M. Wirth, A. Klein and A. Ortiz, ”Risk-Aware Multi-Armed Bandits for Vehicular Communications”, in *2022 IEEE 95th Vehicular Technology Conference (VTC2022-Spring)*, June 2022.

Risk-Aware Multi-Armed Bandits for Vehicular Communications

Maximilian Wirth
Communications Engineering Lab
Technische Universität Darmstadt
m.wirth@nt.tu-darmstadt.de

Anja Klein
Communications Engineering Lab
Technische Universität Darmstadt
a.klein@nt.tu-darmstadt.de

Andrea Ortiz
Communications Engineering Lab
Technische Universität Darmstadt
a.ortiz@nt.tu-darmstadt.de

Abstract*—The importance of vehicular communications has grown significantly in recent years. Potential use cases of vehicular communications are manifold and range from sharing information for driver assistance to entertainment purposes. This means that each connected vehicle has an individual data requirement for the communication infrastructure. However, due to the dynamic wireless environment, the simultaneous fulfillment of such requirements cannot be guaranteed. Therefore, novel solutions should not only consider the requirements of each user but also the risk of not being able to fulfill them. In this paper, we consider a vehicular communication scenario consisting of a base station that serves the vehicles in its coverage area using 5G millimeter wave (mmWave) narrow beams. The problem boils down to finding an optimal policy for the selection of the narrow beams. This should be done carefully, as the choice of the used beams greatly impacts the performance. For this purpose, we propose a risk-aware contextual Multi-Armed Bandit (MAB) online learning algorithm. Using this algorithm, the base station autonomously learns its environment and selects the best set of beams based on the vehicles located in its coverage area. In order to achieve a large risk awareness, this work focuses on two pillars. Firstly, the notion of risk is integrated in the proposed contextual MAB algorithm by exploiting the concepts of Mean-Variance and Conditional Value at Risk for the evaluation of the decisions made by the algorithm. Secondly, we introduce mechanisms that can detect non-stationarities and swiftly adapt to them in order to make the proposed approach robust against volatile environments that violate stationarity assumptions. By using extensive simulations, the effectiveness of the aforementioned approaches are proven numerically.

I. Introduction

In recent years, there has been a rapid increase of interest in the field of vehicular communication. New cars are equipped with a multiplicity of hardware that performs various tasks ranging from convenience features, such as infotainment systems, to safety relevant functionalities, like driver assistance. Furthermore, modern vehicles employ a large number of sensors to measure different parameters, such as road conditions or traffic patterns [1]. Sharing the information gathered from the acquired sensor data can have many advantages, as it can reduce the likelihood of accidents or increase the traffic flow [2]. Nevertheless, the large scale acquisition and exchange of sensor information puts high demands on the communication infrastructure due to the high throughput it requires. Furthermore, the exchange of information needs to be rapid and almost in real time, which translates to a low latency requirement. One potential candidate to serve those high demands is 5G millimeter wave (mmWave) technology [3], because it enables very high data rates and low latency [4]. However, mmWave comes with its own unique drawbacks, as it suffers from a higher path loss compared to older technologies like 4G as well as larger losses due to shadowing [5]. Nevertheless, the fact that in 5G mmWave, the radiated energy is strongly concentrated in one direction can be used to purposely steer the radiated energy into the desired direction [6]. Moreover, the mmWave base stations can transmit using a subset of these narrow beams simultaneously. Thus, mmWave beam technology enables spatial multiplexing by design and can therefore increase the throughput even further.

In this paper, the aforementioned 5G mmWave narrow beams are used for the communication between vehicles and a base station. The cars are assumed to have individual data requirements, and a policy for the selection of a subset of mmWave beams to be used, is proposed. The beam selection problem is challenging because, due to the small width of the beams, a misalignment between transmitter and receiver can cause a severe degradation in performance [7]. Furthermore, due to the constantly changing environment, e.g., the occurrence of obstructions or the continuously varying traffic flow, the simultaneous fulfillment of the data requirements cannot be guaranteed. Moreover, the environment may not even be stationary in the probabilistic sense. To address these challenges, in this work we aim at minimizing the likelihood that the individual requirements of the vehicles are not fulfilled and propose a risk-aware vehicular communication approach capable of quickly learning and adapting to its environment, even if stationarity assumptions are violated.

The subject of mmWave beam selection has been treated in various works in the past. Most of the earlier approaches are based on performing a search over the space of available beams in order to find the optimal choices [8], [9]. Such strategies can be computationally prohibitive and therefore cause substantial latencies. In the context of vehicular communications, large delay times can be critical. More recent works tackle the problem of the beam selection by means of reinforcement learning techniques. In particular, many approaches are based on Multi-Armed Bandit (MAB) theory. The authors of [10] use a MAB framework and exploit the correlation between beams. In addition to that, a priori information about channel fluctuations is incorporated. In [11],

a MAB algorithm is used in conjunction with channel tracking based on Bayesian learning and Kalman filtering. The authors of [12] propose a contextual MAB algorithm for the beam selection in a vehicular communication scenario. One common shortcoming of all the aforementioned approaches is that they fail to incorporate the notion of risk in the beam selection process. These works mainly focus on maximizing the overall performance without evaluating if the requirements of an individual user are satisfied. Thus, in the context of vehicular communications, their applicability is reduced.

Similarly to [12], in our work, the considered scenario consists of a mmWave base station that serves multiple cars in its coverage area. It is assumed that a fixed set of directional beams is available and the goal is to select the best subset of beams, depending on the environment and the vehicles located in the coverage area. The main goal of this work is to achieve a safe and reliable communication, which is achieved by the following contributions of our paper:

- We include the risk that some vehicles are not able to fulfill their communication requirements and consider environments for which stationarity assumptions are violated.
- We propose a contextual MAB algorithm that aims for risk-aware decisions.
- We introduce mechanisms that enable the MAB algorithm to quickly adapt to non-stationary environments.
- We show that our proposed risk-aware algorithm outperforms risk-neutral formulations.

As it is generally very hard or even impossible to give strict guarantees for the communication reliability, this work tackles this challenge by integrating the risk awareness already in the evaluation of the decisions made by the algorithm, rather than by trying to satisfy some Quality-of-Service side constraint. For this purpose, the concepts of Mean-Variance and Conditional Value at Risk are exploited. In the field of MABs, risk aversion has been treated by using Mean-Variance [13], [14] and Conditional Value at Risk [15], [16]. However, to the best of our knowledge, there is no work that treated the subject of risk aversion in combination with contextual MABs.

In the following, Section II introduces the system model. In Section III, the optimization problem is formulated and concepts for a risk aware assessment of the performance of the beams are explained. After that, in Section IV, the risk aware contextual MAB algorithm is explained, including the mechanisms that enable an adaptation for non-stationary environments. A performance evaluation based on numerical results is shown in Section V.

II. System Model

In this section, the system model is introduced. For this work, only the downlink case, i.e., the transmission from the base station to the vehicles, is considered. However, all principles described can also be applied to the uplink case in a similar fashion. The considered scenario consists of a mmWave base station and moving vehicles that are located inside the coverage area of the base station, as shown in Figure 1. For the communication with the vehicles, the mmWave base station has a set $\mathcal{B}$ comprised of $B = |\mathcal{B}|$ orthogonal directional beams available. For each time step $t = 1, \dots, T$ under the considered time horizon T, the mmWave base station selects $m \leq B$ beams out of $\mathcal{B}$, in dependence on the V_t vehicles currently located in its coverage area. The number of vehicles $V_t \in \mathbb{N}$ per time step t is upper bounded by V_{max} for every time step, i.e., $V_t \leq V_{max}$. Additionally, $v_{t,i},\ i = 1, \dots, V_t$ describes the V_t vehicles in the coverage area during time step t and $\mathcal{V}_t = \{v_{t,i}\}_{i=1,\dots,V_t}$ summarizes them in one set. It is assumed that a beam is not directly allocated to a certain vehicle. Instead, all m selected beams can transmit to all V_t vehicles at the same time. In order to achieve this, the mmWave base station may transmit a combined message, which contains a superposition of all the individual messages belonging to the different vehicles. The vehicles can then extract their part of the combined message by employing a decoding technique, e.g., successive interference cancellation. Furthermore, every vehicle demands a certain minimum amount of received data per time step, measured in bits, represented by a Quality-of-Service constraint. Specifically, for every time step $t = 1, \dots, T$, each of the vehicles $v_{t,i},\ i = 1, \dots, V_t$ that are in the coverage area during that time step are assumed to require a certain minimum amount of data $r_{t,i}^{\min}$. This amount of data corresponds to the information that the vehicle requests from the base station in time slot t. This data can range from some driving related information, such as the traffic, the road conditions or the weather, to entertainment content, like a music or movie stream.

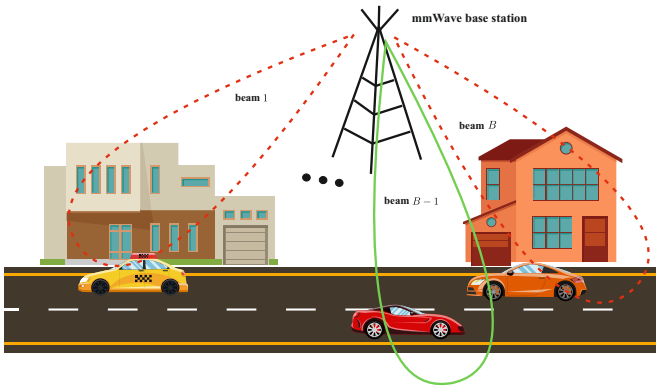

Fig. 1. Exemplary vehicular communication scenario

The beam selection aims for two goals: maximizing the average received amount of data for all V_t vehicles and reducing the risk that some vehicles do not receive enough data. In section III, mathematical concepts for metrics that capture the aforementioned aspects are introduced. The m selected beams in time step t are denoted by $s_{t,j},\ j = 1, \dots, m$ and $\mathcal{S}_t = \{s_{t,j}\}_{j=1,\dots,m} \subseteq \mathcal{B}$ describes the resulting subset of selected beams. The number of beams m that the mmWave base station selects per time step is limited by external factors such as the channel characteristics, or restrictions imposed by the used hardware. Furthermore, the selection of the beams

strongly depends on the environment in the coverage area, e.g., permanent blockages due to buildings or temporary obstructions resulting from moving objects, e.g., other vehicles. It is assumed that the mmWave base station has no prior knowledge about its environment. Thus, it has to learn and adapt to its environment completely autonomously.

It is assumed that each vehicle $v_{t,i}$ is characterized by a context vector $x_{t,i}$. This context vector $x_{t,i}$ is taken from the X-dimensional context space $\mathcal{X} = [0,1]^X$, where $X \in \mathbb{N}$. Each context dimension reflects certain attributes that describe the vehicles, which can be discrete or continuous quantities. Examples for these attributes are the vehicle's velocity, its location, the direction towards it is moving or the type of vehicle. Note that regardless of the type of quantity the context dimension describes, it is always mapped to the interval $[0,1]$.

The quality of a beam is assessed by the amount of data received by the vehicles. Let $r_{t,j}(x_{t,i},t)$ denote the actual amount of data that vehicle $v_{t,i}$ with context $x_{t,i}$ has received from the beam $s_{t,j} \in \mathcal{B}$ in time slot t. Note that $r_{t,j}(x_{t,i},t)$ is not a deterministic quantity. The instantaneous received amount of data $r_{t,j}(x_{t,i},t)$ depends on random occurrences such as fluctuations of the channel or the sudden appearance of temporary obstructions. Thus, the received amount of data $r_{t,j}(x_{t,i},t)$ is a random variable. Specifically, $r_b(x)$ describes how much data a vehicle with context x receives from beam $b \in \mathcal{B}$ in a single time slot. It is assumed that the amount of received data is upper bounded by $R_{\max}$, i.e., $r_b(x) \leq R_{\max}$. $R_{\max}$ depends on the environment and the length of one time slot, as this determines the maximum contact time, i.e., the time that a vehicle is served by a certain beam. Additionally, $R_{\max}$ is also influenced by additional factors, such as the employed modulation scheme. Consequently, the random variable $r_b(x)$ is confined to the interval $[0, R_{\max}]$, since the minimum amount of received data cannot be smaller than zero. Finally, $r_b(x)$ is characterized by statistical quantities. $\mu_b(x) = \mathbb{E}[r_b(x)]$ is the expected received amount of data from beam b for a vehicle with context x, where $\mathbb{E}[\cdot]$ is the expectation operator. Furthermore, $\sigma_b^2(x) = \mathbb{E}[(r_b(x) - \mu_b(x))^2]$ is the variance of the received amount of data.

III. Problem Formulation

This section treats the formal definition of the beam selection problem. At first, an optimization problem is formulated. In the next step, two risk aware metrics for evaluating the beam performance are introduced.

A. Optimization Problem

Our goal is to find a beam selection policy that maximizes the received amount of data for all the vehicles connected to the mmWave base station during each time step t. At the same time, the risk of selecting beams that lead to poor performance should be minimized. We consider a general function $f_b(x)$ to assess the performance of beam b for a vehicle with context x. The advantage of this general approach is that it allows us to incorporate any relevant statistical property of the random variables $r_b(x)$ into the beam performance assessment. Specifically, $f_b(x)$ is used to evaluate the risk associated with selecting a certain beam.

In order to formulate the problem, firstly, consider the optimization variable $y_{t,b}$, which is a binary decision variable that represents the beam selection. If beam b is selected in time slot t, then $y_{t,b} = 1$, otherwise it is defined to be zero. The objective function of the optimization problem (1) is defined as the expected cumulative beam performance. The summation is done over all vehicles and all time steps in the considered time horizon T. For each time step t, the optimum is achieved by selecting the m beams $\{s_{t,j}\}_{j=1,\dots,m}$ that maximize the sum of the expected beam performances. Note that the beam selection can be made independently for each time step, since for every t, a new subset of beams is selected. The resulting optimization problem is given by

$$\max_{\{y_{t,b}\}_{b\in\mathcal{B},\ t=1,\dots,T}} \sum_{t=1}^{T}\sum_{b\in\mathcal{B}} y_{t,b} \sum_{i=1}^{V_t} f_b(x_{t,i}) \tag{1}$$

$$\text{subject to} \sum_{j=1}^{|\mathcal{S}_t|} r_{t,j}(x_{t,i},t) \geq r_{t,i}^{\min}, \qquad \forall\, t,\ \forall\, i \tag{2}$$

$$\sum_{b\in\mathcal{B}} y_{t,b} \leq m, \qquad \forall\, t \tag{3}$$

$$y_{t,b} \in \{0,1\}, \qquad \forall\, b,\ \forall t. \tag{4}$$

where (2) corresponds to the Quality-of-Service constraint, i.e., the minimum required amount of data. Constraint (3) ensures that no more than m beams are selected for each time step t. At the optimum, this inequality is fulfilled with equality, since more selected beams will always increase the received amount of data further and we neglected any possible interference between the beams. Finally, the binarity of the decision variable $y_{t,b}$ is enforced by constraint (4).

In order to satisfy constraint (2), for every time step t, perfect knowledge about the received amount of data $r_{t,j}(x_{t,i},t)$ is required prior to solving the optimization problem. However, this information can never be accessed in advance. Thus, the optimization problem cannot be solved optimally, as it is impossible to give a strict guarantee for the satisfaction of constraint (2). Therefore, we propose an online learning approach that assesses and selects the beams in a risk-aware fashion, such that the risk of not fulfilling constraint (2) is reduced. In this context, it is important to choose the beam assessment function $f_b(x)$ in such a way, that it incorporates the risk associated with selecting a certain beam.

B. Risk Aware Beam Assessment

In the following, two choices for the general beam performance assessment function $f_b(x)$ are discussed, the Mean-Variance and the Conditional Value at Risk. As already mentioned, $f_b(x)$ should encourage a risk averse behavior, i.e., beams that have a high likelihood of providing small amounts of received data should be avoided.

1) Mean-Variance: The idea behind the Mean Variance is very intuitive. From a risk averse perspective, not only the average performance, i.e, the statistical mean, is important, but

also the likelihood of deviations from it. This is expressed by utilizing the variance of the random variable in the sense that the larger the variance, the higher the uncertainty of a certain beam. Thus, the higher the risk for a very small amount of received data. Hence, the Mean-Variance combines both, the expectation and the variance. Formally, the Mean-Variance of the beam performance for a beam b and a vehicle with context x can be defined similarly to [17] as

$$f_b(x) = \mathrm{MV}_b(x) = \mu_b(x) - \rho\sigma_b(x), \qquad \rho \geq 0. \tag{5}$$

Ideally, a beam with a low risk has a large expected received amount of data $\mu_b(x)$ and only a small variance $\sigma_b(x)$, i.e., a low uncertainty. Hence, from a risk aware perspective, the larger the Mean-Variance of a beam, the better. The parameter ρ models the trade-off between the focus on high average performance and low uncertainty. Thus, it determines the risk tolerance. The larger ρ, the more risk averse the behavior and vice versa. The special case of $\rho = 0$ corresponds to the risk neutral case. Furthermore, it should be noted that very large values of ρ might not be reasonable, since in this case, there would not be any attention paid to the average performance.

2) Conditional Value at Risk: The intuition behind the Conditional Value at Risk is to take into account only those values of the random variable that are below a certain quantile of the distribution. In particular, the expectation below that quantile is considered. Thus, the Conditional Value at Risk can be seen as a measure for the likelihood of obtaining small samples from a random variable. Formally, the Conditional Value at Risk $\mathrm{CVaR}_b(x)$ of beam b for a vehicle with context x is defined as follows,

$$f_b(x) = \mathrm{CVaR}_b(x) = \mathbb{E}[r_b(x) | r_b(x) < \mathrm{VaR}_b(x)] \tag{6}$$

where,

$$\mathbb{P}[r_b(x) \leq \mathrm{VaR}_b(x)] = \alpha \text{ and } \alpha \in (0, 1]. \tag{7}$$

In this definition, the Value at Risk $\mathrm{VaR}_b(x)$ is equal to the α-quantile of the distribution. The larger the Conditional Value at Risk, the smaller the risk of receiving a small amount of data. That is because in that case, the expected amount of received data in the $\alpha \times \%$ worst cases is large. The parameter α can be understood as a risk tolerance parameter. The smaller α, the higher the risk awareness and vice versa. Hence, for $\alpha = 1$, the risk neutral case is reached.

IV. Proposed Algorithm

In this section, the proposed online learning algorithm for the beam selection is explained. To this aim, the *Fast Machine Learning* (FML) algorithm from [12] is used as a basis. We then include the notion of risk, using the concepts shown in Section III, and augment the algorithm by giving it the capability to quickly adapt to non-stationary environments.

A. Overview

The proposed algorithm is based on a MAB framework. In this context, the multiple arms are the set of available beams $\mathcal{B}$. Furthermore, the reward feedback is given by the instantaneous amount of received data $r_{t,j}(x_{t,i}, t)$ that the vehicles inform the mmWave base station about after each time step. The beams are selected either by means of exploration or exploitation actions. During an exploration step, the MAB algorithm randomly selects beams that have not been used sufficiently many times in previous time steps. For the exploitation, the beams that have performed best in previous time steps are selected.

As explained in Section III, it is impossible to guarantee that the Quality-of-Service constraint (2) is always fulfilled when perfect knowledge is not available. Therefore, our proposed algorithm uses risk-aware metrics for evaluating the performance of the beams and makings its decisions. The intuition behind this is that these metrics already encourage a beam selection policy that reduces the risk of not satisfying constraint (2).

The vehicle context is incorporated by partitioning the context space into a discrete set of hypercubes and estimating the beam performance for each hypercube separately. This can be understood as a sampling of the context space under the assumption that similar contexts will have similar beam performances [18]. The algorithm starts with the partitioning of the context space $\mathcal{X} = [0, 1]^X$. For that sake, each dimension of $\mathcal{X}$ is subdivided into p_T smaller fractions of equal length $\frac{1}{p_T}$, where p_T is an input parameter of the algorithm. In total, $(p_T)^X$ separate hypercubes are obtained. The resulting partition containing these hypercubes is denoted as $\mathcal{P}_T$.

In contrast to [12], our algorithm keeps track of two different counters. The external counter $N^{\mathrm{ex}}_{b,h}(t)$ counts how many times a certain beam b was used in combination with the context hypercube $h \in \mathcal{P}_T$. It is needed for the control of the exploration phases. The internal counter $N^{\mathrm{in}}_{b,h}(t)$ keeps track of the number of samples used to calculate the beam performance estimates $\hat{f}_{b,h}(t)$, i.e., the last $N^{\mathrm{in}}_{b,h}(t)$ reward samples.

For every time step t, the exploration is done by first finding the set of underexplored beams, which contains beams that have not been used often enough in previous time steps. The set of underexplored beams is denoted by $\mathcal{B}^{\mathrm{ue}}_{\mathcal{H}_t}(t)$, where $\mathcal{H}_t = \{h_{t,i}\}_{i=1,\dots,V_t}$ denotes a set that contains all the hypercubes, inside which the contexts of the vehicles in time step t are located. $\mathcal{B}^{\mathrm{ue}}_{\mathcal{H}_t}(t)$ is given by

$$\mathcal{B}^{\mathrm{ue}}_{\mathcal{H}_t}(t) := \bigcup_{i=1}^{V_t} \{b \in \mathcal{B} : N^{\mathrm{ex}}_{b,h_{t,i}}(t) \leq K(t)\}, \tag{8}$$

where $K(t)$ is a control function, which controls the trade-off between exploration and exploitation. It is given to the algorithm as an input. Note that we use the external counter $N^{\mathrm{ex}}_{b,h_{t,i}}$ for the determination of underexplored beams. If $\mathcal{B}^{\mathrm{ue}}_{\mathcal{H}_t}(t)$ is not empty, up to m beams are drawn randomly from this set in time step t, depending on how many beams it contains. If $\mathcal{B}^{\mathrm{ue}}_{\mathcal{H}_t}(t)$ contains less than m beams, then the remaining $m - |\mathcal{B}^{\mathrm{ue}}_{\mathcal{H}_t}(t)|$ beams are chosen by means of an exploitation step, i.e., they are selected according to

$$\hat{b}_{j,\mathcal{H}_t}(t) \in \underset{b \in \mathcal{B} \setminus (\mathcal{B}^{\mathrm{ue}}_{\mathcal{H}_t}(t) \bigcup_{k=1}^{j-1} \{\hat{b}_{k,\mathcal{H}_t}(t)\})}{\arg\max} \sum_{i=1}^{V_t} \hat{f}_{b,h_{t,i}}(t). \tag{9}$$

For the estimates for the beam performance $\hat{f}_{b,h_{t,i}}(t)$, we use one of the two risk-aware performance metrics introduced in Section III, namely, either the Mean-Variance or the Conditional Value of Risk. Estimates for the Mean-Variance can be obtained by using the consistent estimators sample mean and sample variance for the expectation and standard deviation, respectively. According to [19], a consistent estimator for the Conditional Value at Risk for beam b and vehicle context hypercube h is given by

$$\hat{\text{CVaR}}_{b,h}(t) = \frac{1}{\lceil \alpha N^{\text{in}}_{b,h}(t) \rceil} \sum_{\tau=1}^{\lceil \alpha N^{\text{in}}_{b,h}(t) \rceil} \tilde{r}^{\tau}_{b,h}, \tag{10}$$

where $\tilde{r}^{1}_{b,h} \leq \ldots \leq \tilde{r}^{N^{\text{in}}_{b,h}(t)}_{b,h}$ denote the last $N^{\text{in}}_{b,h}(t)$ received rewards for beam hypercube combination b and h until time slot t ordered in an increasing fashion.

B. Adaptation to Non-Stationary Environments

The added mechanisms for increasing the adaptation speed are based on the idea of comparing the short-term behavior of the beam performance with its long-term behavior. We refer to long-term memory as the last $N^{\text{in}}_{b,h}(t)$ samples of the rewards, which are also used to compute the estimates $\hat{f}_{b,h}(t)$ of the beam performance. The short-term memory only considers the last t_{st} samples, where t_{st} is an input parameter for the algorithm that models the sensitivity to track non-stationarities. In order to track sudden changes in the statistical properties of the beam performance, we compare estimates of the long-term standard deviation $\hat{\sigma}_{b,h}(t)$ with the short-term standard deviation $\hat{\sigma}^{\text{st}}_{b,h}(t)$, computed with respect to the long-term mean. Specifically, if the short-term standard deviation $\hat{\sigma}^{\text{st}}_{b,h}(t)$ is significantly larger than the long-term standard deviation $\hat{\sigma}_{b,h}(t)$, an adaptation step is triggered. This is expressed by the condition $\hat{\sigma}^{\text{st}}_{b,h}(t) \geq \delta \hat{\sigma}_{b,h}(t)$, where $\delta > 0$ is an input parameter that tunes the adaptation-sensitivity. In case an adaptation step is triggered, the beam performance estimate $\hat{f}_{b,h}(t)$ is updated only taking into account the reward samples in the short-term memory and the internal counter is set to the short-term memory size, i.e., $N^{\text{in}}_{b,h} = t_{\text{st}}$. Furthermore, the external counter is set to $N^{\text{ex}}_{b,h} = \lceil K(t) \rceil$. This avoids that beam b gets re-explored in many succeeding time steps. This is especially important if it has undergone a change in statistical properties that caused a severe performance degradation. We name our algorithm *Adaptive Risk-Aware FML Algorithm* (ARA-FML) and summarize it in Algorithm 1.

Algorithm 1 ARA-FML Algorithm

Input Parameters: T, p_T, $K(t)$, t_{st} and δ
Create partition $\mathcal{P}_T$ of context space $\mathcal{X}$
Internal counter initialization: $N^{\text{in}}_{b,h} = 0, \quad \forall\, b, \forall\, h$
External counter initialization: $N^{\text{ex}}_{b,h} = 0, \quad \forall\, b, \forall\, h$
Estimate initialization: $\hat{\mu}_{b,h} = 0, \quad \forall\, b, \forall\, h$
for each $t = 1, \ldots, T$ **do**
  Receive vehicle contexts $x_{t,i}, \ i = 1, \ldots, V_t$
  Find set of hypercubes $\mathcal{H}_t = \{h_{t,i}\}_{i=1,\ldots,V_t}$ that matches vehicle contexts, i.e., $x_{t,i} \in h_{t,i} \in \mathcal{P}_T, \ \forall\, i$
  Find set of underexplored beams $\mathcal{B}^{\text{ue}}_{\mathcal{H}_t}(t)$ according to (8)
  if $\mathcal{B}^{\text{ue}}_{\mathcal{H}_t}(t) \neq \emptyset$ **then**
    Set $u = |\mathcal{B}^{\text{ue}}_{\mathcal{H}_t}(t)|$
    if $u \geq m$ **then**
      Exploration: Choose m beams $s_{t,1}, \ldots, s_{t,m}$ at random from $\mathcal{B}^{\text{ue}}_{\mathcal{H}_t}(t)$
    else
      Exploration: Choose u beams $s_{t,1}, \ldots, s_{t,u}$ at random from $\mathcal{B}^{\text{ue}}_{\mathcal{H}_t}(t)$
      Exploitation: Choose $(m-u)$ beams $s_{t,u+1}, \ldots, s_{t,m}$ according to (9)
    end if
  else
    Exploitation: Choose m beams $s_{t,1}, \ldots, s_{t,m}$ according to (9)
  end if
  Obtain information about amount of data $r_{j,i}$ that each vehicle $v_{t,i}, \ i = 1, \ldots, V_t$ received from each selected beam $s_{t,j}, \ j = 1, \ldots, m$
  for $i = 1, \ldots, V_t$ **do**
    for $j = 1, \ldots, m$ **do**
      Update long-term variance $\hat{\sigma}^2_{s_{t,j},h_{t,i}}(t)$
      Update short-term variance $\left(\hat{\sigma}^{\text{st}}_{s_{t,j},h_{t,i}}(t)\right)^2$
      if $\hat{\sigma}^{\text{st}}_{s_{t,j},h_{t,i}}(t) \geq \delta \hat{\sigma}_{s_{t,j},h_{t,i}}(t)$ **AND** $N^{\text{in}}_{s_{t,j},h_{t,i}} \geq t_{\text{st}}$ **then**
        Update estimate $\hat{f}_{s_{t,j},h_{t,i}}$ only based on the previous t_{st} samples
        Discard all samples for beam hypercube combination $s_{t,j}, h_{t,i}$ expect the last t_{st} samples
        Reset internal counter to t_{st}: $N^{\text{in}}_{s_{t,j},h_{t,i}} = t_{\text{st}}$
        Reset external counter to $\lceil K(t) \rceil$: $N^{\text{ex}}_{s_{t,j},h_{t,i}} = \lceil K(t) \rceil$
      else
        Update estimate: $\hat{f}_{s_{t,j},h_{t,i}}$
        Update internal counter: $N^{\text{in}}_{s_{t,j},h_{t,i}} = N^{\text{in}}_{s_{t,j},h_{t,i}} + 1$
        Update external counter: $N^{\text{ex}}_{s_{t,j},h_{t,i}} = N^{\text{ex}}_{s_{t,j},h_{t,i}} + 1$
      end if
    end for
  end for
end for

V. Numerical Results

A. Simulation Setup

The scenario consists of a mmWave base station that is assumed to be positioned at 50 meters distance centered to a 250 meter long road-stretch, that is located inside its coverage area. The base station can choose from $B = 8$ beams and selects $m = 2$ per time step. The carrier frequency of the system is set to 28 GHz and the system bandwidth is assumed to be 1 GHz. Additionally, the mmWave base station transmits with a power of 30 dBm and the considered noise power density is $-204 \frac{\text{dBm}}{\text{Hz}}$ according to [12]. Furthermore, the vehicle speed is assumed to be on average between 30 and 70 kilometers per hour. The maximum number of vehicles per time step is set to $V_{\max} = 10$. The vehicle context space $\mathcal{X}$ is assumed to be discrete and one-dimensional. This single dimension describes the approximate location of the vehicle on the road stretch at the beginning of a time step. For that sake, the road is divided into four sections of equal length and the center of that section is used as the approximate location of the respective vehicle. The context is drawn at random for every vehicle at each time step. The time horizon is set to $T = 400$ time steps, where each time step has a duration of 10 seconds. In order to model the random variables that describe the random amount of received data, we use truncated Gaussian distributions. The aim of this choice is to capture the various dynamics that can occur in a vehicular communication scenario, such as random blockages or varying vehicle speeds. The expectations of the Gaussian distributions are calculated with Shannon's channel capacity formula, where the channel gain is determined using the path loss model. The variances of the distributions are drawn randomly for every simulation.

As a performance benchmark, we compare our proposed

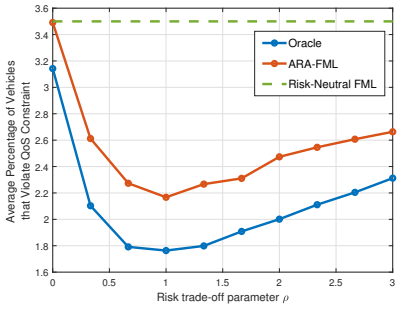

Fig. 2. Mean-Variance: Average percentage of vehicles that violate their Quality-of-Service constraint

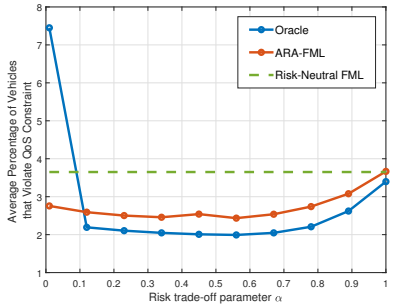

Fig. 3. Conditional Value at Risk: Average percentage of vehicles that violate their Quality-of-Service constraint

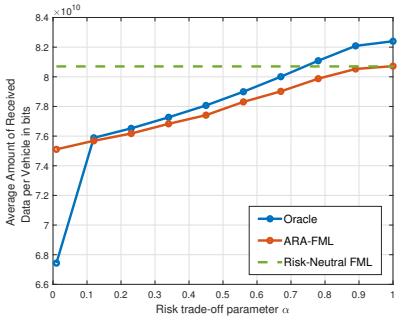

Fig. 4. Conditional Value at Risk: Average amount of received data per vehicle

ARA-FML algorithm to the risk neutral FML algorithm from [12], which does not account for the risk that the vehicles do not fulfill their communication requirements. Additionally, we show the oracle solution. The oracle is an optimal solution because it is an omniscient entity that has perfect knowledge about the statistical properties of the random processes.

B. Risk-Aversion

At first, the effectiveness with respect to the risk-aversion is evaluated. For that sake, the average percentage of vehicles that violate their Quality-of-Service constraint (2) per time step is shown. The averaging of the percentage of vehicles that do not satisfy their Quality-of-Service constraint is performed over all time steps and simulation realizations. Fig. 2 displays the average percentage of vehicles that violate the Quality-of-Service constraint as a function of the risk trade-off parameter ρ if the Mean-Variance is used for the risk-aware assessment of the beam performance. One observes that our proposed risk-aware algorithm ARA-FML is capable of reducing the percentage of vehicles that do not receive enough data from 3.5% to 2.2% compared to the risk-neutral algorithm from [12], for a risk trade-off parameter of $\rho = 1$. Furthermore, ρ should be chosen carefully, as it has a great impact on the performance. For too large values of ρ, the performance degrades, as there is less attention paid to the expectation part of the Mean-Variance. Moreover, our ARA-FML is capable of achieving a performance close to the optimal oracle solution, without having the unrealistic advantage of perfect knowledge about the statistical properties of the beam performance. The performance gap of roughly 23% is caused by the fact that ARA-FML additionally has to perform exploration steps, where it randomly selects suboptimal beams.

A similar behavior can be seen if the Conditional Value at Risk is chosen for the beam performance assessment. This is shown in Fig. 3, which shows the average percentage of vehicles that violate their Quality-of-Service constraint as a function of the risk trade-off parameter α. Note that the poor performance for the oracle solution for small values of the risk trade-off parameter α stems from the fact that in this case, only the expected amount of received data for very unlikely edge cases is considered in the beam performance assessment. This degrades the performance, as the beams are selected for extreme situations that almost never occur. The ARA-FML algorithm outperforms the oracle in these instances, as it does not have sufficiently many samples for a reliable estimation.

Our proposed risk-aversion strategies come with some performance penalty. This is illustrated in Fig. 4, which shows the average received amount of data per vehicle with respect to the risk trade-off parameter α for the Conditional Value at Risk. As the ARA-FML algorithm aims for avoiding potentially poor performing beams, it sacrifices some average performance. However, in Fig. 4, it can be seen that that this loss in average performance compared to risk-neutral FML is at worst 9% for $\alpha =$. In proportion to the gains in terms of risk-aversion, this is small. A similar behavior can be observed for the Mean-Variance, but is not shown here due to space constraints.

C. Adaptation Speed

In the following, the adaptation behavior for non-stationary environments is examined. For this purpose, at $t = 200$, two of the beams undergo an abrupt change of their statistical properties. Specifically, the two beams with the highest mean received amount of data are assumed to suffer from a 70% drop of their expected average amount of received data.

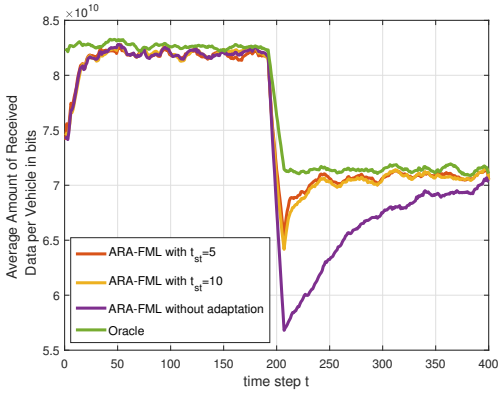

Fig. 5. Average amount of received data per vehicle in dependence of the time step t for the Conditional Value at Risk with $\alpha = 0.4$.

Fig. 5 shows how the average amount of received data per

vehicle evolves over time in case the Conditional Value at Risk with $\alpha = 0.4$ is used to assess the beam performance. In the figure, we compare the oracle solution, our proposed ARA-FML algorithm with a short-term memory size of $t_{\text{st}} = 5$ and $t_{\text{st}} = 10$, as well as the ARA-FML algorithm without the adaptation mechanisms. It can be observed that at $t = 200$, due to the sudden change in the statistical properties, the average amount of received data for the optimal oracle solutions drops significantly. Our proposed ARA-FML algorithm is capable of following this change very quickly, as convergence is reached in approximately the same time as for the initial start of the considered time horizon. Moreover, one observes slight advantages if a smaller short-time memory size is used. Compared to the non-adaptive version of the algorithm, the ARA-FML shows a significantly better behavior, as the initial drop in performance is roughly 50% smaller and the convergence back to the oracle solution happens in about 25% of the time steps needed as compared to the case that the non-adaptive version of the algorithm is used. This can be explained by the fact that the non-adaptive version does not discard outdated samples from before the abrupt performance degradation at $t = 200$. Thus, since it incorporates outdated information for the beam performance estimation, it needs more time to recognize that the performance of the two beams has degraded.

VI. Conclusion

A vehicular communication scenario was considered, in which a 5G mmWave base station serves vehicles with mmWave narrow beams. The problem boils down to finding a policy for the optimal selection of the beams. Our goal was to make the communication reliable and reduce the risk that the individual requirements of the vehicles get not fulfilled. To this aim, we introduced risk-aware metrics for the assessment of the quality of the narrow beams and integrated them into a contextual MAB algorithm. Furthermore, our proposed ARA-FML algorithm is equipped with mechanisms that allow a swift adaptation if the environment is non-stationary. Numerical simulations show that our risk-aversion approach is capable of reducing the likelihood that a vehicle does not receive enough data compared to risk-neutral formulations. Moreover, we show that our proposed algorithm exhibits a high adaptation speed in case of non-stationary environments.

Acknowledgement

This work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Projektnummer 210487104 - SFB 1053 MAKI and the BMBF project Open6GHub, grant number 16KISK014.

References

[1] J. Ni, A. Zhang, X. Lin, and X. S. Shen, "Security, privacy, and fairness in fog-based vehicular crowdsensing," *IEEE Communications Magazine*, vol. 55, no. 6, pp. 146–152, 2017.

[2] M. Boban, A. Kousaridas, K. Manolakis, J. Eichinger, and W. Xu, "Connected roads of the future: Use cases, requirements, and design considerations for vehicle-to-everything communications," *IEEE vehicular technology magazine*, vol. 13, no. 3, pp. 110–123, 2018.

[3] C. R. Storck and F. Duarte-Figueiredo, "A survey of 5G technology evolution, standards, and infrastructure associated with vehicle-to-everything communications by internet of vehicles," *IEEE Access*, vol. 8, pp. 117 593–117 614, 2020.

[4] J. Choi, V. Va, N. Gonzalez-Prelcic, R. Daniels, C. R. Bhat, and R. W. Heath, "Millimeter-wave vehicular communication to support massive automotive sensing," *IEEE Communications Magazine*, vol. 54, no. 12, pp. 160–167, 2016.

[5] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter-wave cellular wireless networks: Potentials and challenges," *Proceedings of the IEEE*, vol. 102, no. 3, pp. 366–385, 2014.

[6] A. Alkhateeb, Y.-H. Nam, M. S. Rahman, J. Zhang, and R. W. Heath, "Initial beam association in millimeter wave cellular systems: Analysis and design insights," *IEEE Transactions on Wireless Communications*, vol. 16, no. 5, pp. 2807–2821, 2017.

[7] M. Hashemi, A. Sabharwal, C. E. Koksal, and N. B. Shroff, "Efficient beam alignment in millimeter wave systems using contextual bandits," in *IEEE Conference on Computer Communications*. IEEE, 2018, pp. 2393–2401.

[8] J. Wang, Z. Lan, C.-w. Pyo, T. Baykas, C.-s. Sum, M. A. Rahman, J. Gao, R. Funada, F. Kojima, H. Harada *et al.*, "Beam codebook based beamforming protocol for multi-gbps millimeter-wave WPAN systems," *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 8, pp. 1390–1399, 2009.

[9] Z. Xiao, T. He, P. Xia, and X.-G. Xia, "Hierarchical codebook design for beamforming training in millimeter-wave communication," *IEEE Transactions on Wireless Communications*, vol. 15, no. 5, pp. 3380–3392, 2016.

[10] W. Wu, N. Cheng, N. Zhang, P. Yang, W. Zhuang, and X. Shen, "Fast mmwave beam alignment via correlated bandit learning," *IEEE Transactions on Wireless Communications*, vol. 18, no. 12, pp. 5894–5908, 2019.

[11] M. B. Booth, V. Suresh, N. Michelusi, and D. J. Love, "Multi-armed bandit beam alignment and tracking for mobile millimeter wave communications," *IEEE Communications Letters*, vol. 23, no. 7, pp. 1244–1248, 2019.

[12] G. H. Sim, S. Klos, A. Asadi, A. Klein, and M. Hollick, "An online context-aware machine learning algorithm for 5G mmwave vehicular communications," *IEEE/ACM Transactions on Networking*, vol. 26, no. 6, pp. 2487–2500, 2018.

[13] A. Sani, A. Lazaric, and R. Munos, "Risk-aversion in multi-armed bandits," *Advances in Neural Information Processing Systems*, vol. 25, pp. 3275–3283, 2012.

[14] X. Liu, M. Derakhshani, S. Lambotharan, and M. Van der Schaar, "Risk-aware multi-armed bandits with refined upper confidence bounds," *IEEE Signal Processing Letters*, vol. 28, pp. 269–273, 2020.

[15] N. Galichet, M. Sebag, and O. Teytaud, "Exploration vs exploitation vs safety: Risk-aware multi-armed bandits," in *Asian Conference on Machine Learning*. PMLR, 2013, pp. 245–260.

[16] S. Vakili and Q. Zhao, "Mean-variance and value at risk in multi-armed bandit problems," in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2015, pp. 1330–1335.

[17] H. Markowitz, "Portfolio selection*," *The Journal of Finance*, vol. 7, no. 1, pp. 77–91, 1952. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1952.tb01525.x

[18] S. Müller, O. Atan, M. van der Schaar, and A. Klein, "Context-aware proactive content caching with service differentiation in wireless networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 2, pp. 1024–1036, 2016.

[19] S. X. Chen, "Nonparametric estimation of expected shortfall," *Journal of financial econometrics*, vol. 6, no. 1, pp. 87–107, 2008.