} q_{i,j} * l_{i,j} * \prod_{k\in\pi_{i,j}} (1-p_k) + a_3 * \varphi + a_4}}. \tag{4}$$

The term α is a constant parameter to align QoE to utility. The cost of CP is directly related to its source coding parameter settings, such as the compression ratio and rate-distortion truncation of bit-planes:

$$C_{cp} = \beta * \sum_{i=1}^{N} \sum_{j=1}^{M_i} q_{i,j} * l_{i,j}, \tag{5}$$

where the β is the system parameter to align the CP's cost to utility. The cost of WC is approximated as wireless channel quality, i.e., the packet error rate it can provide in the wireless channels. The constant parameter γ is to align the packet error rate to utility:

$$C_{wc} = \gamma * \sum_{i=1}^{N} \sum_{j=1}^{M_i} log \prod_{k\in\pi_{i,j}} (1-p_k). \tag{6}$$

C. System Objective

The goal of our work is to maximize the social utility U_{social} by finding the proper amount of data that UE purchase from CP, and the optimal price setting for each multimedia PG. Regards to the optimal per-bit price y_j, the feasibility of achieving such an optimal solution would be impractical for a large amount of multimedia packets within a number of PGs. Let y_0 denote the normalized base price, i.e., the unit quality gain price for each bit. Then the per-bit multimedia gain price of packet j could be presented as $y_j = y_0 * \sum_{k\in\pi_j} q_k$. Then, instead of finding the optimal price for each packet, we change our goal to find the optimal base price y_0 and optimal multimedia transaction bits $\sum l$, for the purpose of maximizing the social utility U_{social}.

$$\{y_0, l\} = argmax\{U_{social}\} \tag{7}$$

IV. Reinforcement Q-Learning QoE Pricing Solutions

A. Q-Learning Pricing Algorithm Design

As one of the important machine learning paradigms, the reinforcement learning has been widely studied to address the exploration vs. exploitation trade-off in the finite Markov Decision Process (MDP) scenarios. A typical reinforcement learning model includes 5 features. 1): A set of environment and agent states, S. 2): A set of actions of the agent, A. 3): The probability of transition from state s to s', $P_a(s,s') = P_r(s_{t+1} = s'|s_t = s, a_t = a)$. 4): After the transition from state s to s', the agent gains an immediate reward $R_a(s,s')$. 5): The rules (a.k.a policy) that describe what the agent observes, P. Recall the wireless service scenario in our work, where the CP-WC-UE system is treated as the agent in the learning procedure. The state observed by agent for characterizing the environment includes all the factors that affect social utility, i.e., the quality-throughput contribution $\sum_i q_i * l_i * \prod_{k\in\pi_i}(1-p_k)$, the content preference of UE φ, the packet error rate in the channel p_k, the consumed data length of UE $\sum l$, and the CP's base price base price y_0. Thus, the state is expressed as $s_t = [\sum_i q_i * l_i * \prod_{k\in\pi_i}(1-p_k), \varphi, p_k, \sum l, y_0]_t$. At each time t, the agent observes a state s_t from the state space S, takes an action a_t from the action space A, gains the reward r_t and reaches the new state s_{t+1}.

It is worth pointing out that the transition probability and reward function are not available in our proposed wireless service scenario. The rationale behind this is that we assume the system factors, i.e., channel quality, UE's requested data, are changing stochastically with time. The system will randomly move to next state if any factor in current state $s_t = [q_i, \varphi, p_k, \sum l, y_0]_t$ changes. Thus, the model-free reinforcement learning will be utilized in our work, i.e., the Q-learning approach. The goal of Q-learning is to learn a policy, which guides the agent to take the optimal action under any circumstances. The Q-learning model adopted in this paper works well in handling problems with stochastic transitions and rewards.

Since to the reward function is not available in the proposed wireless service scenario, we would use instant social utility gain (i.e., U_{social} of next state) as the reward of agent when we adapt the Q-learning into our work. More specifically,

$$r_t = U_{social}\left([q_i, \varphi, p_k, \sum l, y_0]_{t+1}\right) \tag{8}$$

where r_t denotes the reward the agent obtains by taking action a_t at state s_t. The calculation of $U_{social}([q_i, \varphi, p_k, \sum l, y_0]_{t+1})$ is given in Equation (3), which implies the social utility gain at state s_{t+1}. The state transition and reward are stochastic by following the MPD. They only depend on the state of environment and the actions taken by the agent. The Q-learning obtains the optimal agent actions policy to maximize the social utility:

$$argmax\{G_t\}|G_t = E[\sum_{n=0}^{\infty} d_f^n r_{t+n}] \tag{9}$$

where d_f denotes the discount factor of the previous reward. The core factors that would affect the agent's reward (a.k.a social utility gain) include channel quality p_k, base price of data y_0, and user's consumed data amount $\sum l$. Assume there are X possible channel conditions, i.e., $\forall p_k \in [p_k^1, ..., p_k^X]$, Y available base price options for CP, i.e., $\forall y_o \in [y_0^1, ..., y_0^Y]$, and Z data request options for UE, i.e., $\forall \sum l \in [L^1, ..., L^Z]$. Then the size of agent's action space is $X * Y * Z$. The goal of Q-learning is to obtain the optimal action policy P^*, so that the agent would achieve the maximum Q value from any state: $a_t = argmaxQ(s_t, a_t)$ through $a_t \in P^*$. The Q value for a given state-action pair (s_t, a_t), denoted as $Q(s_t, a_t)$, can be calculated and updated according to the dynamic Bellman Q-function [14] without any knowledge of the system.

$$\begin{aligned} &Q_{new}(s_t, a_t) = Q_{old}(s_t, a_t) + \\ &l_r * [r_t + d_f * max\{Q_{old}(s, a_t)\} - Q_{old}(s, a_t)] . \end{aligned} \tag{10}$$

where the l_r denotes the learning rate of the process. It has been proved that the Q value at action-state pair (s_t, a_t) will converge to the optimal $Q^*(s_t, a_t)$ if each action-state is visited infinite times with properly setting up the learning rate l_r [16].

Generally speaking, the channel condition in the physical environment keeps stable in certain period of time. Thus we would cut down the size of agent's action space to $Y * Z$ in our actual Q-learning application by assuming the channel keeps constant. The rationale behind such simplification is that reduction of action space would dramatically decrease the iterations of Q-learning procedure, where all exploration steps are stochastically chosen. It is important pointing out the extra cost the simplification is that we need to re-run the Q-learning to obtain the new optimal action policy P^* when the channel condition of UE occurs significant changes, i.e., packet error rate changes from p_k^i to p_k^j.

B. Algorithm Analysis

The learning procedure of the agent (i.e., the CP-WC-UE system) is illustrated in Algorithm 1, where the Epsilon-greedy strategy is utilized. In order to reduce the complexity of Q matrix (i.e., the Q matrix is with size of Q[length, action]) and to improve the learning efficiency, the two dimensional (2D) agent action space (i.e., $Y * Z$) is transformed into one dimension (1D) when the agent starts exploring. The action space transformation has the following features: 1) The total number of states in two scenarios doesn't change; 2) In the 2D scenario, there would be up to 4 options for the next move, while there are only 2 options in the 1D scenario. The convergence speed in the 1D would be faster than it in 2D. 3) The output optimal policies are different in two scenarios, but they would lead the agent converge to the same $Q^*(s_t, a_t)$.

At the beginning of learning, we initialize the Q matrix with zeros, as shown in step 3. The value of length(state) is $Y * Z$ and the value of length(action) is 2 (either moving forward or backward). The maximum iterations K in step 4 in decided by the convergence speed of algorithm. The algorithm would not stop running without K, even the Q matrix already converged. Thus, we will decide the K after multiple attempts to ensure the convergence of algorithm. Regarding to the terminal states mentioned in step 13, we assume the state (L^1, y_0^1) and state (L^Z, y_0^Y) are the two terminal states for the learning. The agent would gain zero reward if its action leads it to terminal states.

Algorithm 1 The Epsilon-Greedy Q-Learning Algorithm

```
1: Inputs: (1) The system parameters to determine the QoE gain of UE, i.e., $a_1 \sim a_4$, $\varphi$. (2) The system preset parameters $\alpha$, $\beta$, and $\gamma$ for UE, CP, and WC, in order to calculate the social utility. (3) The parameters for Q-learning process, e.g., learning rate $l_r$, discount factor $d_f$, and exploration probability $\epsilon$. (4) The parameters in multimedia service, i.e., the packet error rate $p_k$, the multimedia quality $q_k$, the range of base price $y_0$, and the range of data request $\sum l$.
2: Outputs: (1) The optimal action policy $P^*$ for the agent. (2) The state-action set that achieves the optimal $Q^*(s_t, a_t)$, where we obtain the maximum $U_{social}$ .
3: Initialize the Q matrix: Q = zeros(length(state), length(action)); The maximum number of iterations $K$. The initial state to start (randomly).
4: For $s = 1 : K$
5:     $X = rand()$; Get one uniform random number;
6:     $X = sum(X >= cumsum([0, 1 - \epsilon, \epsilon])$.
7:     If $X == 1$
8:         Next action: choose exploitation.
9:     Else
10:        Next action: choose exploration.
11:    Based on reward definition $r_t = U_{social}([q_i, \varphi, p_k, \sum l, y_0]_{t+1})$ and Equation (1) $\sim$ (6), calculate the reward of agent at next action. Update the state;
12:    Update the Q matrix using the Q-learning rule: $Q_{new}(s_t, a_t) = Q_{old}(s_t, a_t) + l_r * [r_t + d_f * max\{Q_{old}(s, a_t)\} - Q_{old}(s, a_t)]$.
13:    If current state is terminal state
14:        Restart the episode with a new (random) state;
15: End for
16: Output the optimal action policy based on the Q matrix;
```

V. Simulation

In this section, we carry out simulations to evaluate the system performance. In our simulations, the standard test video *Harbour* from H.265 codec is used in the multimedia transmission. More system parameters and their values are listed in Table II.

TABLE II
MAJOR PARAMETERS AND THEIR VALUES IN THE SIMULATIONS.

Symbol	Comments
q_i	$35 \sim 39dB$
y_i	$0.1 \sim 0.5$
N	$2 \sim 5$
M	17
a_1	20
$a_2 \sim a_4$	$1 \sim 6$
d_f	0.9
l_r	0.85
ϵ	0.95

In order to properly adjust the QoE model in our work, we evaluate the impact of parameters a_2 and a_3 on the UE's QoE gain. From the QoE definition Equation (1) we know the a_2 is related to the "objective" service quality (i.e., q_i, l_i, and p_i) and the a_3 is related to the UE's "subjective" experience (i.e., content preference, psychological factor). As illustrated in Fig. 2, the content reference factor would not change the maximum QoE gain of UE (in fact, the maximum QoE gain is determined by the preset parameter a_1). While a_3 has significantly impact on the amount of UE's consumed data in order to reach the maximum QoE gain. If UE would request more data to obtain high QoE gain if the UE has high preference on the multimedia data.

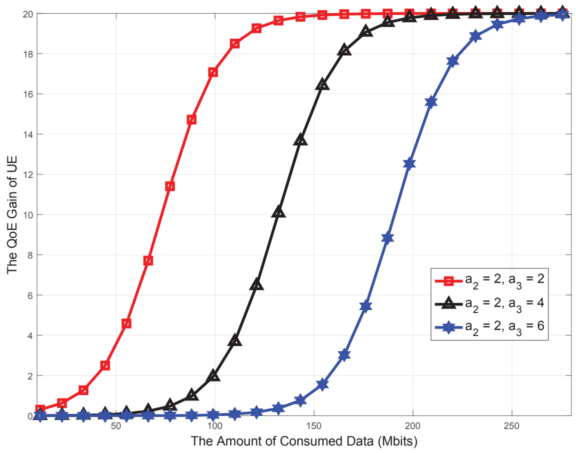

Fig. 2. The illustration of how the subjective system parameter a_3 impacts the QoE gain of UE.

In addition, we test the parameter of objective service quality part and the result is shown in Fig. 3. We can observe that the increase of a_2 would not change the maximum QoE gain of UE as well. But the bigger a_2 set up, the faster that UE would reach the maximum QoE with respect to the consumed data. We observe that the parameters a_2 and a_3 have similar influence on the QoE gain of UE. That is to say, both "service quality" and "content preference" will impact the speed of UE to reach its maximum QoE gain. While when UE has high content preference on certain data, the larger amount of data is need.

As we mentioned before, the social utility gain would be affected by a lot of factors. In Fig. 4, we evaluate the social utility gain under different sets up of environment parameters. We simulate different service situations by taking various values of cost coefficients of CP and WC (i.e., β and γ) and the base price y_0. It is worth pointing out that the value of γ is negative, and thus the $\gamma * log \prod_{k \in \pi_{i,j}} (1 - p_k)$ in Equation (6) is always negative. From the results we observe that the curve of social utility gain is concave, meaning there is an optimal amount of purchased data. In addition, comparing with the cost coefficient of CP β, the WC's cost coefficient has higher influence on the social utility. Generally speaking, higher cost coefficients lead to the decreased social utility.

As one of the most important factor in the business, we evaluate the impact of base price on the social utility gain in Fig. 5. From the result we observe that with the increase of base price of the data, the social utility gain will keep decreasing. The optimal amount of consumed data decreases as well. The rationale behind this is that when the data's price is high, UE would like to consume less data, and vice versa. When other system parameters keep stable, the base price and amount of consumed data will affect the UE's QoE and social utility. We can observe there is only one optimal base price and consumed data pair that maximizes the social utility.

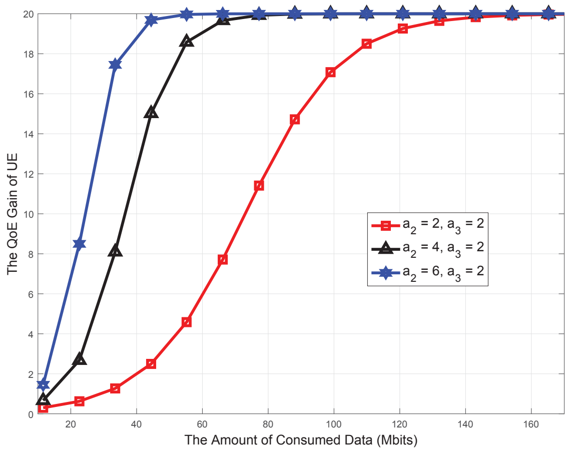

Fig. 3. The illustration of how the objective system parameter a_2 impacts the QoE gain of UE.

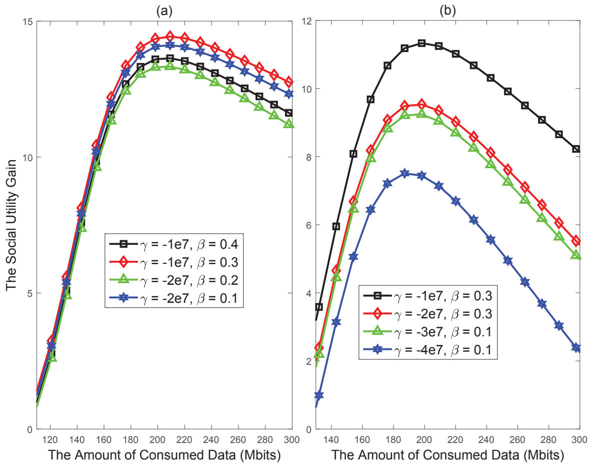

Fig. 4. The illustration of social utility gain under different environment set up. The base price in panel (a) and (b) are set as $y_0 = 0.1$ and $y_0 = 0.5$, respectively.

In Fig. 6 we compare the social utility performances under two schemes: in the proposed reinforcement learning approach (marked as RL method), and the fixed price approach, in different service scenarios. In the RL method, the base price and amount of consumed data are obtained from the Q-learning algorithm. While in the fixed price method, the base price is pre-set, and the amount of consumed data is chosen from the simulation in previous figure where the data amount

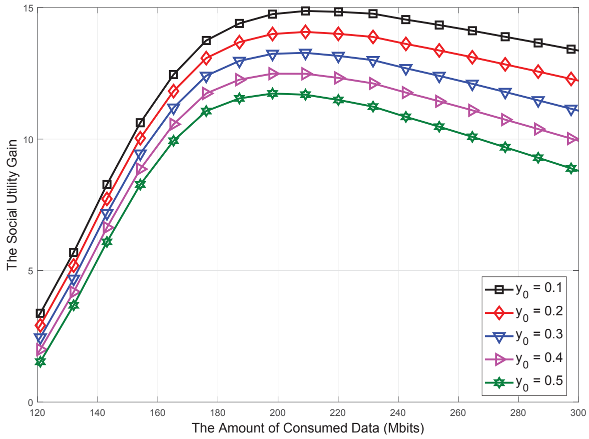

Fig. 5. The social utility gain performance under different sets of base price y_0 and amount of consumed data $\sum l$, in the simulation we set up $\beta = 0.1$, $\gamma = -2e7$.

achieves the maximization of social utility. We can observe from that the proposed Q-learning method outperforms the traditional method in all environment parameters set up. The rationale behind this is that: comparing with the traditional method, both base price and consumed data are dynamic along with the changing of environment factors in the learning method. The proposed Q-learning approach has better social utility performance.

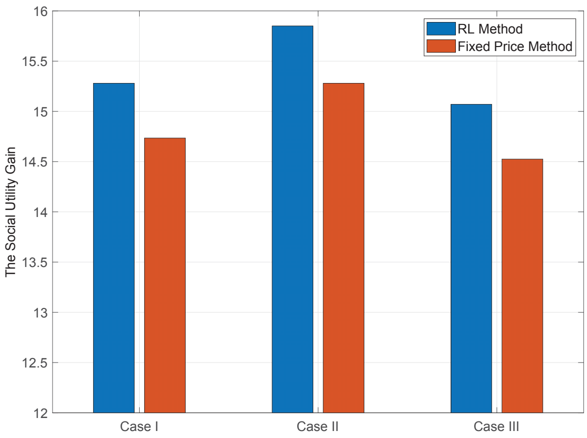

Fig. 6. The illustration of the social utility performance with proposed reinforcement learning method and conventional fixed price method, in different service environments: Case I: $\gamma = -1e7, \beta = 0.4$; Case II, $\gamma = -1e7, \beta = 0.3$; Case III: $\gamma = -2e7, \beta = 0.2$.

VI. Conclusion

In this paper, we proposed a reinforcement learning-based QoE pricing model for wireless multimedia communications in emerging networks such as Tactile Internet with Mixture Reality. In the economic pricing model we considered the media QoE, the amount of purchased data, the packet error rate in the wireless channel, and the multimedia content preference of users. Furthermore, instead of analyzing individual entity's financial cost and reward, we described the economic concept of social utility to unite the QoE of UE and the operating costs of UE-CP-WC together. Facing the fact that factors such as channel conditions, multimedia content preference, and data price are variance from time to time, we develop a reinforcement learning solution to handle the complex and dynamic aspects of the economic pricing model. We evaluated the QoE and utility with different parameters and demonstrated the performance improvement of the proposed Q-learning QoE pricing model through simulations.

VII. Acknowledgement

This research was supported in part by National Science Foundation Grant CNS-1744182.

References

[1] S. Bell, "Human-centric smart cities: service providers the essential glue." *White Paper, Heavy Reading Reports,* CISCO, October, 2018.

[2] W. Wang, and Q. Wang, "Price the QoE, not the data: SMP-economic resource allocation in wireless multimedia Internet of Things." *IEEE Communications Magazine,* vol. 56, no. 9, pp. 74-79, 2018.

[3] X. Tao, C. Jiang, J. Liu, A. Xiao, Y. Qian, and J. Lu, "QoE Driven Resource Allocation in Next Generation Wireless Networks." *IEEE Wireless Communications*, 2018.

[4] C. Y. Chang, H. C. Yen, C. C. Lin, and D. J. Deng, "QoS/QoE support for H. 264/AVC video stream in IEEE 802.11 ac WLANs." *IEEE Systems Journal,* vol. 11, no. 4, pp. 2546-2555, 2017.

[5] S. He, and W. Wang, "Wireless image relay: Prioritized QoE scheduling with simplified space-time coding mode selection." *In International Conference on Wireless Algorithms, Systems, and Applications (WASA)*, pp. 605-616, Springer, Cham, 2017.

[6] S. He, and W. Wang, "Context-aware QoE-price equilibrium for wireless multimedia relay communications using Stackelberg game." In *Proc. IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2017.

[7] V. Pandey, D. Ghosal, and B. Mukherjee, "Pricing-based approaches in the design of next-generation wireless networks: A review and a unified proposal." *IEEE Communications Surveys and Tutorials.* Vol. 9, no. 1-4, pp. 88-101, 2007.

[8] C. A. Gizelis, and D. D. Vergados, "A survey of pricing schemes in wireless networks." *IEEE Communications Surveys & Tutorials.* Vol. 13, no. 1, pp. 126-145, 2011.

[9] Q. Wang, W. Wang, S. Jin, H. Zhu, and N. T. Zhang, "Smart media pricing (SMP): Non-uniform packet pricing game for wireless multimedia communications." In *Proc. 2016 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2016.

[10] S. He, and W. Wang, "A Generalized Best-Response Smart Media Pricing Economic Model for Wireless Multimedia Communications." In *Proc. 2019 16th IEEE Annual Consumer Communications & Networking Conference (CCNC),* 2019.

[11] W. Ji, P. Frossard, B. W. Chen, and Y. Chen, "Profit optimization for wireless video broadcasting systems based on polymatroidal analysis." *IEEE Transactions on Multimedia,* vol. 17, no. 12, pp. 2310-2327, 2015.

[12] C. Jiang, H. Zhang, Y. Ren, Z. Han, K. C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks." *IEEE Wireless Communications,* vol. 24, no. 2, pp. 98-105, 2017.

[13] P. Geurts, I. El Khayat, and G. Leduc, "A machine learning approach to improve congestion control over wireless computer networks." In *Proc. IEEE International Conference on Data Mining,* November, 2004.

[14] H. Ye, and G. Y. Li, "Deep reinforcement learning for resource allocation in V2V communications," in *Proc. IEEE International Conference on Communications (ICC),* 2018.

[15] Y. Wang, P. Li, L. Jiao, Z. Su, N. Cheng, XS. Shen, and P. Zhang, "A data-driven architecture for personalized QoE management in 5G wireless networks." *IEEE Wireless Communications,* vol. 24, no. 1, pp. 102-110, 2017.

[16] Y. Luo, Z. Shi, X. Zhou, Q. Liu, and Q. Yi, "Dynamic resource allocations based on Q-learning for D2D communication in cellular networks." In *Proc. 2014 11th International Computer Conference Wavelet Active Media Technology and Information Processing (ICCWAMTIP),* 2014.