Smart Vehicular Communication via 5G mmWaves

Xiaotong Li[a], Ruiting Zhou[b,*], Ying-Jun Angela Zhang[c], Lei Jiao[d], Zongpeng Li[a]

[a] *School of Computer Science, Wuhan University, Wuhan, China*
[b] *Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan, China*
[c] *Department of Information Engineering, the Chinese University of Hong Kong, Shatin, New Territories, Hong Kong*
[d] *Department of Computer and Information Science, University of Oregon, Eugene, OR, USA*

Abstract

Millimeter-Wave (mmWave) communication is conceived as a viable approach for 5G vehicular communication systems, where vehicles are equipped with more sensors that generate Gbps data for future autonomous driving. However, such directional mmWave communication relies on accurate beam alignment and is sensitive to blockage. Dense deployment of mmWave base stations (mmBSs) and high mobility of vehicles also lead to frequent handovers and complex beam alignment calculation. 5G mmWave vehicular communication calls for a smart and stable solution. To this end, we propose an online learning scheme to address the problem of beam selection with blockage-free guarantee in 5G mmWave vehicular networks. We first model this problem as a contextual combinatorial multi-armed bandit (MAB) problem with QoS constraints and delayed feedback. Next, we propose an online learning algorithm, BPG, to predict beam directions, with provable sub-linear regret and blockage-free bounds. BPG exploits the context space and learns the expected weight of each beam from arrived vehicles' contexts and the delayed feedback. To validate the efficiency of BPG, we also conduct trace-driven simulations based on real-world traffic patterns. Simulation results show that BPG achieves close-to-optimal throughput with low violation and outperforms other benchmark algorithms.

1. Introduction

Vehicle-to-everything (V2X) represents a critical architecture for vehicular communication, paving the way for the upcoming fully autonomous driving era. Future intelligent transportation systems emphasize the need for communication links with high bandwidth, low latency and high reliability. For example, cooperative collision avoidance and high-density platooning leverage enormous Gbps of data generated from sensors [1], which requires new levels of communication reliability and bandwidth; V2X augmented reality transmits large amounts of real-time video for navigation systems [2], which challenges existing crowded sub-6GHz bands. Since lower bands are heavily congested with Wi-Fi, Bluetooth and current 4G LTE signals, mmWaves (30G-300GHz) with significantly larger bandwidth have become a viable candidate to power 5G networks. However, mmWave communication is still challenging in terms of severe propagation loss and sensitivity to blockage [3]. To compensate for these impairments, beamforming (BF) is adopted as an essential technique, utilizing directional beams for mmWave communications. Dense deployment of mmWave base stations (mmBSs) brings even more potential for wide coverage. All the developments have triggered the interest in developing a high reliability and low latency 5G mmWave vehicular communication system.

However, supporting high mobility vehicular networks via mmWaves still faces severe challenges. First, directional mmWave communication between mmBSs and vehicles requires accurate beam alignments to ensure successful data transmission. Different beams need to be chosen to deal with different traffic pattern. Second, mmWaves are sensitive to blockages due to weak diffraction ability [3]. Surrounding buildings, moving vehicles and passengers constitute pos-

[*] Corresponding author

Email addresses: xtlee@whu.edu.cn (Xiaotong Li), ruitingzhou@whu.edu.cn (Ruiting Zhou), yjzhang@ie.cuhk.edu.hk. (Ying-Jun Angela Zhang), jiao@cs.uoregon.edu (Lei Jiao), zongpeng@whu.edu.cn (Zongpeng Li) sible obstacles during communication. Once blockage happens, the communication link is interrupted, which severely harms vehicular communication. In order to establish stable communication links, it is extremely important to take blockage into consideration. Thus, a minimum threshold which guarantees the average blockage-free probability needs to be met. Therefore, mmBSs should balance between maximizing the throughput and avoiding blockages as much as possible to better guarantee system performance. Third, given that mmWave has high propagation loss and poor penetration, the effective communication range of a mmBS remains around 100m at best [4]. 5G mobile communication no longer relies on the deployment structure of large base stations, and a large number of small base stations will become a new trend, which can cover the peripheral communication that cannot be touched by large base stations. Vehicles with high-mobility in a dense mmWave deployments need to frequently hand over between neighbouring mmBSs (from the source mmBS to the target mmBS). Frequent handovers bring about high beam realignment overhead and decision latency for the target mmBS. Therefore, we come up with an idea that the source mmBS can make beam alignment predictions ahead for the target mmBS according to current traffic pattern. As a result, smart beam prediction methods, which can autonomously adapt to the real-time traffic situations and blockages with low complexity and latency, are in urgent need.

A fundamental problem for the source mmBS exists. *How to select beams for the target mmBS such that the throughput is maximized and the average blockage free probability satisfies the minimum threshold?* To effectively handle handovers and keep the vehicular communication stable, we propose a smart vehicular communication scheme to predict the beam directions of the target mmBS, from the perspective of the source mmBS.

The basic idea of our method is that the source mmBS (which is communicating with vehicles currently) infers the beam directions for target mmBS in advance according to current traffic pattern. We carefully design an online beam prediction algorithm with performance guarantee, BPG, leveraging the basic idea of contextual multi-armed bandit. Different from existing beam selection algorithms which either neglect the system QoS requirement [5][6] or hardly deal with the handover situations [7][8], our algorithm, BPG, shows better performance especially in the following aspects: (i) as an online learning method, BPG autonomously explores and exploits the search space, adapts to the environment and improves its prediction strategy based on the information learned from past choices. It can reduce the decision latency caused by frequent handovers and converges to near-optimal solutions faster than other state-of-the-art beam selection methods; (ii) BPG simultaneously and separately learns the best beam selection strategy for different types of vehicles, thus it can deal with dynamic traffic patterns; (iii) BPG especially focuses on the communication stability requirement of the vehicular communication system, and strikes a good balance between maximizing the total throughput and minimizing the violation of average blockage-free guarantee; (iv) to eliminate the impact of delayed feedback, BPG transfers the original problem into a non-delayed setting by splitting time slots into consecutive and overlapped segments. The technical contributions of our algorithm BPG are summarized as follows:

First, we model the beam prediction problem for the target mmBS as an online learning problem under the 5G network framework. We propose a new contextual-combinatorial MAB framework, taking blockage-free requirements and delayed feedback into consideration. The proposed policy, BPG, successfully balances between maximizing the throughput and controlling the violation of blockage-free constraints. It can also make specific beam predictions under different traffic patterns and blockage situations. Our model can be adapted to various contexts and can be extended to fit more complicated constrained settings. Different from existing beam selection policies aiming to maximize the overall aggregated received data, our algorithm further tries to satisfy the minimum blockage-free guarantee threshold to keep vehicular communication stable.

Second, we tackle three main challenges when designing BPG: (i) we leverage the theory of Lagrangian method in constrained optimization and carefully utilize an adjustable penalty coefficient to balance between maximizing the throughput and lowering the violation; (ii) we partition the context space into sub-hypercubes based on the natural assumption that vehicles with similar contexts tend to reveal the same feedback. Each sub-hypercube maintains a set of weights for each beam, which can be used to calculate the *expected weight* under inferred traffic pattern. We obtain the selection probabilities based on the expected weights and use a dependent rounding algorithm to conduct beam predictions; (iii) by splitting the entire time span into consecutive overlapped time segments, the delayed problem can be transformed into a standard non-delayed setting.

Third, through rigorous theoretical analysis, we obtain sub-linear bounds for both *regret* and *violation* by taking blockage constraints into consideration. The sub-linear bounds suggest that our online learning algorithm makes asymptotically optimal predictions. We further conduct extensive simulation studies to verify the effectiveness of BPG by using real-world trace of taxi cabs in Rome [9]. The results show that BPG achieves 95% of the optimal throughput and always generates $25\% - 70\%$ less violations than other benchmark algorithms. In summary, BPG significantly outperforms other strategies and always shows better system performance in terms of maximizing the throughput while maintaining the lowest performance violation in vehicular communication systems.

In the rest of the paper, related work is reviewed in Sec. 2. We present the system model in Sec. 3. The detailed online beam prediction algorithm and the theoretical analysis are presented in Sec. 4. Sec. 5 is the simulation study and Sec. 6 concludes the paper.

2. Related Work

Beam Selection Problems. Traditional beam selection solutions rely heavily on accurate localization information and complex transceiver chain[10][11][12], resulting in high overhead and latency. Several learning-based architectures have been proposed recently for beam selection. Ali *et al.* [13] leverage out-of-band side information, such as similarity extracted from sub-6 GHz and mmWave, to help beam training. Va *et al.* [7] make beam alignment using a risk-aware online learning approach, which utilizes sensor data about position for beam selections. Wang *et al.* [5] leverage a situation-aware machine learning method to obtain beam information. Asadi *et al.* [14] propose an online learning method to solve the environment-aware beam-selection problem. However, the above literature fails to deal with requirements of system performance guarantee, and only focuses on maximizing the system throughput. Therefore, the above methods may lead to great violation of mmWave communication requirements, which severely harms the system performance in the long term.

Due to the high mobility of vehicles, the aforementioned methods can barely handle handover situations. Alkhateeb *et al.* [15] leverage a deep learning based beamforming method to enable highly-mobile vehicular systems. Mavromatis *et al.* [16] propose a MAC-layer based smart motion-prediction beam alignment algorithm. Given the CSI of source BS obtained previously, Chen *et al.* [6] leverage a sequence-to-sequence neural network to make beam predictions at target BS. Different from the above literature, our beam prediction framework can adapt to dynamic traffic systems. This is because our algorithm, BPG, can autonomously and separately learns beam prediction strategy for different types of vehicles from the environment each time.

Multi-armed bandit (MAB) optimization. MAB is a classic online learning method to address sequential decision making problems under partial feedback. The basic MAB focuses on learning how to choose the best single arm among a finite set of arms to maximize the total reward, without considering any constraints. Our bandit formulation in BPG is related to combinatorial-contextual bandits with delayed feedback. Especially, combinatorial bandits allows multiple plays each round [17][18], while contextual bandits considers context-dependent rewards [19][20]. Müller *et al.* [21] propose a bandit model combining contextual bandits together with combinatorial bandits. Nue *et al.* [22] show a multiplicative regret for the adversarial bandit with delayed feedback. Joulani *et al.* [23] provide a systematic algorithm framework to handle delayed feedback for different online learning algorithms. Chen *et al.* [24], rather recently, combine these concerns together.

However, the above algorithms are all formulated without considering communication stability. When applying MAB to realistic applications, communication QoS constraints cannot be neglected. Therefore, constrained bandits becomes another concern. Mahdavi *et al.* [25] design an efficient online learning algorithm under stochastic constraints. This algorithm leverages the theory of Lagrangian method in constrained optimization and takes both the regret and violation into consideration. Cai *et al.* [26] extend the former algorithm to the multi-play setting and propose a MAB framework with multi-level rewards for several network applications. Inspired but different from the literature above, our algorithm BPG extends the constrained combinatorial MAB into contextual and delayed settings, which is more suitable for vehicular communication system requirements, since we do need to consider vehicles with different types to adjust beam predictions. Our algorithm achieves near-to-optimal performance when applied in vehicular communication system, make a good balance between maximizing the aggregated received data and satisfying the system performance guarantee.

3. Problem Model

We consider a scenario where multiple mmBSs densely overlay the coverage area of a 5G NR gNB (New Radio gNodeB). The mmBSs are connected to the gNB via a backhaul link, and each mmBS's information can be learned from the gNB. Vehicles with high mobility arrive from time to time, coming across multiple coverage areas of neighbouring mmBSs. Each vehicle is equipped with two types of interfaces for communication: an NR interface connected to the gNB and an mmWave interface for ultra-low latency directional mmWave communication with mmBSs. Limited by physical constraints such as RF chains, mmBSs can simultaneously transmit a limited number of beams. Both the gNB and the mmBSs have no prior knowledge of their surroundings, the only information they have access to is vehicles' contexts (*i.e.,* directions and speeds) upon arrival. In this work, we focus on the downlink model, while the solution can also be adapted to the uplink situation, except that the mmBS needs to identify the owner of each signal it receives. Another problem in the uplink scenario is collision avoidance. If more than one user transmits in the same beam, then interference occurs. To avoid collision or to cancel interference is beyond the scope of the current paper.

Once the gNB detects that a group of vehicles are going to hand over from one mm-cell to another, it informs the source mmBS the target mmBS's beam information and current vehicle contexts. The source mmBS analyzes the future traffic pattern according to current vehicle contexts, and predicts the most promising beam directions for the target mmBS. In this way, the performance of target mmBS can be guaranteed especially during peak hours. After choosing specific beams according to the predictions, the target mmBS receives feedback from vehicles and sends it back to the source mmBS with a fixed delay (one time slot).

As shown in Fig. 1, the red car is going to hand over from the source mmBS to the target mmBS. The source mmBS sends predicted beam directions to target mmBS according to the red car's current context. After choosing the suggested beams, target mmBS sends the feedback to source mmBS in the next time slot.

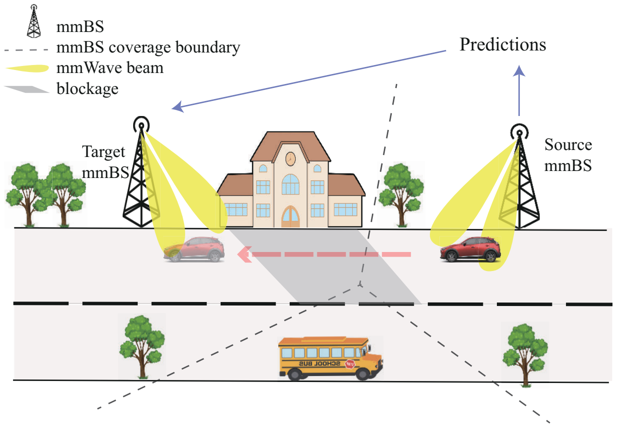

Figure 1: System model.

To better describe our problem, we assume that the target mmBS has M available distinct orthogonal beams, but can only simultaneously select m beams $(1 \leq m \leq M)$ each round due to physical constraints. Let $[X]$ denote the integer set $\{1, 2, \dots, X\}$. During system life span T, let I_t denote the set of selected beams at time $t \in [T]$, *i.e.,* $I_t \subseteq [M], |I_t| = m$. Let $\boldsymbol{p_t} = (p_1^t, ..., p_i^t, ..., p_M^t)$ be *chosen probability vector* for target mmBS's beams at time t. Here, p_i^t refers to the probability of selecting beam i at t. We use a dependent rounding technique [27] to choose m beams in time slot t, *i.e.,* $\mathbf{1}^\intercal \boldsymbol{p_t} = m$, where $\mathbf{1} = (1, \dots, 1)$. There are total N_t vehicles arriving at time t. Each vehicle $j \in [N_t]$ registers to the mmBS via the gNB, coming with its own context $x_{j,t}$. Let χ_t denote the context set received from vehicles at time t, *i.e.,* $\chi_t = \{\{x_{j,t}\}_{j=1,\dots,N_t}\}$, where each element $x_{j,t}$ is a vector with d_X dimensions taken from the bounded context space $\chi = [0, 1]^{d_X}$.

Suppose that each beam i is associated with an unknown *random process* $V_x^i(t), \forall t \in [T]$, which characterizes achievable throughput (without blockage) at each time slot t for each context $x \in \chi$. $V_x^i(t)$ is not necessarily stationary but bounded across i. Let $v_{i,x}^t$ be the realization of $V_x^i(t)$, which denotes the throughput of beam i at time t under context x. Due to the property that mmWave signals can be easily affected by static or moving blockages, only beams without blockage can successfully transmit data to vehicles. However, mmBSs have no knowledge of their surroundings (the pattern of arrivals and the probability of existence of blockage), thus for each beam i, it is also combined with an unknown *random process*, $U_x^i(t), \forall t \in [T]$, which characterizes blockage-free probability with context x at time t. Let $u_{i,x}^t$ be the realization of $U_x^i(t)$. Given that the environment is relatively stable, the traffic pattern is regular in the long run. We make an assumption that $U_x^i(t)$ is stationary. Suppose that $U_x^i(t)$ is stationary and independent across i with *unknown* mean $u_{i,x} = \mathbb{E}[U_x^i(t)]$, $\boldsymbol{u_x} = \{u_{1,x}, ..., u_{M,x}\}$. In order to guarantee the system performance, there exists a minimum blockage-free threshold $\rho > 0$. For each context $x \in \chi_t$, the average of the sum of blockage-free probability needs to be above this threshold, *i.e.,* $\mathbb{E}[\boldsymbol{u_x^\intercal p_t}] > \rho$. We normalize $U_x^i(t) \in [0, 1]$ and $V_x^i(t) \in [0, 1]$ and assume that they are independent of each other. Therefore, the successful throughput at time t with context x is characterized by a *compound throughput vector* $\boldsymbol{g}_x^t = \{g_{1,x}^t, ..., g_{M,x}^t\}$, where each compound throughput $g_{i,x}^t$ of a beam i at time t is generated by the random process $G_x^i(t) = U_x^i(t)V_x^i(t)$, thus $g_{i,x}^t = u_{i,x}^t v_{i,x}^t$. At time t, the expected total compound throughput is $\mathbb{E}[\sum_{x\in\chi_t} \boldsymbol{g_x^{t\intercal} p_t}]$. Since the source mmBS can only observes the performance of predictions at the end of the next time slot, one-slot delay feedback exists.

Specifically, the offline optimization problem is formulated into a linear program (LP) as follows:

$$\text{maximize} \sum_{t\in[T]} \sum_{x\in\chi_t} \sum_{i\in[M]} g_{i,x}^t p_i^t \tag{1}$$

subject to:

$$\sum_{i\in[M]} p_i^t = m, \forall t \in [T], \tag{1a}$$

$$\sum_{i\in[M]} u_{i,x}^t p_i^t > \rho, \forall t \in [T], \forall x \in \chi_t, \tag{1b}$$

$$p_i^t \in [0, 1], \forall i \in [M], t \in [T]. \tag{1c}$$

Note that in the online setting, LP (1) is nontrivial to solve since feedback can be observed only when beams are selected, which means the value of $u_{i,x}^t$ and $v_{i,x}^t$ are observed one time-slot later. Our model can be seen as an online learning problem, in which data becomes available in a sequential order and is used to update the decision policy for future data during each time slot.

The fundamental problem for each source mmBS is to learn how to efficiently predict a subset of m beams for target mmBS, taking vehicles' contexts, performance constraints and delayed feedback into consideration. Therefore, we formulate the beam prediction problem into a *contextual-combinatorial MAB* problem with *performance constraints* and *delayed feedback*. Selecting beams with large throughput helps to achieve better network performance, and choosing blockage-free beams helps to maintain stable vehicular communication.

We consider the optimal solution of LP (1) as an *oracle*, in which we know all the information in advance and make the best selection each round from God's perspective. However, learning a policy to maximize the compound reward is challenging in realistic settings without full knowledge in prior. We change our goal into designing a beam prediction policy π, which updates the beam chosen vectors $\boldsymbol{p_t}$ for target mmBS, such that the *regret*, which is referred to a loss compared to the *Oracle*, is as small as possible. The regret for a learning policy π is defined as follows:

$$R(T) = \max_{\boldsymbol{u_x^\intercal p_t} > \rho} \sum_{t\in[T]} \sum_{x\in\chi_t} \boldsymbol{g_x^{t\intercal} p_t} - \mathbb{E}[\sum_{t\in[T]} \sum_{x\in\chi_t} \boldsymbol{g_x^{t\intercal} p_t^\pi}]. \tag{2}$$

where p_t denotes the optimal beam selection vector while p_t^π denote the beam prediction decision made by our policy π. In order to achieve close-to-oracle reward, the source mmBS needs to balance between *exploration* and *exploitation* in each time slot to learn from each choice to improve its policy. Note that during early time exploration, source mmBS might make beam predictions that violate the blockage constraint, since it has little knowledge of its surroundings. If the source mmBS violates the constraint during exploration, a low throughput or even zero throughput may be obtained. To make sure that the total violation of constraints after T time slots as small as possible, we define the violation as follows:

$$V(T) = \mathbb{E}[\sum_{t\in[T]} \sum_{x\in\chi_t} (\rho - \boldsymbol{u_x^\intercal p_t^\pi})]^+, \tag{3}$$

where $[x]^+ = \max\{0, x\}$. A lower regret means that π gets closer to the Oracle. A smaller violation means that π becomes better in satisfying the constraint as time t increases. We next design an algorithm in order to balance between the regret and the violation.

4. An Online Learning Algorithm

In this section, we design an online learning algorithm BPG to predict beam directions. We first

Table 1: Notation

M	# of beams	T	# of time slots
m	# of selected beams	X	integer set $\{1, 2, ...X\}$
N_t	# of vehicles at time t	I_t	beam set selected at t
d_X	# of context dimensions	P_T	# of sub-hypercubes
n_T	# of parts each dimension can be divided into		
p_i^t	chosen probability of beam i at t		
$x_{j,t}$	vehicle j's context at t		
χ_t	context set received from vehicles at time t		
$V_x^i(t)$	random process of beam i at t which characterizes varying throughput		
$U_x^i(t)$	random process of beam i at t which characterizes blockage-free probability		
$G_x^i(t)$	$= U_x^i(t)V_x^i(t)$, random process of compound throughput		
$v_x^i(t)$	throughput of beam i at t		
$u_x^i(t)$	blockage-free probability of beam i at t		
$g_x^i(t)$	compound throughput of beam i at t		
ρ	minimum blockage-free threshold		
$w_{i,h}^t$	weight for beam i in sub-hypercube h at t		
h_j^t	sub-hypercube which vehicular j belongs to at t		
$\mathcal{H}_t$	set of sub-hypercubes at t		
$\boldsymbol{\xi^t}$	collection of arrival distributions of hypercube h at t		
ξ_h^t	the percentage of contexts which belong to hypercube h at t		
S_t^h	set of weights which need to be reduced in sub-hypercube h at t		
λ_t^h	Lagrange multiplier in sub-hypercube h at t		

present several algorithm design challenges and propose corresponding solutions in Sec. 4.1, and we elaborate the details of the algorithm design in Sec. 4.2. Finally, we carry out rigorous theoretical analysis in Sec. 4.3.

4.1. Algorithm Design Challenges

Towards designing the beam prediction policy, a key challenge exists. *How to balance between maximizing the aggregated received data and satisfying the performance constraint?* Inspired by [25][26], we leverage the theory of Lagrangian method in constrained optimization. Our target is to minimize a modified regret function with the component of violation and an adjustable Lagrange multiplier $\lambda(T)$. If the constraint is substantially violated, our algorithm places more weight on the Lagrange multiplier $\lambda(T)$; it lowers the weight when the constraint is satisfied reasonably. Our algorithm introduces a sub-linear bound for the modified regret function as follows:

$$R(T) + \lambda(T)(V(T))^2 \leq T^{1-\epsilon}, 0 < \epsilon < 1. \qquad (4)$$

Since $-R(T) \leq O(mT)$ for any policy π, we derive a bound for $R(T)$ and $V(T)$:

$$R(T) \leq O(T^{1-\epsilon}), V(T) \leq \sqrt{O(T^{1-\epsilon} + mT)/\lambda(T)}, \qquad (5)$$

We can achieve sub-linear bounds for both $R(T)$ and $V(T)$ if we properly adjust $\lambda(T)$.

Considering the basic idea of our algorithm is to choose m beams according to the chosen probability vector $\boldsymbol{p_t}$, the second challenge is *how to update each time slot*. Our solution is to calculate the chosen probabilities by maintaining a set of M weights for M beams. Due to the fact that vehicles with different contexts tend to have different beam preference, even the same beam's weight varies when facing different contexts. *How to set and update the weights for M beams each time* becomes another challenge. A straightforward approach to tackle this problem is to maintain separate weights for different contexts. However, considering the large amount of contexts, this method incurs high computation complexity. In order to deal with the large context space, we propose two basic assumptions: (i) vehicles with similar contextual information reveal similar feedback under the same condition; (ii) vehicles with different contexts prefer different sets of beams. Under these assumptions, similarities between contexts can be exploited online for future predictions.

Our algorithm BPG partitions the large context space into uniform sub-hypercubes, each representing a certain type of vehicles. Each sub-hypercube in our algorithm maintains a set of weights for M beams, which can be used to compute the chosen probabilities under specific context and will be modified in each time slot according to hitherto feedback. The accuracy of weights for each sub-hypercube increases as time elapses. However, since we need to take the combination of all the arrival contexts into consideration, we propose an *expected combination hypercube*, which maintains a set of *expected weights* for M beams. The expected weights are calculated by current context distribution (traffic pattern) and corresponding sub-hypercubes's weights (see Fig. 3). This is easy to understand, since the more a certain type of vehicles come, the more biased it is to choose preferred beam directions for that type. Based on the expected weights, our policy calculates the probabilistic distribution $\tilde{\boldsymbol{p}}_t$ for M arms. According to $\tilde{\boldsymbol{p}}_t$, BPG selects m beams from M beams using the dependent rounding method. Then the feedback of chosen beams are used to modify each sub-hypercube's weights. Finally, we use calculated weights to compute each beam's chosen probability.

Traditional MAB problems receive feedback (reward) at the current time slot. In contrast, in our beam prediction problem, *dealing with the delayed feedback* becomes the last challenge. Rewards and violations can be observed only one time slot later, which will influence the analysis of the regret bound and violation bound. Nonetheless, we can transform the delayed setting into non-delayed setting by combining two continuous time slots into one *compound slot*. We split T time slot sequence into $\lceil K(T) \rceil$ consecutive overlapped segments $\{[t'_n, t'_{n+2})\}_{n=1}^{\lceil K(T) \rceil}$,$n \in \mathbb{N}^+$, and $K(T) = t^{\frac{2}{3+d_X}} log(t)$, $\lceil K(T) \rceil = n$ for any $t \in [t'_n, t'_{n+2})$. In this way, this problem can be seen as a standard non-delayed problem, where we make beam predictions at t'_n and receive feedback at the end of t'_{n+1}, $n \in \lceil K(T) \rceil$.

4.2. Algorithm Design

The general idea of our algorithm BPG is as follows: First, we partition the context space into uniform sets of similar contexts (sub-hypercubes). In each time period, BPG observes the arriving vehicles' contexts and classifies them into different sub-hypercubes. The algorithm then calculates the expected weights according to the context distribution and corresponding sub-hypercube's weights. The chosen probability vector $\boldsymbol{p_t}$ is computed based on the expected weights. Finally, BPG updates each sub-hypercube's weights and Lagrange multiplier according to the delayed feedback.

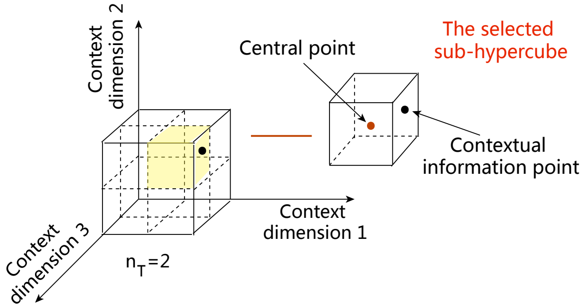

Figure 2: Context Partion Model.

The details of BPG are shown in Alg. 1. In its initialize phase, BPG partitions the contextual space $\chi = [0,1]^{d_X}$ into $(n_T)^{d_X}$ hypercubes. Each sub-hypercube has an identical size $\frac{1}{(n_T)^{d_X}}$, where n_T is an input constant which determines the number of parts each dimension can be divided into (see Fig. 2). During each time slot, we first observe the number of arriving vehicles, N_t, and their contextual information $\{x_{j,t}\}_{j \in N_t}$. For each context $x_{j,t}$, BPG determines which hypercube it belongs to. As shown in Fig. 3, we will use the center point x^h to represent all context information points in each sub-hypercube $h \in P_T$ for computation. Let $\mathcal{H}_t = \{h_{j,t}\}_{j \in N_t}$ denote the collection of hypercubes at time t. Since we need to choose the beams given the combination of all these contexts, we also observe the arrival distribution of each sub-hypercubes $\boldsymbol{\xi^t} = \{\xi_h^t\}_{h \in \mathcal{H}_t}$, where $\xi_h^t = \frac{|\sum_{j:\{x_{j,t} \in h\}}|}{N_t}$ denotes the percentage of contexts which belong to hypercube h. (Lines 4-5). For each sub-hypercube $h \in P_T$, BPG maintains a set of weights for M beams respectively, *i.e.*, $\boldsymbol{w_h^t} = \{w_{1,h}^t, ..., w_{M,h}^t\}$, which are initialized into $\mathbf{1}$ at the beginning.

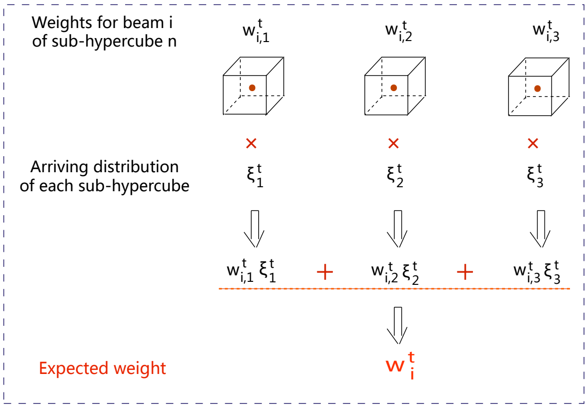

Figure 3: Expected Weight Model.

Since we need to choose m beams for the combination of these vehicles with different contexts together, we transform the set of hypercubes into one *expected hypercube*. As can be seen in Fig. 3, the expected weights for M beams are calculated following $w_i^t = \sum_{h \in \mathcal{H}_t} \xi_h^t w_{i,h}^t$. Let $\boldsymbol{w}_t = \{w_1^t, w_2^t, ..., w_M^t\}$ denote the set of expected weights for M arms at time t, which is used to calculate the chosen probability vector $\boldsymbol{\tilde{p}_t}$. At line 17, we show the similar idea to Exp3 [28] for exploration and exploitation by using $(1-\gamma)\tilde{w}_i^t / \sum_{i=1}^M \tilde{w}_i^t$ for exploitation and γ/M for exploration. Lines 6- 14 guarantee that the probabilities in $\boldsymbol{\tilde{p}_t}$ are less than or equal to 1.

At line 19, we utilize a dependent rounding algorithm (Alg. 2) to update $\boldsymbol{\tilde{p}_t}$ until $\mathbf{1}^\intercal \boldsymbol{\tilde{p}_t} = m, p_i^t \in \{0,1\}$. Since the feedback of predictions can only be obtained after one time slot, we obtain each vehicle's feedback $u_{i,h_{j,t}}^{t-1}$ and $v_{i,h_{j,t}}^{t-1}$ at time $t(t \geq 2)$. For each sub-hypercube, we calculate the unbiased estimates of $\hat{u}_{i,h}^t$ and $\hat{g}_{i,h}^t$ for each arm $i \in M$ at line 24, where $u_{i,x^h}^t = (\sum_{j:x_{j,t}\in h} u_{i,x_{j,t}}^t)/\sum_{j:x_{j,t}\in h} \cdot 1$ and $v_{i,x^h}^t = (\sum_{j:x_{j,t}\in h} v_{i,x_{j,t}}^t)/\sum_{j:x_{j,t}\in h} \cdot 1$. It is easy to verify that $\mathbb{E}[\hat{u}_{i,h}^t] = u_{i,x^h}^t$ and $\mathbb{E}[\hat{g}_i^t] = u_{i,x^h}^t v_{i,x^h}^t$. Let $\mathbb{1}(K)$ be the indicator function: $\mathbb{1}(K) = 1$ if the event K happens and 0 otherwise. Finally, for each sub-hypercube h, the weight vector $\{w_{i,h}^t\}_{i\in M}$ and the Lagrange multiplier λ_t^h are updated using the estimations at the end of each iteration as shown in lines 26-27.

Algorithm 1 Online Beam Direction Prediction with Performance Guarantee: BPG

```
1: Initialize context partition: Create partition $P_T$ of context space into $(n_T)^{d_X}$ hypercubes of identical size
2: Initialize $\boldsymbol{w_h^1} = \boldsymbol{1}, \lambda_1 = 0, \rho > 0, \beta = (\frac{1}{m} - \frac{\gamma}{M})(1-\gamma), \eta = \frac{\gamma\delta m}{(\delta+m)M}, \delta = \frac{4(e-2)\gamma m}{1-\gamma}, S_t^h = 0, I_t = 0$
3: for $t = 1, 2, ..., T$ do
4:     Observe vehicle contexts set $\chi_t = \{x_{j,t}\}_{j=1,...,N_t}$
5:     Find $\mathcal{H}_t = \{h_{j,t}\}_{j=1,...,N_t}$ such that $x_{j,t} \in h_{j,t} \in P_T$, observe $\boldsymbol{\xi}^t = \{\xi_h^t\}_{h\in\mathcal{H}_t}$
6:     for each sub-hypercube $h \in \mathcal{H}_t$ do
7:         if $\max_{i\in M} w_{i,h}^t \geq \beta \sum_{i=1}^M w_{i,h}^t$ then
8:             Find $\alpha_{h,t}$ such that $\alpha_{h,t}/\sum_{i=1,w_{i,h}^t\geq\alpha_{h,t}}^M \alpha_{h,t} + \sum_{i=1,w_{i,h}^t\leq\alpha_{h,t}}^M w_{i,h}^t = \beta$
9:             $S_t^h = \{i : w_{i,h}^t \geq \alpha_{h,t}\}$
10:        end if
11:        for $i = 1, ..., M$ do
12:            $\tilde{w}_{i,h}^t = \alpha_{h,t}$ if $i \in S_t^h$, otherwise, $\tilde{w}_{i,h}^t = w_{i,h}^t$
13:        end for
14:    end for
15:    The expected weight $\tilde{w}_i^t = \sum_{h\in\mathcal{H}_t} \xi_h^t \tilde{w}_{i,h}^t$
16:    for $i = 1, ..., M$ do
17:        $\tilde{p}_i^t = m[(1-\gamma)\tilde{w}_i^t/\sum_{i=1}^M \tilde{w}_i^t + \gamma/M]$
18:    end for
19:    $I_t = DRA(m, \tilde{\boldsymbol{p}}_t)$
20:    if $t \geq 2$ then
21:        Receive vehicles' rewards $\boldsymbol{u}_{t-1}$ and $\boldsymbol{v}_{t-1}$
22:        for each hypercube $h \in \mathcal{H}_{t-1}$ do
23:            for each beam $i \in M$ do
24:                $\hat{u}_{i,h}^{t-1} = u_{i,x^h}^{t-1}/\tilde{p}_i^{t-1}\mathbb{1}(i \in I_{t-1})$,
                   $\hat{g}_{i,h}^{t-1} = (u_{i,x^h}^{t-1} v_{i,x^h}^{t-1})/\tilde{p}_i^{t-1}\mathbb{1}(i \in I_{t-1})$
25:            end for
26:            $w_{i,h}^{t+1} = \begin{cases} w_{i,h}^t & i \in S_t^h \\ w_{i,h}^t \circ \exp(\eta(\hat{g}_{i,h}^{t-1} + \lambda_t \hat{u}_{i,h}^{t-1})) & i \notin S_t^h \end{cases}$
27:            $\lambda_{t+1}^h = [(1-\delta\eta)\lambda_t^h - \eta(\frac{\hat{u}_h^{t-1}\tilde{p}_{t-1}}{1-\gamma} - \rho)]^+$
28:        end for
29:    end if
30: end for
```

Algorithm 2 Dependent Rounding Algorithm: DRA

```
Initialize $I_t = \varnothing$
while exist $i \wedge p_i^t \in (0,1)$ do
    Randomly select two jobs $i_1, i_2, i_1 \neq i_2$, $p_{i_1} p_{i_2} \in (0,1)$
    Define $\Psi_1 \triangleq \min\{1 - p_{i_1}^t, p_{i_2}^t\}$
    Define $\Psi_2 \triangleq \min\{p_{i_1}^t, 1 - p_{i_2}^t\}$
    With probability $\frac{\Psi_2}{\Psi_1+\Psi_2}$, set
    $p_{i_1}^t = p_{i_1}^t + \Psi_1, p_{i_2}^t = p_{i_2}^t - \Psi_1$
    With probability $\frac{\Psi_1}{\Psi_1+\Psi_2}$, set
    $p_{i_1}^t = p_{i_1}^t - \Psi_2, p_{i_2}^t = p_{i_2}^t + \Psi_2$
end while
return $I_t = \{i \in [M] : p_i^t = 1\}$
```

4.3. Regret Analysis

The theorem below shows that both the regret and violation of BPG are sub-linear in the time horizon T, *i.e.*, $\lim_{T\to\infty} \frac{R(T)}{T} = 0$ and $\lim_{T\to\infty} \frac{V(T)}{T} = 0$. It guarantees that the algorithm converges to the optimal beam prediction policy over time, and has a near-to-optimal performance when T is large enough. The regret and violation of BPG can be bounded as follows:

Theorem 1. *Let* $\eta = \frac{\gamma\delta m}{(\delta+m)M}$, $n_T = \lceil T^{\frac{1}{9d_X}} \rceil$ *and* $\gamma = \min(1, \sqrt{\frac{2(e-2)M+Mm}{mln(M/m)T^{1/2}}})$. *By running the policy* $\tilde{\pi}$, *we achieve sub-linear bounds for both the regret and violation as follows:*

$$\begin{aligned} R_D(T) &\leq O(Ld_X^{\frac{\alpha}{2}} mM \ln(M) T^{\frac{5}{6} - \frac{\alpha}{6d_X}}) \\ V_D(T) &\leq O(L^{\frac{1}{2}} d_X^{\frac{\alpha}{4}} m^{\frac{1}{2}} M^{\frac{1}{2}} T^{\frac{5}{6}}). \end{aligned} \tag{6}$$

Proof. We first consider the regret and violation bound in one hypercube $h \in P_T$, without considering the influence of delayed feedback, and then obtain its summation over the number of sub-hypercubes $(n_T)^{d_X}$. Finally, we discuss the influence of the bound due to delayed feedback.

Before considering the upper-bound in each hypercube, we use the Lipschitz condition to define the dissimilarity between different contexts:

Assumption 1. *There exists constant* $L > 0$ *such that for all context* $x_i, x_j \in \chi$, *we have* $D_X(x_i, x_j) \leq L||x_i - x_j||^\alpha$, *where* $||\,\boldsymbol{.}\,||$ *denoted the Euclidian norm in* $\mathbb{R}^{d_X}$.

Note that the Lipschitz constants L and the similarity degree parameter α are not required to be known by our prediction algorithms. They will only be used when quantifying BPG' performance and both of them will appear in our regret and violation bounds.

Lemma 1. *For each sub-hypercube h, let $\hat{\boldsymbol{r}}_h^t = \hat{\boldsymbol{g}}_h^t + \lambda_t^h \hat{\boldsymbol{u}}_h^t$, where $\hat{\boldsymbol{g}}_h^t, \hat{\boldsymbol{u}}_h^t \in \mathbb{R}_+^M$, $w_{t+1}^h = w_t^h \cdot \exp(\eta \hat{\boldsymbol{r}}_t^h)$. Let $\boldsymbol{p}_t$ be an arbitrary probabilistic selection vector which satisfied $p_i^t \in [0,1], i \in [M], \mathbf{1}^\intercal \boldsymbol{p}_t = m, \boldsymbol{u}_h^{t\intercal} \boldsymbol{p}_t \geq \rho$. Let $\tilde{\boldsymbol{p}}_t$ denote the chosen probability vector of the prediction policy $\tilde{\pi}$ at time t, we can get the following inequality:*

$$\begin{aligned}
&\mathbb{E}[\sum_{t=1}^T \hat{\boldsymbol{r}}_h^{t\intercal} \boldsymbol{p}_t - \frac{1}{1-\gamma} \sum_{t=1}^T \hat{\boldsymbol{r}}_h^{t\intercal} \tilde{\boldsymbol{p}}_t] \\
\leq& \frac{m}{\eta} \ln \frac{M}{m} + \frac{2(e-2)\eta M}{1-\gamma} T + \frac{2(e-2)\eta M}{1-\gamma} \sum_{t=1}^T (\lambda_t^h)^2. \quad (7)
\end{aligned}$$

Proof. Let $\lambda_{t+1}^h = [(1-\delta\eta)\lambda_t^h - \eta(\frac{(\hat{\boldsymbol{u}}_h^{t-1})^\intercal \tilde{\boldsymbol{p}}_i^{t-1}}{1-\gamma} - \rho)]^+ \leq [(1-\delta\eta)\lambda_t^h + \eta\rho]_+$, by induction on λ_t^h, we can get $\lambda_t^h \leq \frac{\rho}{\delta}$. we first show an upper-bound and a lower bound of $\sum_{t=1}^T \ln(\frac{W_{t+1}^h}{W_t^h})$ for the sequence of selected I_t at $t = 1, ..., T$.

$$\begin{aligned}
&\sum_{t=1}^T \ln(\frac{W_t^h}{W_t^h}) = \ln(\frac{W_t^h}{W_1}) = \ln \sum_{i=1}^M w_{i,h}^{T+1} - \ln M \\
\geq& \ln \sum_{i=1}^M p_i^t w_{i,h}^{T+1} - \ln M \geq \sum_{i=1}^M \frac{p_i^t}{m} \sum_{t=1}^T \eta \hat{r}_{i,h}^t - \ln \frac{M}{m} \\
=& \frac{\eta}{m} \sum_{i=1}^M p_i^t \sum_{t:i \notin S_t} \hat{r}_{i,h}^t - \ln \frac{M}{m}. \quad (8)
\end{aligned}$$

Since $\eta = \frac{\gamma\delta m}{(\delta+m)M}$, $\lambda_t^h \leq \frac{\rho}{\delta}$, we have $\eta \hat{r}_{i,h}^t \leq 1$. According to the fact that $e^x \leq 1 + x + (e-2)x^2$ for $x \leq 1$, we have

$$\begin{aligned}
&\frac{W_t^h}{W_t^h} = \sum_{i \in M/S_t^h} \frac{w_{i,h}^{t+1}}{W_t^h} + \sum_{i \in S_t^h} \frac{w_{i,h}^{t+1}}{W_t^h} \\
=& \sum_{i \in M/S_t^h} \frac{w_{i,h}^t \exp(\eta \hat{r}_{i,h}^t)}{W_t^h} + \sum_{i \in S_t^h} \frac{w_{i,h}^t}{W_t^h} \\
\leq& \sum_{i \in M/S_t^h} \frac{w_{i,h}^t}{W_t^h}(1 + \eta \hat{r}_{i,h}^t + (e-2)(\eta \hat{r}_{i,h}^t)^2) + \sum_{i \in S_t^h} \frac{w_{i,h}^t}{W_t^h} \\
=& \frac{\tilde{W}_t^h}{W_t^h}(\sum_{i \in M/S_t^h} \frac{w_{i,h}^t}{\tilde{W}_t^h} + \sum_{i \in S_t^h} \frac{w_{i,h}^t}{\tilde{W}_t^h}) + \frac{\tilde{W}_t^h}{W_t^h} \sum_{i \in S_t^h} \frac{w_{i,h}^t}{\tilde{W}_t^h} \eta \hat{r}_{i,h}^t \quad (9) \\
&+ \frac{\tilde{W}_t^h}{W_t^h}(e-2)(\eta \hat{r}_{i,h}^t)^2 \\
\leq& 1 + \frac{1}{m(1-\gamma)} \sum_{i \in M/S_t^h} \eta \hat{r}_{i,h}^t + \frac{e-2}{m(1-\gamma)} \sum_{i \in M/S_t^h} \eta^2 (\hat{r}_{i,h}^t)^2.
\end{aligned}$$

Due to the fact that $\ln(1+x) \leq x$ and $\tilde{p}_i^t \hat{r}_{i,h}^t = \hat{r}_{i,h}^t \leq 1 + \lambda_t^h$, we can transform the above inequality as follows:

$$\ln \frac{W_{t+1}^h}{W_t^h} \leq \frac{\eta}{m(1-\gamma)} \sum_{i \in M/S_t^h} \tilde{p}_i^t \hat{r}_{i,h}^t + \frac{(e-2)\eta^2}{m(1-\gamma)} \sum_{i \in M/S_t^h} (1+\lambda_t^h) \hat{r}_{i,h}^t.$$

Simultaneously summing t on both sides of the above inequality and combining with (8), we have

$$\begin{aligned}
&\frac{\eta}{m} \sum_{i=1}^M p_i^t \sum_{t:i \notin S_t} \hat{r}_{i,h}^t - \ln \frac{M}{m} \\
\leq& \sum_{t=1}^T (\frac{\eta}{m(1-\gamma)} \sum_{i \in M/S_t} \tilde{p}_i^t \hat{r}_{i,h}^t + \frac{(e-2)\eta^2}{m(1-\gamma)} \sum_{i \in M/S_t} (1+\lambda_t) \hat{r}_{i,h}^t). \quad (10)
\end{aligned}$$

Since $\tilde{p}_i^t = 1$ for $i \in S_t^h \subseteq I_t$, we have $\sum_{i=1}^M p_i^t \sum_{t:i \in S_t^h} \hat{r}_{i,h}^t \leq \frac{1}{1-\gamma} \sum_{t=1}^T \sum_{i \in S_t^h} \hat{r}_{i,h}^t$. Combining this inequality with (10) and multiplying both sides with $\frac{m}{\eta}$, we can get

$$\begin{aligned}
&\sum_{t=1}^T (\hat{\boldsymbol{r}}_h^t)^\intercal \boldsymbol{p}_t - \frac{m}{\eta} \ln \frac{M}{m} \\
\leq& \frac{\sum_{t=1}^T (\hat{\boldsymbol{r}}_h^t)^\intercal \tilde{p}_t}{1-\gamma} + \frac{(e-2)\eta}{1-\gamma} \sum_{t=1}^T \sum_{i=1}^M (1+\lambda_t^h) \hat{r}_{i,h}^t.
\end{aligned}$$

Taking expectation on both sides, we can get

$$\begin{aligned}
&\mathbb{E}[\sum_{t=1}^T (\hat{\boldsymbol{r}}_h^t)^\intercal \boldsymbol{p}_t - \frac{1}{1-\gamma} \sum_{t=1}^T (\hat{\boldsymbol{r}}_h^t)^\intercal \boldsymbol{p}_t] \\
\leq& \frac{m}{\eta} \ln \frac{M}{m} + \frac{(e-2)\eta}{1-\gamma} \mathbb{E}[\sum_{t=1}^T \sum_{i=1}^M (1 + \lambda_t^h \lambda_t^h) \hat{r}_{i,h}^t] \\
\leq& \frac{m}{\eta} \ln \frac{M}{m} + \frac{(e-2)\eta}{1-\gamma} \mathbb{E}[\sum_{t=1}^T \sum_{i=1}^M (1+\lambda_t^h)(g_{i,h}^t + \lambda_t^h u_{i,h}^t)] \\
\leq& \frac{m}{\eta} \ln \frac{M}{m} + \frac{2(e-2)\eta M}{1-\gamma} T + \frac{2(e-2)\eta M}{1-\gamma} \sum_{t=1}^T (\lambda_t^h)^2,
\end{aligned}$$

where the last inequality holds due to the fact that $g_i^t, u_i^t \leq 1$ and $(x+y)^2 \leq 2x^2 + 2y^2$, which means $\mathbb{E}[\sum_{i=1}^M (1+\lambda_t^h)(g_i^t + \lambda_t^h u_i^t)] \leq M(1+\lambda_t^h)^2 \leq 2M + 2M(\lambda_t^h)^2$. □

Lemma 2. *In each sub-hypercube h, let $f_{h,t}(\lambda) = \frac{\delta}{2}\lambda^2 + \lambda(\frac{(\hat{\boldsymbol{u}}_h^t)^\intercal \tilde{\boldsymbol{p}}_t}{1-\gamma})$, $\lambda_{t+1}^h = [\lambda_t^h - \eta \nabla f_{h,t}(\lambda_t^h)]_+$, and $\lambda_1 = 0$. Assuming $\eta = \frac{\gamma\delta m}{\delta+m}$, $\frac{(\hat{\boldsymbol{u}}_h^t)^\intercal \tilde{\boldsymbol{p}}_t}{1-\gamma} \geq \rho$, we can get*

$$\begin{aligned}
&\mathbb{E}[\frac{\delta}{2} \sum_{t=1}^T ((\lambda_t^h)^2 - \lambda^2) + \sum_{t=1}^T (\lambda_t^h - \lambda)(\frac{(\hat{\boldsymbol{u}}_h^t)^\intercal \tilde{\boldsymbol{p}}_t}{1-\gamma} - \rho)] \\
\leq& \frac{\lambda^2}{2\eta} + (m^2 + \frac{\eta m M}{(1-\gamma)^2}) \eta T. \quad (11)
\end{aligned}$$

Proof. $\lambda_{t+1}^h = [\lambda_t^h - \eta\nabla f_{h,t}(\lambda_t^h)]_+ = [(1-\delta\eta)\lambda_t^h - \eta(\frac{(\hat{\boldsymbol{u}}_h^t)^\intercal \tilde{\boldsymbol{p}}_t}{1-\gamma} - \rho)]_+ \leq [(1-\delta\eta)\lambda_t^h + \eta\rho]_+$. For arbitrary λ, we have

$$
\begin{aligned}
&(\lambda_{t+1}^h - \lambda) = [\lambda_t^h - \eta(\delta\lambda_t^h + \frac{(\hat{\boldsymbol{u}}_h^t)^\intercal \tilde{\boldsymbol{p}}_t}{1-\gamma} - \rho) - \lambda]^2 \\
\leq&(\lambda_t^h - \lambda)^2 + (\eta(\delta\lambda_t^h - \rho) + \eta\frac{(\hat{\boldsymbol{u}}_h^t)^\intercal \tilde{\boldsymbol{p}}_t}{1-\gamma})^2 \\
&- 2(\lambda_t^h - \lambda)(\eta\nabla f_{h,t}(\lambda_t^h)) \\
\leq&(\lambda_t^h - \lambda)^2 + 2\eta^2\rho^2 + 2\eta^2(\frac{(\hat{\boldsymbol{u}}_h^t)^\intercal \tilde{\boldsymbol{p}}_t}{1-\gamma})^+ 2\eta(f_{h,t}(\lambda) - f_{h,t}(\lambda_t^h)).
\end{aligned}
$$

Then, by rearranging the terms we get

$$
\begin{aligned}
&f_{h,t}(\lambda_t^h) - f_{h,t}\lambda \\
\leq&\frac{1}{2\eta}[(\lambda_{t+1}^h - \lambda)^2 - (\lambda_t^h - \lambda)^2] + \eta(\rho^2 + (\frac{(\hat{\boldsymbol{u}}_h^t)^\intercal \tilde{\boldsymbol{p}}_t}{1-\gamma})^2) \\
\leq&\frac{1}{2\eta}[(\lambda_{t+1}^h - \lambda)^2 - (\lambda_t^h - \lambda)^2] + \eta m^2 + \frac{\eta m^2}{(1-\gamma)^2 m}\sum_{i=1}^{M}(\tilde{p}_i^t \hat{u}_{i,h}^t)^2 \\
\leq&\frac{1}{2\eta}[(\lambda_{t+1}^h - \lambda)^2 - (\lambda_t^h - \lambda)^2] + \eta m^2 + \frac{\eta m}{(1-\gamma)^2}\sum_{i=1}^{M}\hat{u}_{i,h}^t.
\end{aligned}
$$

Taking expectation over $\sum_{t=1}^T[f_{h,t}\lambda_t^h - f_{h,t}(\lambda)]$, then Lemma 2 holds. □

Applying $\boldsymbol{r}_h^t = \boldsymbol{g}_h^t + \lambda_t^h \boldsymbol{u}_h^t$ to (7), and combining (7) and (11) together gives

$$
\begin{aligned}
&\sum_{t=1}^T (\boldsymbol{g}_h^t)^\intercal \boldsymbol{p}_t - \frac{1}{1-\gamma}\mathbb{E}[\sum_{t=1}^T (\boldsymbol{g}_h^t)^\intercal \tilde{\boldsymbol{p}}_t] + \mathbb{E}[\lambda\sum_{t=1}^T(\rho - \frac{(\boldsymbol{g}_h^t)^\intercal \boldsymbol{p}_t}{1-\gamma})] \\
&- \mathbb{E}[(\frac{\delta T}{2} + \frac{1}{2\eta})\lambda^2] \\
\leq&\frac{m}{\eta}\ln\frac{M}{m} + \frac{2(e-2)\eta MT}{1-\gamma} + \eta m^2 T + \frac{\eta mMT}{(1-\gamma)^2} \\
+&(\frac{2(e-2)\eta M}{1-\gamma} - \frac{\delta}{2})\sum_{t=1}^T(\lambda_t^h)^2 + \mathbb{E}[\sum_{t=1}^T \lambda_t^h(\rho - \frac{(\hat{\boldsymbol{u}}_h^t)^\intercal \tilde{\boldsymbol{p}}_t}{1-\gamma})].
\end{aligned}
$$

Since $\eta = \frac{\gamma\delta m}{(\delta+m)M}$ and $\delta = \frac{4(e-2)\gamma m}{1-\gamma} \geq \frac{4(e-2)\gamma m}{1-\gamma} - m$, thus we have $\frac{2(e-2)\eta M}{1-\gamma} - \frac{\delta}{2} \leq 0$. Also, $\frac{(\hat{\boldsymbol{u}}_h^t)^\intercal \tilde{\boldsymbol{p}}_t}{1-\gamma} \geq \rho$. Multiplying both sides with $(1-\gamma)$, we have

$$
\begin{aligned}
&(1-\gamma)\sum_{t=1}^T (\boldsymbol{g}_h^t)^\intercal \boldsymbol{p}_t - \mathbb{E}[\sum_{t=1}^T (\boldsymbol{g}_h^t)^\intercal \tilde{\boldsymbol{p}}_t] - \mathbb{E}[(\frac{\delta T}{2} + \frac{1}{2\eta})\lambda^2] \\
&+ \mathbb{E}[\lambda\sum_{t=1}^T((1-\gamma)\rho - (\boldsymbol{g}_h^t)^\intercal \boldsymbol{p}_t)] \\
\leq&(1-\gamma)\sum_{t=1}^T (\boldsymbol{g}_h^t)^\intercal \boldsymbol{p}_t - \mathbb{E}[\sum_{t=1}^T (\boldsymbol{g}_h^t)^\intercal \tilde{\boldsymbol{p}}_t] - (1-\gamma)\mathbb{E}[(\frac{\delta T}{2} + \frac{1}{2\eta})\lambda^2] \\
&+ \mathbb{E}[\lambda\sum_{t=1}^T(1-\gamma)(\rho - (\boldsymbol{g}_h^t)^\intercal \boldsymbol{p}_t)] \\
\leq&(1-\gamma)\frac{m}{\eta}\ln\frac{M}{m} + 2(e-2)\eta MT + (1-\gamma)\eta m^2 T + \frac{\eta mMT}{(1-\gamma)} \\
\leq&\frac{m}{\eta}\ln\frac{M}{m} + 2(e-2)\eta MT + \eta m^2 T + \frac{\eta mMT}{(1-\gamma)}.
\end{aligned}
$$

Let $\lambda = \frac{\sum_{t=1}^T((1-\gamma)\rho - (\hat{\boldsymbol{u}}_h^t)^\intercal \tilde{\boldsymbol{p}}_t)}{\delta T + 1/\eta}$. By taking maximization over $\boldsymbol{p}_t$, we have

$$
\begin{aligned}
&\max_{(\hat{\boldsymbol{u}}_h^t)^\intercal \boldsymbol{p}_t \geq \rho}\sum_{t=1}^T (\boldsymbol{g}_h^t)^\intercal \boldsymbol{p}_t - \mathbb{E}[\sum_{t=1}^T (\boldsymbol{g}_h^t)^\intercal \tilde{p}_t] \\
&+ \mathbb{E}\{\frac{[\sum_{t=1}^T((1-\gamma)\rho - (\hat{\boldsymbol{u}}_h^t)^\intercal \tilde{\boldsymbol{p}}_t)]^{+2}}{2(\delta T + 1\eta)}\} \\
\leq&\frac{m}{\eta}\ln\frac{M}{m} + 2(e-2)\eta MT + \eta m^2 T + \frac{\eta mMT}{(1-\gamma)}. \qquad (12)
\end{aligned}
$$

The above inequation (12) is the upper-bound for $R_{x^h}(T) + \lambda_{x^h}(T)(V_{x^h}(T))^2$ (without considering the impact of delay) when we substitute all the points in sub-hypercube h with the center point x^h. Then we consider the *Context Gap*, which is the difference between the original point $x_{j,t} \in h$ and center point x^h in context sub-hypercube $h \in \mathcal{H}_t$. We utilize $\max\{D_X(x_{j1}^{h_t}, x_{j2}^{h_t})\} = \max\{L||x_{j1} - x_{j2}||^\alpha\} = L(\frac{\sqrt{d_X}}{n_T})^\alpha$ to denote the deviation in the context sub-hypercube $h \in \mathcal{H}_t$, and let $n_T = \lceil T^\theta \rceil \leq 2T^\theta, 0 < \theta < \frac{1}{d_X}$, thus we can get the upper-bound of $R_h(T) + \lambda_h(T)(V_h(T))^2$ for each sub-hypercube $h \in P_T$ (without delay):

$$
\begin{aligned}
&\max_{\hat{\boldsymbol{u}}_t^\intercal \boldsymbol{p}_t \geq \rho}\sum_{t=1}^T \boldsymbol{g}_t^\intercal \boldsymbol{p}_t - \mathbb{E}[\sum_{t=1}^T \boldsymbol{g}_t^\intercal \tilde{p}_t] + \mathbb{E}\{\frac{[\sum_{t=1}^T((1-\gamma)\rho - \hat{\boldsymbol{u}}_t^\intercal \tilde{\boldsymbol{p}}_t)^+]^2}{2(\delta T + 1\eta)}\} \\
&\leq L(\frac{\sqrt{d_X}}{n_T})^\alpha(\frac{m}{\eta}\ln\frac{M}{m} + 2(e-2)\eta MT + \eta m^2 T + \frac{\eta mMT}{(1-\gamma)}) \\
&\leq L d_X^{\frac{\alpha}{2}} T^{-\alpha\theta}[\frac{m(\delta+m)M}{\gamma\delta m}\ln\frac{M}{m} \\
&+ \frac{\gamma\delta m}{(\delta+m)M}(2(e-2)MT + m^2 T + \frac{mMT}{(1-\gamma)})].
\end{aligned}
$$

Summing both sides over $(n_T)^{d_X}$ sub-hypercubes, we can derive the bound of the modified regret function in the form of (4) without delay, $B(T)$:

$$
\begin{aligned}
&\max_{(\hat{\boldsymbol{u}}_h^t)^\intercal \boldsymbol{p}_t \geq \rho}\sum_{t=1}^T\sum_{h\in\mathcal{H}_t} \boldsymbol{g}_t^\intercal \boldsymbol{p}_t - \mathbb{E}[\sum_{t=1}^T\sum_{h\in\mathcal{H}_t} \boldsymbol{g}_t^\intercal \tilde{p}_t] \\
&+ \mathbb{E}\{\frac{[\sum_{t=1}^T\sum_{h\in\mathcal{H}_t}((1-\gamma)\rho - \hat{\boldsymbol{u}}_t^\intercal \tilde{\boldsymbol{p}}_t)]^{+2}}{2(\delta T + 1\eta)}\} \\
\leq&2^{d_X} L d_X^{\frac{\alpha}{2}} T^{(d_X-\alpha)\theta}[\frac{m(\delta+m)M}{\gamma\delta m}\ln\frac{M}{m} \\
&+ \frac{\gamma\delta m}{(\delta+m)M}(2(e-2)MT + m^2 T + \frac{mMT}{(1-\gamma)})]. \qquad (13)
\end{aligned}
$$

The performance of BPG with non-delayed feedback has been discussed above. However, the non-delayed feedback assumption can be easily violated in application since the source mmBS can only observe the feedback of predictions after one time slot. Therefore, we analyze the performance of BPG with delayed feedback. Since we split the time slot sequence $\{1, 2, ..., T\}$ into $\lceil K(T) \rceil$ consecutive segments

$\{[t'_n, t'_{n+2})\}_{n=1}^{\lceil K(T) \rceil}, n \in \mathbb{N}^+$. Such that for any $t \in [t'_n, t'_{n+2})$, we have $\lceil K(t) \rceil = n$, $K(T) = t^{\frac{2}{3+d_X}} log(t)$. That is, we run in parallel $d+1$ (in our model $d = 1$) instances of the BPG for the standard (no delay) setting, and at each time step $t = 2r + s, s \in \{0, 1\}$, for each $r = 1, 2, ..,$ use instance $s + 1$ for the current play. Hence, the non-delayed bound applies to every new *compound slot* and we obtain $B_D(T) \leq \sum_1^2 B(\frac{T}{2})$, where $B_D(T)$ denotes the bound of (4) in delayed setting. Therefore, we obtain $B_D(T)$ as follows:

$$\max_{\hat{\boldsymbol{u}}_t^\intercal \boldsymbol{p}_t \geq \rho} \sum_{t=1}^{T} \sum_{h \in \mathcal{H}_t} \boldsymbol{g}_t^\intercal \boldsymbol{p}_t - \mathbb{E}[\sum_{t=1}^{T} \sum_{h \in \mathcal{H}_t} \boldsymbol{g}_t^\intercal \tilde{p}_t]$$
$$+ \mathbb{E}\{\frac{[\sum_{t=1}^{T} \sum_{h \in \mathcal{H}_t} ((1-\gamma)\rho - \hat{\boldsymbol{u}}_t^\intercal \tilde{\boldsymbol{p}}_t)]^{+2}}{2(\delta T + 1\eta)}\}$$
$$\leq \sum_{n=1}^{2} 2^{d_X} L d_X^{\frac{\alpha}{2}} (\frac{T}{2})^{(d_X - \alpha)\theta} [\frac{m(\delta + m)M}{\gamma \delta m} \ln \frac{M}{m}$$
$$+ \frac{\gamma \delta m}{(\delta + m)M} (2(e-2)M(\frac{T}{2}) + m^2(\frac{T}{2}) + \frac{mM}{(1-\gamma)} \frac{T}{2})]$$
$$= 2^{\alpha\theta + d_X(1-\theta)} L d_X^{\frac{\alpha}{2}} [2 \frac{m(\delta + m)M}{\gamma \delta m} \ln \frac{M}{m} T^{(d_X - \alpha)\theta}$$
$$+ \frac{\gamma \delta m}{(\delta + m)M} (2(e-2)M + m^2 + \frac{mM}{(1-\gamma)}) T^{1 + (d_X - \alpha)\theta}].$$

Let $G(T) = 2^{\alpha\theta + d_X(1-\theta)} L d_X^{\frac{\alpha}{2}} [2 \frac{m(\delta+m)M}{\gamma\delta m} \ln \frac{M}{m} T^{(d_X-\alpha)\theta}$ $+ \frac{\gamma\delta m}{(\delta+m)M} (2(e-2)M + m^2 + \frac{mM}{(1-\gamma)}) T^{1+(d_X-\alpha)\theta}]$. Let $\theta = \frac{1}{9(d_X)}$, we have results in the form of (5): $R(T) \leq G(T)$ and $V(T) \leq \sqrt{2(G(T) + mT)(\delta T + 1/\eta)}$. Since $\gamma = \min(1, \sqrt{\frac{2(e-2)M+Mm}{m \ln(M/m) T^{2/3}}}) = \Theta(T^{-\frac{1}{3}})$ and $\delta = \frac{4(e-2)\gamma m}{1-\gamma} = \Theta(T^{-\frac{1}{3}})$, we have $\eta = \Theta(\frac{1}{M} T^{-\frac{2}{3}})$. Finally, we have sub-linear bounds for the regret and the violation:
$R(T) \leq O(L d_X^{\frac{\alpha}{2}} mM \ln(M) T^{\frac{5}{6} - \frac{\alpha}{6 d_X}})$,
$V(T) \leq O(L^{\frac{1}{2}} d_X^{\frac{\alpha}{4}} m^{\frac{1}{2}} M^{\frac{1}{2}} T^{\frac{5}{6}})$. □

5. Performance Evaluation

Simulation Settings. Our simulation study is conducted based on real-world mobility traces of taxi cabs in Rome [9]. We observe the traffic pattern of a specific street (see Fig. 4), where there are a source mmBS and a target mmBS. We set the distance between them to be 100m [4]. We consider two kinds of blockages in our simulations: static blockages and moving blockages. Static blockages are the buildings and trees along the road; Moving blockages are the randomly appeared objects with large size, *i.e.*, trunks and buses. To better simulate the moving blockage scenario, we randomly select some of these taxi cabs to be trucks and buses. We set the vehicles' speeds between 40 km/h and 80 km/h. In our simulation, a time slot is defined as the maximal time in which a vehicle is within the coverage of source mmBS, and we set $t = 10s$ here.

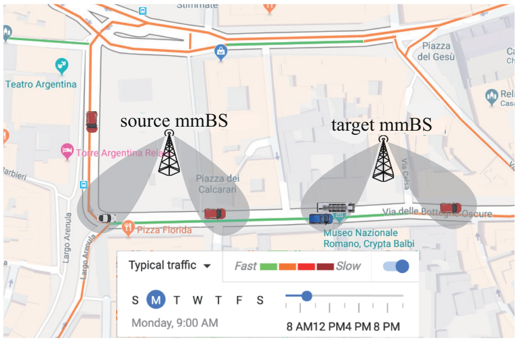

Figure 4: The map of the simulation environment, where the buildings are static blockages and trucks are moving blockages.

We normalize the throughput and blockage-free probability to [0,1]. Even for the same beam, the throughput and blockage-free probability varies when facing different types of vehicles. Follow the similar settings in [14][4], we assume the target mmBS has 16 orthogonal beams, each beam's width varies from 10° to 40° for omni-directional coverage. We set vehicles' beam width to 30°. The number of simultaneously selected beams m is 4 by default. We set transmission power to be 30 dBm, and system bandwidth to be 1GHz. The noise figure of mmBS is 4 dB, and 7dB for vehicles. The thermal noise is -174dBm/Hz. The path loss model, which is scaled in dB, is $32.4 + 17.3 log_{10} d(m) + 20 log_{10}(f_c(GHz)) + \xi$, $\xi \sim \mathcal{N}(0, \delta), \delta = 1.1dB$ [29].

Algorithms for comparison. We compare our online algorithm BPG with four benchmark algorithms:

- Optimum: We assume that this algorithm has a prior knowledge of the whole system. Then it can make the best beam predictions since it knows the blockage situation and expected beam performance in advance.

- FML: We modify the *Fast Machine Learning* algorithm [14] to fit our prediction setting. The modified FML aims to maximize the total compound throughput, *i.e.*, $\sum_{t \in [T]} \sum_{j \in N_t} \sum_{i \in [M]} g_{i,x_{j,t}}^t$, and it receives the feedback with one slot delay.

- Adapted-UCB (AUCB): This is the variant of the classic bandit algorithm UCB [30]. The algorithm maintains a set of UCB bonus, which is

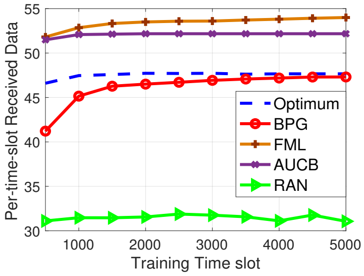

Figure 5: Prediction accuracy of five algorithms after several times of training.

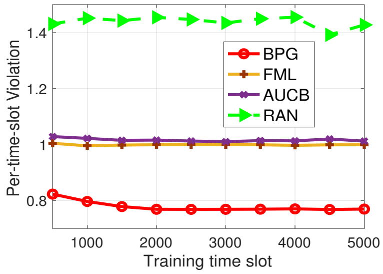

Figure 6: Violation of four algorithms after several times of training.

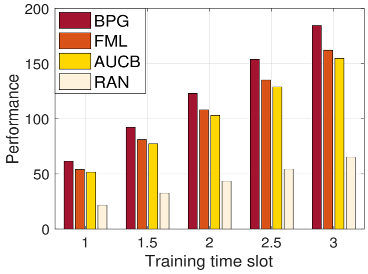

Figure 7: Performance of four algorithms under different QoS weights.

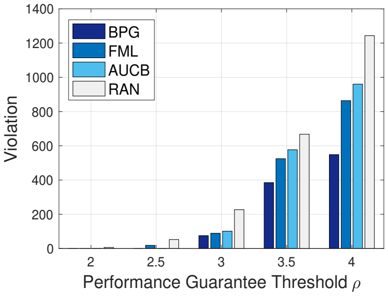

Figure 8: Violation under different value of ρ.

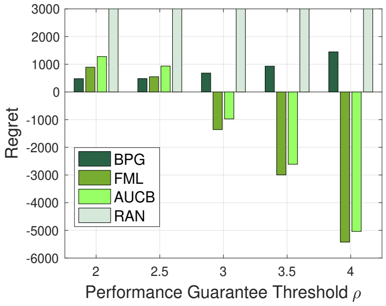

Figure 9: Regret under different value of ρ.

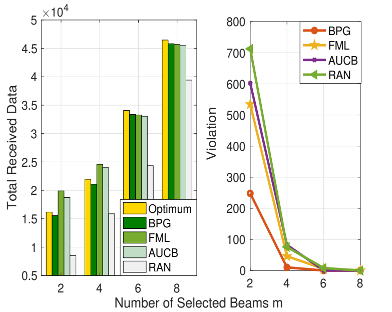

Figure 10: Impact of the number of selected beams (m).

used to characterize each beam's performance. In each round, source mmBS predicts the top-m beams with the highest UCB indices $\hat{g}_i^t + \sqrt{1.5\log(t)/(N_i(t))}$. $N_i(t)$ is the number of times that beam i has been selected before $t-1$.

- Ran: This algorithm randomly selects m beams for target mmBS in each round.

Performance Analysis. We collect the vehicular system information within 5000 time slots. We compute the per-time-slot optimal beam prediction solution accordingly. We utilize the collected system information as a set of training data and feed them to four algorithms for training. We analyze the per-time-slot throughput and the per-time-slot violation after iterations of training (see Fig. 5 and Fig. 6). We utilize the optimal solutions as standards to verify the other four algorithms' performance. As seen in Fig. 5, the accuracy of BPG improves with the increase of training time. During the first 1500 time slots, the throughput obtained by our algorithm is much smaller than the optimum, since BPG may make bad choices in the exploration phrase. The throughput of BPG approaches the optimum after $t = 3500$, and we utilize the obtained model for the following simulations. Note that the throughput of FML and AUCB are always larger than the value of optimum, since these two algorithms merely tend to maximize the total throughput without considering the system QoS constraints, leading to great violation in the long term.

Fig. 6 shows that our algorithm's violation is the smallest, which means BPG's prediction shows the best QoS performance and guarantees the stability of vehicular communication. To better understand the performance of each algorithm under vehicular communication system, we first define the QoS of the system as $QoS = \frac{1}{Violation}$. We then use *Total Received Data* $\cdot$ *(QoS weight* $\cdot$ *QoS)* to evaluate the algorithm performance. As shown in Fig. 7, our algorithm has the best performance under different QoS weights. Thus, It keeps a good balance between a large throughput and a low violation. With the increase of QoS weight, the gap between BPG and other algorithms becomes larger.

We then investigate the impact of the performance guarantee threshold ρ on the regret and violation of each algorithm. In Fig. 8 and Fig. 9, we can see that when we increase the value of ρ, all the algorithms' violations and regrets increase. However, our algo-

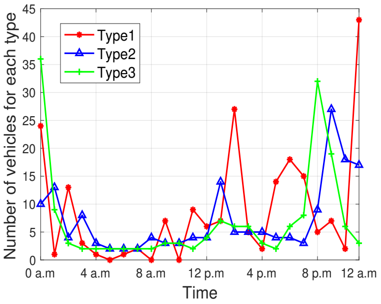

Figure 11: Traffic pattern for 24 hours at the chosen street in Rome [9].

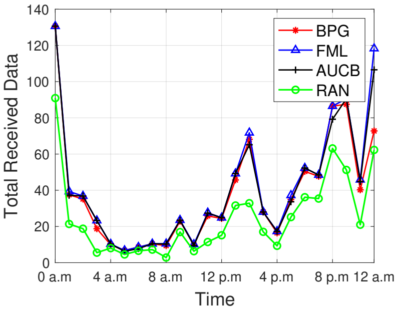

Figure 12: Throughput under dynamic traffic pattern for 24 hours.

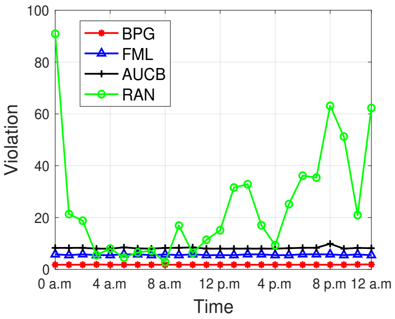

Figure 13: Violation of QoS under dynamic traffic pattern for 24 hours.

rithm BPG keeps the lowest regret level. We can see from Fig. 10, when the number of selected beams increases, each algorithm's throughput increases too. This is due to the increased coverage areas of the system. Also, if the value of m increases, the violation of each algorithm decreases to zero since it gets easier to satisfy the minimum performance guarantee. The difference between BPG and FML, AUCB gets smaller due to the easily satisfied performance requirement.

We plot the dynamic traffic pattern of one specific street for 24 hours in Rome based on the real-world data source [9] in Fig. 11, containing three types of vehicles. Under this traffic pattern, we compare BPG with three other benchmark algorithms. As shown in Fig. 12 and Fig. 13, the throughput of BPG, FML and AUCB are nearly the same, however, our algorithm BPG always has the lowest violation. Therefore, BPG shows the best performance since we strike a good balance between maximizing the throughput while keeping a lowest violation.

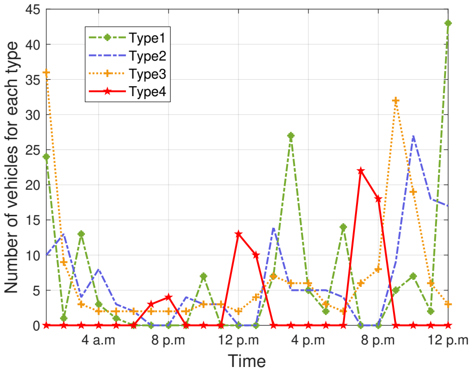

Figure 14: Trafffic pattern with congestion.

In order to deal with specific conditions such as traffic congestion and speeds starting from 0 km/h, we make traffic congestion assumptions based on the realistic congestion situations in Rome [31][32]. We divide the vehicles into four categories according to their speeds. That is, we classify vehicles with speeds of 60-80km/h as type-1, 40-60km/h as type-2, 20-40km/h as type-3 and 0-20km/h as type-4. The number of type-4 vehicles reflects the traffic congestion situations. Especially at rush hours during a day, all vehicles' speeds are slow and the number of type-4 cars increases. The new traffic pattern are shown in Fig. 14. As can be seen in Fig. 15 and Fig. 16, BPG almost receives the largest communication data and the lowest violation even with traffic congestion. Therefore, our algorithm BPG still shows the best performance under this traffic pattern.

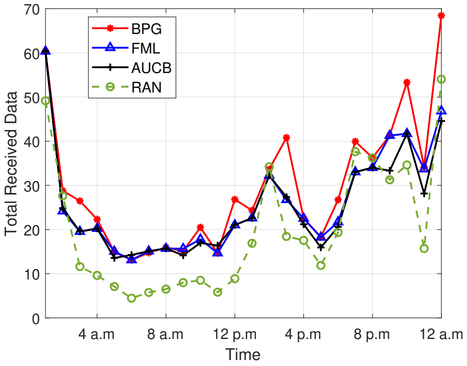

Figure 15: Throughput under traffic pattern with congestion.

As for the computation overhead, we compare the running time of our algorithm BPG with others. We apply the tic and toc functions in MATLAB to measure the running time. We run 20 tests on MacBook Air (I1.4GHz inter Core i5/4GB /1600 MHz DDR3) and present the average result in Table 2, where TS stands for time consumption. Although BPG consumes longer time to compute each round, the difference is not significant. Considering the near-to-

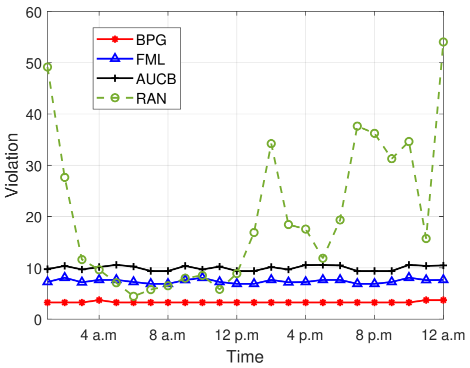

Figure 16: Violation under traffic pattern with congestion.

optimal solution obtained by BPG, our algorithm still shows a good performance compared with others.

Table 2: Algorithm Overhead.

Alg	Optimal	BPG	FML	UCB	RAN
TS (ms)	1.401	0.018	0.012	0.006	0.002

6. Concluding Remarks

This work proposes an online learning algorithm that makes close-to-optimal beam selection for a target mmBS. Our algorithm, BPG, makes predictions under the consideration of vehicle contexts and system QoS constraints. We aim to strike a balance between maximizing the total throughput and satisfying the minimum performance guarantee. Through valid theoretical analysis, we prove that our online learning method achieves sub-linear bounds for regret and violation. We also verify our algorithm's good performance as compared to other benchmark algorithms through simulation studies.

We have mainly focused on the situation that the target mmBS selects beams completely according to the predictions from the source mmBS. Our model can be extended to more complex settings, where the prediction policy is utilized to guide the beam selection strategy of the target mmBS. Our prediction policy will reduce the training overhead and increase the selection accuracy. It can also reduce the beam rearrangement latency, which contributes to better performance of vehicular communication. Note that our beam prediction policy can be adapted to handle additional constraints in more realistic settings: (i) non-orthogonality can be formulated as an additional constraint, *i.e.*, overlapping beams cannot be used simultaneously; (ii) reflection from surrounding objects can be seen as an interference, thus we can add a constraint to keep the interference under a threshold. We will continue studying the multi-level constraints vehicular communication model in our future work.

References

[1] M. R. Hafner, D. Cunningham, L. Caminiti, D. Del Vecchio, Cooperative collision avoidance at intersections: Algorithms and experiments., IEEE trans. intelligent transportation systems 14 (3) (2013) 1162–1175.

[2] Leading the world to 5G: C-V2X technologies, `https://dwz.cn/tAuj7Fp7`.

[3] Y. Niu, Y. Li, D. Jin, L. Su, A. V. Vasilakos, A survey of millimeter wave communications (mmwave) for 5g: opportunities and challenges, Wireless Networks 21 (8) (2015) 2657–2676.

[4] D. Yuan, H.-Y. Lin, J. Widmer, M. Hollick, Optimal joint routing and scheduling in millimeter-wave cellular networks, in: Proc. of IEEE INFOCOM, 2018.

[5] Y. Wang, M. Narasimha, R. W. Heath, Mmwave beam prediction with situational awareness: A machine learning approach, arXiv preprint arXiv:1805.08912 (2018) 1–5.

[6] S. Chen, Z. Jiang, S. Zhou, Z. Niu, Time-sequence channel inference for beam alignment in vehicular networks, arXiv preprint arXiv:1812.01220 (2018).

[7] V. Va, T. Shimizu, G. Bansal, R. W. Heath Jr, Online learning for position-aided millimeter wave beam training, arXiv preprint arXiv:1809.03014 (2018).

[8] G. E. Garcia, G. Seco-Granados, E. Karipidis, H. Wymeersch, Transmitter beam selection in millimeter-wave mimo with in-band position-aiding, IEEE Transactions on Wireless Communications (TWC) PP (99) (2017) 1–1.

[9] L. Bracciale, M. Bonola, P. Loreti, G. Bianchi, R. Amici, A. Rabuffi, CRAWDAD dataset roma/taxi (v. 2014-07-17), Downloaded from `https://crawdad.org/roma/taxi/20140717` (Jul. 2014). doi:10.15783/C7QC7M.

[10] P. Kela, M. Costa, J. Turkka, M. Koivisto, J. Werner, A. Hakkarainen, M. Valkama, R. Jantti, K. Leppanen, Location based beamforming in 5g ultra-dense networks, in: 2016 IEEE VTC, 2016.

[11] P. Kela, M. Costa, K. Leppänen, R. Jäntti, Location-aware beamformed downlink control channel for ultra-dense networks, in: 2017 IEEE CSCN, 2017.

[12] I. Mavromatis, A. Tassi, R. J. Piechocki, A. Nix, mmwave system for future its: A mac-layer approach for v2x beam steering, in: 2017 IEEE VTC, 2017.

[13] A. Ali, N. Gonzlez-Prelcic, J. R. W. Heath, Millimeter wave beam-selection using out-of-band spatial information, IEEE Transactions on Wireless Communications (TWC) 17 (2) (2018) 1038–1052.

[14] A. Asadi, S. Müller, G. H. Sim, A. Klein, M. Hollick, Fml: Fast machine learning for 5g mmwave vehicular communications, in: Proc. of IEEE INFOCOM, 2018.

[15] A. Alkhateeb, S. Alex, P. Varkey, Y. Li, Q. Qu, D. Tujkovic, Deep learning coordinated beamforming for highly-mobile millimeter wave systems, arXiv preprint arXiv:1804.10334 (2018).

[16] I. Mavromatis, A. Tassi, R. J. Piechocki, A. Nix, Mmwave system for future its: A mac-layer approach for v2x beam steering, in: Proc. of VTC, 2018.

[17] W. Chen, Y. Wang, Y. Yuan, Combinatorial multi-armed bandit: General framework and applications, in: Proc. of ICML, 2013.

[18] J. Komiyama, J. Honda, H. Nakagawa, Optimal regret analysis of thompson sampling in stochastic multi-armed bandit problem with multiple plays, arXiv preprint arXiv:1506.00779 (2015).

[19] A. Slivkins, Contextual bandits with similarity information, Journal of Machine Learning Research 15 (1) (2014) 2533–2568.

[20] L. Li, W. Chu, J. Langford, R. E. Schapire, A contextual-bandit approach to personalized news article recommendation, in: Proc. of ACM WWW, 2010.

[21] S. Mller, O. Atan, M. V. D. Schaar, A. Klein, Context-aware proactive content caching with service differentiation in wireless networks, IEEE Transactions on Wireless Communications (TWC) 16 (2) (2017) 1024–1036.

[22] G. Neu, A. Antos, A. György, C. Szepesvári, Online markov decision processes under bandit feedback, in: Proc. of NIPS, 2010.

[23] P. Joulani, A. Gyorgy, C. Szepesvári, Online learning under delayed feedback, in: Proc. of ICML, 2013.

[24] L. Chen, J. Xu, Task offloading and replication for vehicular cloud computing: A multi-armed bandit approach, arXiv preprint arXiv:1812.04575 (2018).

[25] M. Mahdavi, T. Yang, R. Jin, Online decision making under stochastic constraints, in: Proc. of NIPS workshop on Discrete Optimization in Machine Learning, 2012.

[26] K. Cai, X. Liu, Y. Z. J. Chen, J. C. Lui, An online learning approach to network application optimization with guarantee, in: Proc. of IEEE INFOCOM, 2018.

[27] R. Gandhi, S. Khuller, S. Parthasarathy, A. Srinivasan, Dependent rounding and its applications to approximation algorithms, Journal of the ACM (JACM) 53 (3) (2006) 324–360.

[28] P. Auer, N. Cesa-Bianchi, Y. Freund, R. E. Schapire, The nonstochastic multiarmed bandit problem, SIAM journal on computing 32 (1) (2002) 48–77.

[29] 3GPP, Technical specification group radio access network: Study on channel model for frequencies from 0.5 to 100 ghz, in: Tech. Rep., Mar, 2017.

[30] P. Auer, N. Cesa-Bianchi, P. Fischer, Finite-time analysis of the multiarmed bandit problem, Machine learning 47 (2-3) (2002) 235–256.

[31] Rush hour in Rome, `https://dwz.cn/ZGycfEfM`.

[32] Real-time traffic information in Rome, `https://dwz.cn/wVxJCIbi`.