Beam Training and Data Transmission Optimization in Millimeter-Wave Vehicular Networks

Maria Scalabrin[†], Nicolò Michelusi[‡], and Michele Rossi[†]

Abstract—Future vehicular communication networks call for new solutions to support their capacity demands, by leveraging the potential of the millimeter-wave (mm-wave) spectrum. Mobility, in particular, poses severe challenges in their design, and as such shall be accounted for. A key question in mm-wave vehicular networks is how to optimize the trade-off between directive Data Transmission (DT) and directional Beam Training (BT), which enables it. In this paper, learning tools are investigated to optimize this trade-off. In the proposed scenario, a Base Station (BS) uses BT to establish a mm-wave directive link towards a Mobile User (MU) moving along a road. To control the BT/DT trade-off, a Partially Observable (PO) Markov Decision Process (MDP) is formulated, where the system state corresponds to the position of the MU within the road link. The goal is to maximize the number of bits delivered by the BS to the MU over the communication session, under a power constraint. The resulting optimal policies reveal that adaptive BT/DT procedures significantly outperform common-sense heuristic schemes, and that specific mobility features, such as user position estimates, can be effectively used to enhance the overall system performance and optimize the available system resources.

I. INTRODUCTION

The state-of-the-art protocols for vehicular communication address vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication systems, generally termed V2X. Currently, these communication systems enable a maximum data rate of 100 Mbps for high mobility (using 4G) [1], [2], which are not deemed sufficient to support applications such as autonomous driving, augmented reality and infotainment, which will populate next-generation vehicular networks. Therefore, future vehicular communication networks call for new solutions to support their capacity demands, by leveraging the huge amount of bandwidth in the $30 - 300$ GHz band, the so called millimeter-wave (mm-wave) spectrum. While communication at these frequencies is ideal to support high capacity demands, it relies on highly directional transmissions, which are extremely susceptible to the vehicle mobility. Therefore, a key question is: *How do we leverage mobility information to optimize the trade-off between directive Data Transmission (DT) and directional Beam Training (BT), which enables it, to optimize the communication performance? How much do we gain by doing so?* To address these questions and optimize this trade-off, in this paper we envision the use of learning tools. We demonstrate significant gains compared to common-sense beam alignment schemes.

Compared to conventional lower frequencies, propagation at mm-waves poses several challenges, such as high propagation loss and sensitivity to blockage. To counteract these effects, mm-wave systems are expected to use large antenna arrays to achieve a large beamforming gain via directional transmissions. However, these techniques demand extensive beam training, such as beam sweeping, estimation of angles of arrival and of departure, and data-assisted schemes [3], as well as beam tracking [4]. Despite their simplicity, the overhead incurred by these algorithms may ultimately offset the benefits of beamforming in highly mobile environments [1], [2]. While wider beams require less beam training, they result in a lower beamforming gain, hence smaller achievable capacity [5]. While contextual information, such as GPS readings of vehicles [3], may alleviate this overhead, it does not eliminate the need for beam training due to noise and GPS acquisition inaccuracy. Thus, the design of schemes that alleviate this overhead is of great importance.

In all of the aforementioned works, a priori information on the vehicle's mobility is not leveraged in the design of BT/DT protocols. In contrast, *we contend that leveraging such information via adaptive beam design techniques can greatly improve the performance of automotive networks* [6], [7]. In this paper, we bridge this gap by designing adaptive strategies for BT/DT that leverage a priori mobility information via Partially Observable (PO) Markov Decision Processes (MDPs). Our numerical evaluations demonstrate that these optimized policies significantly outperform common-sense heuristic schemes, which are not tailored to the vehicle's observed mobility pattern. Compared to [8], which develops an analytical framework to optimize the BT/DT trade-off and the BT parameters based on the "worst-case" mobility pattern, in this work, we assume a statistical mobility model.

In the proposed scenario, a Base Station (BS) attempts to establish a mm-wave directive link towards a Mobile User (MU) moving along a road. To this end, it alternates between BT and DT. The goal is to maximize the number of bits delivered by the BS to the MU over the communication session, under a power constraint. To manage the BT/DT trade-off, we exploit a POMDP formulation, where the system state corresponds to the position of the MU within the road link. Specifically, we implement a POMDP with temporally extended actions (i.e., actions with different durations) to model the different temporal scales of BT and DT, and a constraint on the available resources of the system. POMDPs model an agent decision process in which the system dynamics are determined by the underlying MDP (in this case, the MU dynamics), but the agent cannot directly observe the system state. Instead, it maintains a probability distribution

(called *belief*) over the world states, based on observations and their distribution, and the underlying MDP. An exact solution to a POMDP yields the optimal action for each possible belief over the world states. POMDPs have been successfully implemented in a variety of real-world sequential decision processes, including robot navigation problems, machine maintenance, and planning under uncertainty [9], [10]. To address the complexity of POMDPs, we use PERSEUS [11], an approximate solution technique which uses a sub-set of belief points as representative of the belief state. However, in contrast to the original formulation using random belief point selection, we tailor it by selecting a deterministic set of belief points representing uncertainty in MU position, and demonstrate significant performance gains. A unified approach for constrained MDP is given by [12], [13]. Notably, there has been relatively little development in the literature for incorporating constraints into the POMDP [14]–[17]. In order to address the resource constraints in our problem, we propose a Lagrangian method, and an online algorithm to optimize the Lagrangian variable based on the target cost constraint.

This paper is organized as follows. In Section II, we introduce the system model, followed by the optimization in Section III. We present numerical results in Section IV, followed by concluding remarks in Section V.

II. SYSTEM MODEL

We consider a scenario where a BS aims at establishing a mm-wave directive link with a MU moving along a road. To this end, it alternates between BT and DT: with BT, the BS refines its knowledge on the position of the MU within the road link, to perform more directive DT. Our goal is to maximize the number of bits that the BS delivers to the MU during a transmission episode, defined as the time interval between the two instants when the MU enters and exits the coverage range of the BS, under a power constraint. The mm-wave link is in Line-Of-Sight (LOS), thus knowledge on the position of the MU is sufficient for aligning the beams [7].

A. Problem formulation

We consider a *dense* cell deployment, as shown in Fig. 1. The MU is associated with its closest BS, at a distance d_0 from the road link. The road link served by the reference BS is divided into S road sub-links of equal length $\Delta_{\mathrm{s}} = 2d_0 \tan(\Theta/2)/S$, where Θ is the maximum coverage range of the BS. We let $\mathcal{S} \equiv \{1, \ldots, S\}$ be the set of indices of the S road sub-links. The BS associates a beam with each one of the S road sub-links, with angular support, for the s-th beam,

$$\Phi_s = \left[\tan^{-1} \frac{-d_0 \tan(\Theta/2) + (s-1)\Delta_{\mathrm{s}}}{d_0}, \right.$$
$$\left. \tan^{-1} \frac{-d_0 \tan(\Theta/2) + s\Delta_{\mathrm{s}}}{d_0} \right] \quad (1)$$

and beamwidth $\theta_s = |\Phi_s|$, so that $\cup_{s \in \mathcal{S}} \Phi_s = [-\Theta/2, \Theta/2]$ and $\sum_{s \in \mathcal{S}} \theta_s = \Theta$.

The time is discretized into micro time-slots of duration Δ_{t}, with Δ_{t} being the time for a Primary Synchronization Signal (PSS), which allows a proper channel estimation at the receiver [5]. At time t, the MU is located in one of the S road
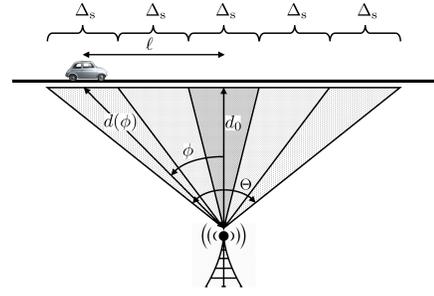

Fig. 1: A *dense* cell deployment.

sub-links, until it exits the coverage area of the BS, denoted by the absorbing state $\bar{s} = S + 1$. We denote the sub-link occupied by the MU at time t as $X_t \in \mathcal{S}$. We assume that the position of the MU within the road link evolves among the S road sub-links following a random walk with probabilities $0 \leq q < p < 1$, with $1 - p - q > 0$, where $p = \mathbb{P}[X_{t+1} = s + 1 \mid X_t = s]$ and $q = \mathbb{P}[X_{t+1} = s - 1 \mid X_t = s]$. Under this model, the MU will exit the BS coverage area at some point. We can view such random walk as an abstraction of the following *physical* mobility model, where the MU moves with average speed $\mathbb{E}[v]$ and speed variance $\mathrm{Var}[v]$: assume that the MU moves at speed v_t at time t, with $v_t \in \{0, v_{\max}, -v_{\max}\}$. Also, let $\mathbb{P}[v_t = v_{\max}] = p$, $\mathbb{P}[v_t = -v_{\max}] = q$, and $\mathbb{P}[v_t = 0] = 1 - p - q$. Note that the maximum speed supported by this model is $v_{\max} \leq \Delta_{\mathrm{s}}/\Delta_{\mathrm{t}}$ (otherwise, the MU may move more than one sub-link within a single micro-slot). It follows that $\mathbb{E}[v] = v_{\max}(p - q) > 0$ and $\mathrm{Var}[v] = v_{\max}^2(p + q) - (\mathbb{E}[v])^2 > 0$. Thus, given average $\mathbb{E}[v]$ and $\mathrm{Var}[v]$, we obtain p and q as

$$p = \frac{\mathrm{Var}[v] + (\mathbb{E}[v])^2}{2v_{\max}^2} + \frac{\mathbb{E}[v]}{2v_{\max}},$$
$$q = \frac{\mathrm{Var}[v] + (\mathbb{E}[v])^2}{2v_{\max}^2} - \frac{\mathbb{E}[v]}{2v_{\max}}. \quad (2)$$

To meet the conditions for the probabilities $0 \leq q < p < 1$, with $1 - p - q > 0$, the following inequalities must hold:

$$p < 1 \rightarrow \mathrm{Var}[v] < -(\mathbb{E}[v])^2 - v_{\max}\mathbb{E}[v] + 2v_{\max}^2,$$
$$q \geq 0 \rightarrow \mathrm{Var}[v] \geq -(\mathbb{E}[v])^2 + v_{\max}\mathbb{E}[v], \quad (3)$$
$$1 - p - q > 0 \rightarrow \mathrm{Var}[v] + (\mathbb{E}[v])^2 < v_{\max}^2,$$

which defines a region of feasible pairs $(\mathbb{E}[v], \mathrm{Var}[v])$. This model can be extended, e.g., to account for memory in the velocity process, although we leave it for future work.

During BT or DT, at time t, the BS transmits using a beam that covers a sub-set of sub-links, $\hat{\mathcal{S}}_t \subseteq \mathcal{S}$, part of our design. Assuming a large antenna array, which allows for arbitrarily sharp beam patterns (i.e., the beams are much larger than the minimum resolution of the antenna array), the beam is designed in order to support a target SNR SNR_t on the beam support, $\mathcal{B}_t \equiv \cup_{s \in \hat{\mathcal{S}}_t} \Phi_s$. To this end, we let $P_t(\phi)$ be the power per radian projected in the angular direction $\phi \in \mathcal{B}_t$, and $P_t(\phi) = 0, \phi \notin \mathcal{B}_t$. To attain the target SNR constraint, we must have that

$$\frac{\Gamma P_t(\phi)}{d(\phi)^2} = \mathrm{SNR}_t, \quad (4)$$

where $\Gamma \triangleq \lambda^2 \xi / (8\pi N_0 W_{\text{tot}})$ is the SNR scaling factor, $\lambda = f_c/c$ is the wavelength, N_0 is the noise power spectral density, ξ is the antenna efficiency, W_{tot} is the bandwidth, and $d(\phi) = d_0 \sqrt{1 + \tan(\phi)^2}$ is the distance of the point in the road link at angular direction ϕ, so that $d(\phi)^{-2}$ models distance dependent path loss. It follows that the total transmit power is given by

$$P_t = \int_{\mathcal{B}_t} P_t(\phi) \mathrm{d}\phi = \text{SNR}_t \sum_{s \in \hat{\mathcal{S}}_t} \int_{\Phi_s} \frac{d_0^2}{\Gamma} [1 + \tan(\phi)^2] \mathrm{d}\phi. \quad (5)$$

Using the change of variables $\phi \rightarrow \ell = d_0 \tan(\phi)$, we obtain

$$P_t = \text{SNR}_t \frac{1}{\Gamma} \sum_{s \in \hat{\mathcal{S}}_t} \int_{-d_0 \tan(\Theta/2) + (s-1)\Delta_s}^{-d_0 \tan(\Theta/2) + s\Delta_s} d_0 \mathrm{d}\ell = \text{SNR}_t \frac{\Delta_s d_0}{\Gamma} |\hat{\mathcal{S}}_t|. \quad (6)$$

In other words, the total transmit power is independent of the sub-link indices, but depends solely on the number of sub-links $|\hat{\mathcal{S}}_t|$ and on the target SNR. This result is in line with the intuition that larger distances are achievable via smaller beamwidths, and vice versa [7].

During DT, assuming isotropic reception at the MU, such target SNR implies an achievable rate given by

$$R_t = W_{\text{tot}} \log_2 \left(1 + \frac{\Gamma}{\Delta_s d_0} \frac{P_t}{|\hat{\mathcal{S}}_t|} \right). \quad (7)$$

During BT, the SNR is set so as to achieve target mis-detection and false-alarm probabilities. To design this parameter, the generic signal detection problem corresponds to receiving a signal $y[l]$, $l = 1, \dots, L$, over a noisy channel. The two hypotheses are

$$\begin{aligned} &\mathcal{H}_0 : y[l] = w[l] &&\text{(no signal at the RX)} \\ &\mathcal{H}_1 : y[l] = x[l] + w[l] &&\text{(signal at the RX)} \end{aligned} \quad (8)$$

where $w[l]$, $l = 1, \dots, L$, are independent random variables, $w[l] \sim \mathcal{N}(0, \sigma_w^2)$, with $\sigma_w^2 = N_0$. Our task is to decide in favor of $\mathcal{H}_0$ or $\mathcal{H}_1$ on the basis of the measurements $y[l]$, $l = 1, \dots, L$, i.e.,

$$\mathbb{P}(y[1], \dots, y[L] | \mathcal{H}_1) \mathbb{P}(\mathcal{H}_1) \gtrless \mathbb{P}(y[1], \dots, y[L] | \mathcal{H}_0) \mathbb{P}(\mathcal{H}_0),$$

or equivalently,

$$\sum_{l=1}^{L} y[l] x[l] \gtrless \sigma_w^2 \ln \left(\frac{\mathbb{P}(\mathcal{H}_0)}{\mathbb{P}(\mathcal{H}_1)} \right) + \frac{1}{2} E_x, \quad (9)$$

where $E_x = \sum_{l=1}^{L} x[l]^2$ is the energy of the pilot signal $x[l]$. If the Neyman-Pearson formulation is used, then the right hand side of Eq. (9) is replaced by a decision threshold $\bar{\rho}$, function of the target error probability. According to the Neyman-Pearson Lemma [18], for a given target error probability, we can derive a decision rule as follows. The false-alarm probability, $\mathbb{P}_{\text{FA}}$ (accept $\mathcal{H}_1$ when $\mathcal{H}_0$ is true), is given as $\mathbb{P}_{\text{FA}} = \int_{\bar{\rho}}^{\infty} \mathbb{P}(y | \mathcal{H}_0) dy = Q(\frac{\bar{\rho}}{\sigma_w \sqrt{E_x}})$, where $Q(\cdot)$ is the Q-function. The mis-detection probability, $\mathbb{P}_{\text{MD}}$ (accept $\mathcal{H}_0$ when $\mathcal{H}_1$ is true), is given as $\mathbb{P}_{\text{MD}} = 1 - \mathbb{P}_D$, where the probability of correct detection is given by $\mathbb{P}_D = \int_{\bar{\rho}}^{\infty} \mathbb{P}(y | \mathcal{H}_1) dy = Q(\frac{\bar{\rho} - E_x}{\sigma_w \sqrt{E_x}}) = Q(Q^{-1}(\mathbb{P}_{\text{FA}}) - \frac{\sqrt{E_x}}{\sigma_w})$, which shows

that $\mathbb{P}_D$ is a function of $\mathbb{P}_{\text{FA}}$. Applying the inverse $Q^{-1}(\cdot)$ to both sides of the last equation, leads to a measure of the SNR required to attain the target error performance:

$$\text{SNR}_t = \frac{E_x}{\sigma_w^2} = (Q^{-1}(\mathbb{P}_{\text{FA}}) - Q^{-1}(\mathbb{P}_D))^2, \quad (10)$$

which is plugged into Eq. (6) to find the transmit power as a function of the number of sub-links covered, $|\hat{\mathcal{S}}_t|$.

B. Partially Observable Markov Decision Process

Next, we define a constrained Partially Observable (PO) Markov Decision Process (MDP).

States: $\bar{\mathcal{S}}$ is a finite set of states describing the position of the MU within the road link, along with the absorbing state $\bar{s}$ when the MU exits the coverage area of the BS. Therefore, $\bar{\mathcal{S}} \equiv \mathcal{S} \cup \{\bar{s}\}$, and $\mathcal{S} \equiv \{1, \dots, S\}$ is the set of road sub-links.

Actions: $\mathcal{A}$ is a finite set of actions that the BS can perform. Specifically, the BS can perform actions for Beam Training (BT) and actions for Data Transmission (DT), which involve selection of the transmission beam, power, and duration. In general, $a \in \mathcal{A}$ is in the form $a = (\hat{\mathcal{S}}, P_c, T_c)$, where: $c = \{\text{BT}, \text{DT}\}$ refers to the action class; $\hat{\mathcal{S}} \subseteq \mathcal{S}$ is a sub-set of sub-links, defining the support of the transmission beam; P_c is the transmission power per beam, such that $P_t = P_c |\hat{\mathcal{S}}|$ in Eq. (6) is the total transmit power, T_c is the transmission duration of action $a \in \mathcal{A}$ (number of micro time-slots of duration Δ_t). If $c = \text{BT}$, then $a = (\hat{\mathcal{S}}, P_{\text{BT}}, T_{\text{BT}})$, where P_{BT} and T_{BT} are fixed parameters of the model. Specifically, we assume that BT actions perform simultaneous beamforming over $\hat{\mathcal{S}}$ in one interval of Δ_t seconds (i.e., $T_{\text{BT}} = 1$). Also, $P_{\text{BT}} = P_{\text{min}}$, where P_{min} is the power per beam required to attain the target SNR constraint, i.e., $P_{\text{min}} = \text{SNR}\frac{\Delta_s d_0}{\Gamma}$, and SNR is a function of false-alarm and mis-detection probabilities $\mathbb{P}_{\text{FA}}$ and $\mathbb{P}_{\text{MD}}$, which are also fixed parameters of the model, via Eq. (10). If $c = \text{DT}$, then $a = (\hat{\mathcal{S}}, P_{\text{DT}}, T_{\text{DT}})$, where P_{DT} and T_{DT} are part of the optimization. Specifically, we assume that DT actions perform simultaneous data communication over $\hat{\mathcal{S}}$ for $T_{\text{DT}} - 1$ micro time-slots, where the last interval of Δ_t seconds is dedicated to the ACK/NACK feedback transmission from the MU to the BS. During DT, the transmission rate follows from Eq. (7). Note that the action space grows as $|\hat{\mathcal{S}}| = \sum_{s=1}^{S} S! / s! / (S-s)! = 2^S - 1$. To reduce its cardinality, we restrict $\mathcal{A}$ such that $\hat{\mathcal{S}}$ is a sub-set of *consecutive* indices in $\mathcal{S}$, i.e., the beam directions specified by $\hat{\mathcal{S}}$ define a compact range of transmission for the BS. Thus, $|\hat{\mathcal{S}}| = S(S+1)/2$.

Observations: $\mathcal{O}$ is a finite set of observations, defined as $\mathcal{O} \equiv \{\text{ACK}, \text{NACK}, \bar{s}\}$. Specifically, $o = \bar{s}$ means that the MU exited the coverage area of the BS; for simplicity, in this work we assume that such event is observable, i.e., the BS knows when the MU exited its coverage area.

Transition probabilities: $\mathbb{P}(s'|s, a)$ is the transition probability from $s \in \bar{\mathcal{S}}$ to $s' \in \bar{\mathcal{S}}$ given $a \in \mathcal{A}$. Note that these probabilities are a function of the duration T_c of action a. If the transmission duration of $a \in \mathcal{A}$ is $T_c = 1$, then we store the 1-step transition probabilities into matrix $\mathbf{M} = [\mathbf{M}_{ss'}]$, with elements $\mathbf{M}_{ss'} = \mathbb{P}(s'|s, a)$ given by the 1-step mobility

model, as:

$$\mathbb{P}(s'|s,a) = p \qquad\qquad s' = s+1,\ s = 1,\ldots,S$$
$$\mathbb{P}(s'|s,a) = q \qquad\qquad s' = s-1,\ s = 2,\ldots,S$$
$$\mathbb{P}(s'|s,a) = 1-p-q \qquad s' = s,\ s = 2,\ldots,S \qquad (11)$$
$$\mathbb{P}(s'|s,a) = 1-p \qquad\qquad s' = s,\ s = 1$$
$$\mathbb{P}(s'|s,a) = 1 \qquad\qquad\quad s' = s,\ s = \bar{s}.$$

If the transmission duration of $a \in \mathcal{A}$ is $T_c = N$, then we compute the N-step transition probabilities into matrix $\mathbf{M}^N$, i.e., we take the N-th power of matrix $\mathbf{M}$ so as to account for the N-step evolution of the system state under $a \in \mathcal{A}$ with transmission duration T_c.

Observation model: $\mathbb{P}(o|s,a,s')$ is the probability of observing $o \in \mathcal{O}$ given $s \in \bar{\mathcal{S}}$ and $a \in \mathcal{A}$ with transmission duration T_c, ending in $s' \in \bar{\mathcal{S}}$. We assume that the BS can successfully perform $a = (\hat{\mathcal{S}}, P_c, T_c)$ if the MU remains within $\hat{\mathcal{S}}$ for T_c subsequent micro time-slots, i.e., the MU does not exit from the beam support, so that all signal is received. In this case, the MU feeds back an ACK to the BS, $o = \text{ACK}$. Therefore, we define $X_0^{T_c} = \{X_0, \ldots, X_{T_c}\}$, as the system state path from time 0 to time T_c, and the event $E_{s,s'}^N = \{X_0^{T_c} \in \hat{\mathcal{S}}^{T_c+1} \mid X_0 = s, X_{T_c} = s', T_c = N\}$, meaning that the system state path $X_0^{T_c}$ remains within $\hat{\mathcal{S}}$ for T_c subsequent micro time-slots, given that $X_0 = s$, $X_{T_c} = s'$, $T_c = N$. In order to compute it, we also define matrix $\tilde{\mathbf{M}}$ as the transition probability matrix restricted to the beam support $\hat{\mathcal{S}}$, i.e., $\tilde{\mathbf{M}} = [\tilde{\mathbf{M}}_{ss'}]$, with elements $\tilde{\mathbf{M}}_{ss'} = \mathbf{M}_{ss'}$ if $\{s, s'\} \in \hat{\mathcal{S}}$, otherwise $\tilde{\mathbf{M}}_{ss'} = 0$. We derive $\mathbb{P}(E_{s,s'}^N)$ as:

$$\mathbb{P}(E_{s,s'}^N) = \mathbb{P}(X_0^{T_c} \in \hat{\mathcal{S}}^{T_c+1} \mid X_0 = s, X_{T_c} = s', T_c = N)$$
$$= \frac{1}{\mathbf{M}_{ss'}^N}\mathbb{P}(X_0^{T_c} \in \hat{\mathcal{S}}^{T_c+1}, X_{T_c} = s' \mid X_0 = s, T_c = N)$$
$$= \frac{1}{\mathbf{M}_{ss'}^N} \sum_{s_0^N \in \hat{\mathcal{S}}^{N+1}} \prod_{z=0}^{N-1} \mathbb{P}(X_{z+1}=s_{z+1} \mid X_z=s_z) = \frac{\tilde{\mathbf{M}}_{ss'}^N}{\mathbf{M}_{ss'}^N}. \quad (12)$$

Given $a \in \mathcal{A}$ with transmission duration T_c, $\mathbb{P}(o|s,a,s')$ is defined as follows. If $c = \text{BT}$, we account for false-alarm and mis-detection errors in the beam detection process. In particular, if $\{s, s'\} \in \hat{\mathcal{S}}$ (i.e., the MU is within the beam support during the duration of BT) then $\mathbb{P}(\text{ACK}|s,a,s') = \mathbb{P}_\text{D}$ (correct detection) and $\mathbb{P}(\text{NACK}|s,a,s') = \mathbb{P}_\text{MD}$ (mis-detection); on the other hand, if $\{s, s'\} \notin \hat{\mathcal{S}}$ (i.e., the MU is outside of the beam support during the duration of BT), then $\mathbb{P}(\text{ACK}|s,a,s') = \mathbb{P}_\text{FA}$ (false-alarm) and $\mathbb{P}(\text{NACK}|s,a,s') = 1 - \mathbb{P}_\text{FA}$. If $c = \text{DT}$, then the transmission is successful if the event $E_{s,s'}^N$ occurs, so that $\mathbb{P}(\text{ACK}|s,a,s') = \mathbb{P}(E_{s,s'}^N)$ for $\{s, s'\} \in \hat{\mathcal{S}}$, and $\mathbb{P}(\text{NACK}|s,a,s') = 1 - \mathbb{P}(\text{ACK}|s,a,s')$. Finally, $\mathbb{P}(\bar{s}|s,a,s') = 1$ whenever either $s = \bar{s}$ or $s' = \bar{s}$, i.e., the BS knows when the MU exited its coverage area.

Rewards: $r(s,a)$ is the expected reward given $s \in \bar{\mathcal{S}}$ and $a \in \mathcal{A}$, defined as the transmission rate (number of bits transmitted from the BS to the MU) during DT if the MU remains within $\hat{\mathcal{S}}$ for T_c subsequent micro time-slots. Formally, $r(s,a) = \sum_{s' \in \bar{\mathcal{S}}} \mathbb{P}(s'|s,a)r(s,a,s') = \sum_{s' \in \bar{\mathcal{S}}} \mathbf{M}_{ss'}^N \mathbb{E}[R(T_\text{DT}-1)\mathcal{X}(E_{s,s'}^N)]$, where $\mathcal{X}(E_{s,s'}^N) = 1$ iff

the event $E_{s,s'}^N$ is true (thus $r(s,a) = 0$ if the MU exits from the beam support). Note that $\mathbb{E}[\mathcal{X}(E_{s,s'}^N)] = \mathbb{P}(E_{s,s'}^N)$, which is computed in Eq. (12). The transmission rate R follows from Eq. (7) when $P_t = P_\text{DT}|\hat{\mathcal{S}}|$. Finally, $r(s,a) = R(T_\text{DT}-1)\sum_{s' \in \bar{\mathcal{S}}} \tilde{\mathbf{M}}_{ss'}^N$, where $T_{DT}-1$ refers to the fact that we reserve one micro time-slot over the total DT duration for the feedback transmission. If $c = \text{BT}$, then $r(s,a) = 0$, as no bits of data are transmitted.

Costs: $c(s,a)$ is the expected energy cost given $s \in \bar{\mathcal{S}}$ and $a \in \mathcal{A}$. The total expected cost during a transmission episode is subject to the constraint C. If $c = \text{DT}$, then $c(s,a) = P_\text{DT}|\hat{\mathcal{S}}|(T_\text{DT} - 1)$, $\forall s \in \mathcal{S}$ (we reserve one micro time-slot for the feedback transmission). If $c = \text{BT}$, then $c(s,a) = P_\text{BT}|\hat{\mathcal{S}}|$. In this POMDP formulation, the overhead cost given by BT with respect to DT is implicitly modelled with the total fraction of time spent by BT with respect to DT.

III. OPTIMIZATION PROBLEM

Since the agent cannot directly observe the system state, we introduce the notion of *belief*. A belief $b \in \mathcal{B}$ is a probability distribution over $\bar{\mathcal{S}}$. The state estimator must compute a new belief, $b' \in \mathcal{B}$, given an old belief $b \in \mathcal{B}$, an action $a \in \mathcal{A}$, and an observation $o \in \mathcal{O}$, i.e., $b' = f(o, a, b)$. It can be obtained via Bayes' rule as:

$$b'(s') = \mathbb{P}(s' \mid o, a, b)$$
$$= \frac{\sum_{s \in \bar{\mathcal{S}}} \mathbb{P}(o|s,a,s')\mathbb{P}(s'|s,a)b(s)}{\mathbb{P}(o|a,b)}. \quad (13)$$

where $\mathbb{P}(o \mid a, b)$ is a normalizing factor, $\sum_{s' \in \bar{\mathcal{S}}} b'(s') = 1$.

Our goal is to determine a policy π (i.e., a map from beliefs to actions) that maximizes the total expected reward the agent can gather, under a constraint on the total expected cost during a transmission episode, following π and starting from $b_0 = b$:

$$\max_\pi \mathbb{E}_\pi\left[\sum_{t=0}^\infty r_t(b_t, \pi(b_t)) \mid b_0 = b\right],$$
$$\text{s.t. } \mathbb{E}_\pi\left[\sum_{t=0}^\infty c_t(b_t, \pi(b_t)) \mid b_0 = b\right] \leq C \qquad (14)$$

where we have defined expected rate and cost metrics under belief b_t as

$$r_t(b_t, \pi(b_t)) = \sum_{s \in \bar{\mathcal{S}}} r_t(s, \pi(b_t))b_t(s)$$
$$c_t(b_t, \pi(b_t)) = \sum_{s \in \bar{\mathcal{S}}} c_t(s, \pi(b_t))b_t(s). \qquad (15)$$

At this point, we opt for a Lagrangian relaxation approach such that $\mathcal{L}(s,a) = r(s,a) - \lambda c(s,a)$ is the metric to be maximized, for some Lagrangian multiplier $\lambda \geq 0$, and the total expected cost during a transmission episode is subject to the constraint C. Hereinafter, according to the notation, $\mathcal{L}_t(b_t, \pi(b_t)) = \sum_{s \in \bar{\mathcal{S}}} \mathcal{L}_t(s, \pi(b_t))b_t(s)$. At the end of Section III-B, we will consider an online algorithm to optimize parameter λ so as to solve the original problem in Eq. (14).

A. Value Iteration for POMDPs

In POMDPs, a policy π is a function over a continuous set $\mathcal{B}$ of probability distributions over $\bar{\mathcal{S}}$. A policy π is characterized

by a value function $V^\pi : \mathcal{B} \to \mathbb{R}$, which is defined as:

$$V^\pi(b) = \mathbb{E}_\pi \left[\sum_{t=0}^\infty \mathcal{L}_t(b_t, \pi(b_t)) \mid b_0 = b \right]. \qquad (16)$$

A policy π that maximizes V^π is called an optimal policy π^*. The value of an optimal policy π^* is the optimal value function V^*, that satisfies the Bellman optimality equation $V^* = HV^*$ (with Bellman backup operator H):

$$V^*(b) = \max_{a \in \mathcal{A}} \left[\sum_{s \in \bar{\mathcal{S}}} b(s) \mathcal{L}(s, a) + \sum_{o \in \mathcal{O}} \mathbb{P}(o|a, b) V^*(b') \right], \quad (17)$$

where $b' = f(o, a, b)$ (see Eq. (13)). When Eq. (17) holds for every belief $b \in \mathcal{B}$ we are ensured the solution is optimal. V^* can be arbitrarily well approximated by iterating over a number of stages, at each stage considering a step further into the future. Also, for problems with an infinite planning horizon, V^* can be approximated, to any degree of accuracy, by a PieceWise Linear and Convex (PWLC) value function [11]. Thus, $V_{n+1} = HV_n$ and we parameterize a value function V_n at stage n by a finite set of vectors (hyperplanes) $\{\alpha_n^i\}_{i=1}^{|V_n|}$, such that $V_n(b) = \max_{\{\alpha_n^i\}_{i=1}^{|V_n|}} b \cdot \alpha_n^i$, where $(\cdot)$ denotes inner product. Each vector in $\{\alpha_n^i\}_{i=1}^{|V_n|}$, is associated with an action $a(\alpha_n^i) \in \mathcal{A}$, which is the optimal one to take at stage n, and defines a region in the belief space for which this vector is the maximizing element of V_n (thus $\pi(b) = a(\alpha_n^i)$). The key idea is that for a given value function V_n at stage n and a belief $b \in \mathcal{B}$, we can compute the vector α_{n+1}^b in $\{\alpha_{n+1}^i\}_{i=1}^{|HV_n|}$ such that:

$$\alpha_{n+1}^b = \underset{\{\alpha_{n+1}^i\}_i^{|HV_n|}}{\arg\max} \; b \cdot \alpha_{n+1}^i , \qquad (18)$$

where $\{\alpha_{n+1}^i\}_{i=1}^{|HV_n|}$ is the (unknown) set of vectors for HV_n. We will denote this operation $\alpha_{n+1}^b = \text{backup}(b)$. It computes the optimal vector for a given belief $b \in \mathcal{B}$ by back-projecting all vectors in the current horizon value function one step from the future and returning the vector that maximizes the value of $b \in \mathcal{B}$. Defining vectors $\mathcal{L}_a$ such that $\mathcal{L}_a(s) = \mathcal{L}(s, a)$ and $g_{a,o}^i$ such that $g_{a,o}^i(s) = \sum_{s' \in \bar{\mathcal{S}}} \mathbb{P}(o|s, a, s') \mathbb{P}(s'|s, a) \alpha_n^i(s')$ ($g_{a,o}^i$ is a projected vector given action a, observation o, and current horizon vector α_n^i), we have [11]:

$$\begin{aligned}
\text{backup}(b) &= \underset{\{g_a^b\}_{a \in \mathcal{A}}}{\arg\max} \, b \cdot g_a^b \\
&= \underset{\{g_a^b\}_{a \in \mathcal{A}}}{\arg\max} \, b \cdot \left[\mathcal{L}_a + \sum_{o \in \mathcal{O}} \underset{\{g_{a,o}^i\}_i}{\arg\max} \, b \cdot g_{a,o}^i \right].
\end{aligned} \qquad (19)$$

In general, computing optimal planning solutions for POMDPs is an intractable problem for any reasonably sized task. This calls for approximate solution techniques, e.g., PERSEUS [11], which we introduce next.

B. Randomized Point-based Value Iteration for POMDPs

PERSEUS is an approximate Point-Based Value Iteration (PBVI) algorithm for POMDPs. It implements a randomized approximate backup operator $\tilde{H}$ that increases (or at least does not decrease) the value of all beliefs $b \in \tilde{\mathcal{B}} \subset \mathcal{B}$. The key idea is that for a given value function V_n at stage n, we can build

a value function $V_{n+1} = \tilde{H}V_n$ that improves the value of all beliefs $b \in \tilde{\mathcal{B}} \subset \mathcal{B}$ by only updating the value of a (randomly selected) subset of beliefs $\check{\mathcal{B}} \subset \mathcal{B}$, i.e., we can build a value function $V_{n+1} = \tilde{H}V_n$ that upper bounds V_n over $\tilde{\mathcal{B}} \subset \mathcal{B}$ (but not necessarily over $\mathcal{B}$): $V_n(b) \leq V_{n+1}(b), \forall b \in \tilde{\mathcal{B}} \subset \mathcal{B}$. Starting with V_0, PERSEUS performs a number of backup stages until some convergence criterion is met. Each backup stage is defined as in Algorithm 1 (where $\tilde{\mathcal{B}}_{\text{temp}}$ is an auxiliary set containing the non-improved beliefs).

Algorithm 1 function PERSEUS

1: **function** PERSEUS($\tilde{\mathcal{B}}, V_n$)
2:      Set $V_{n+1} = 0$. Initialize $\tilde{\mathcal{B}}_{\text{temp}}$ to $\tilde{\mathcal{B}}$.
3:      **while** $\tilde{\mathcal{B}}_{\text{temp}} \neq \emptyset$ **do**
4:          Sample b from $\tilde{\mathcal{B}}_{\text{temp}}$
5:          Compute $\alpha = \text{backup}(b)$ (see Eq. (19))
6:          **if** $b \cdot \alpha \geq V_n(b)$ **then**
7:              Add α to V_{n+1}
8:          **else**
9:              Add $\alpha' = \arg\max_{\{\alpha_n^i\}_i^{|V_n|}} b \cdot \alpha_n^i$ to V_{n+1}
10:          **end if**
11:          Compute set $\tilde{\mathcal{B}}_{\text{temp}} = \{b \in \tilde{\mathcal{B}} : V_{n+1}(b) < V_n(b)\}$
12:      **end while**
13:      **return** V_{n+1}
14: **end function**

Algorithm 2 function BELIEFS

1: **function** BELIEFS
2:      $\tilde{\mathcal{B}} = \emptyset$
3:      **for** $w = 1 : W$ **do**
4:          **for** $i = 1 : S + 1 - w$ **do**
5:              Build b such that $b_{i:i+w-1} = 1/w$ and $\tilde{\mathcal{B}} \leftarrow b$
6:          **end for**
7:      **end for**
8:      **return** $\tilde{\mathcal{B}}$
9: **end function**

Key to the performance of PERSEUS is the design of $\tilde{\mathcal{B}}$. Several standard schemes to select beliefs have been proposed for PBVI, mainly based on grids of points in the belief space. A different option to select beliefs is to simulate the model, i.e., sampling random actions and observations, and generating trajectories through the belief space, as suggested in [11]. Although this approach may seem reasonable, one may argue that the probability distributions collected in $\tilde{\mathcal{B}}$ are not very representative of the system dynamics history, where actions and observations must also depend on beliefs. Hereinafter, we leverage the structure of the POMDP presented in Section II and provide an algorithm (Algorithm 2) to collect beliefs in $\tilde{\mathcal{B}}$ in a smarter fashion. Our approach is simple but effective, and does not require any prior knowledge of the system dynamics: according to Algorithm 2, $\tilde{\mathcal{B}}$ is made of uniform probability distributions over $\bar{\mathcal{S}}$, which are uniformly distributed over at most W *consecutive* road sub-links. Then, this design of $\tilde{\mathcal{B}}$ reflects the compact range of transmission for the BS, where the BS degree of uncertainty on the MU state scales with W.

The basic routine for PBVI is given in Algorithm 3, where $V_{n+1} = V_{n+1}^r - \lambda V_{n+1}^c$ approximates the optimal value function for a given value of λ. Note that we are interested in the optimal policy π^* when b_0 is such that $b_0(s) = \delta(s-1)$, i.e., the agent knows when the MU enters the maximum coverage range of the BS.

Algorithm 3 Point-Based Value Iteration (PBVI)

1: $\tilde{\mathcal{B}} = $ BELIEFS
2: Set $n = 0$, $V_0 = 0$, $V_0^c = 0$, $\lambda_0 = \lambda$, $V_{\text{opt}} = -\infty$.
3: Define $\mathcal{L}(s,a) = r(s,a) - \lambda_0 c(s,a)$
4: **for** $n = 0, \ldots$ **do**
5: $\quad V_{n+1} = $ PERSEUS$(\tilde{\mathcal{B}}, V_n)$
6: $\quad$ **if** $|\sum_{b \in \tilde{\mathcal{B}}} V_{n+1}(b) / \sum_{b \in \tilde{\mathcal{B}}} V_n(b) - 1| < \epsilon_V$ **then**
7: $\quad\quad$ Break
8: $\quad$ **end if**
9: **end for**
10: $V^*(b_0) = V_{n+1}(b_0)$
11: $V_c^*(b_0) = V_{n+1}^c(b_0)$
12: **if** $V_c^*(b_0) < C$ and $V^*(b_0) > V_{\text{opt}}$ **then**
13: $\quad \lambda_{\text{opt}} = \lambda_0$
14: $\quad V_{\text{opt}} = V^*(b_0)$
15: **end if**

Algorithm 4 Point-Based Value Iteration (PBVI) - ONLINE

1: $\tilde{\mathcal{B}} = $ BELIEFS
2: Set $n = 0$, $V_0 = 0$, $V_0^c = 0$, $\lambda_0 = \lambda$, $\alpha_0 = \alpha$.
3: Define $\mathcal{L}(s,a) = r(s,a) - \lambda_0 c(s,a)$
4: **for** $n = 0, \ldots$ **do**
5: $\quad V_{n+1} = $ PERSEUS$(\tilde{\mathcal{B}}, V_n)$
6: $\quad$ **if** $|\sum_{b \in \tilde{\mathcal{B}}} V_{n+1}(b) / \sum_{b \in \tilde{\mathcal{B}}} V_n(b) - 1| < \epsilon_V$ **then**
7: $\quad\quad$ **if** $(V_{n+1}^c(b_0) - C)/C < \epsilon_c$ **then**
8: $\quad\quad\quad$ Break
9: $\quad\quad$ **end if**
10: $\quad$ **end if**
11: $\quad \lambda_{n+1} = \max(0, \lambda_n + \alpha_n(V_{n+1}^c(b_0) - C))$
12: $\quad$ Define $\mathcal{L}(s,a) = r(s,a) - \lambda_{n+1} c(s,a)$
13: **end for**
14: $V^*(b_0) = V_{n+1}(b_0)$
15: $V_c^*(b_0) = V_{n+1}^c(b_0)$
16: $\lambda_{\text{opt}} = \lambda_n$
17: $V_{\text{opt}} = V^*(b_0)$

To find the optimal multiplier λ_{opt}, we have to run the routine for different values of λ. PERSEUS performs a number of backup stages until some convergence criterion is met. At this point, we check if the constraint $V_c^*(b_0) < C$ is satisfied, update λ_{opt} if $V^*(b_0) > V_{\text{opt}}$, and repeat the routine for different values of λ. These values of λ can be sequentially selected from a sorted sequence or properly tuned at the end of the routine in a smarter fashion. However, in both cases we have to wait until convergence. To speed up the search for the optimal multiplier λ_{opt}, we formulate an online version of Algorithm 3, which is presented in Algorithm 4. Here, λ is properly tuned within the main loop

of the routine according to a gradient descent technique [19] [1]: $\lambda_{n+1} = \max(0, \lambda_n + \alpha_n(V_{n+1}^c(b_0) - C))$, where the discount factor is $\alpha_n = \alpha_0/(n+1)$. Finally, given λ_{n+1}, we update the Lagrangian relaxation as $\mathcal{L}(s,a) = r(s,a) - \lambda_{n+1} c(s,a)$. In addition to the convergence criterion of the standard PBVI, we also consider the requirement $(V_{n+1}^c(b_0) - C)/C < \epsilon_c$.

IV. NUMERICAL RESULTS

We set $\Delta_t = 10\ \mu$s. We consider the following parameters: $\Theta = 120°$, $d_0 = 10$ m, $W_{\text{tot}} = 400$ MHz, $f_c = 60$ GHz, $\xi = 1$, $N_0 = -174$ dBm, $S = 10$, $\mathbb{E}[v] = 20$ m/s, $P_{\text{DT}} \in \{10, 20, 30\}$ dBm, $T_{\text{DT}} \in \{1000, 2000, 3000\}$ (number of micro time-slots, i.e., $\{10, 20, 30\}$ ms), $T_{\text{BT}} = 1$, $P_{\text{BT}} = P_{\min}$, where $P_{\min}$ follows from $\mathbb{P}_{\text{FA}} = \mathbb{P}_{\text{MD}} = \epsilon$ (specified below). Finally, we compare different sets $\tilde{\mathcal{B}}$. Let D be the average duration of a transmission episode. The average rate (bit/s) and power (dBm) are computed as V_{n+1}^r/D and V_{n+1}^c/D.
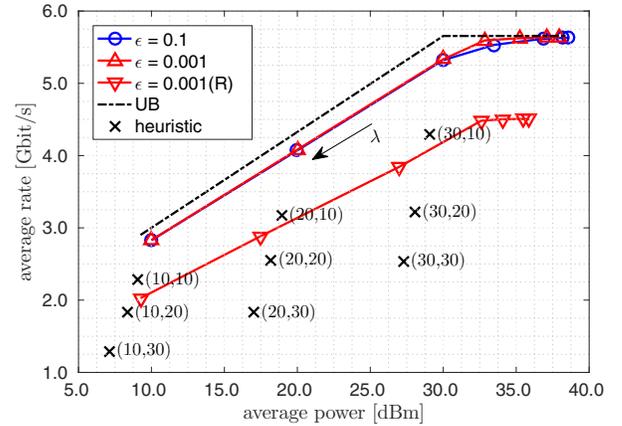

Fig. 2: Rate and power as a function of λ and comparison with the heuristic π_{H} for different pairs $(P_{\text{DT}}, T_{\text{DT}})$ (black crosses) and PBVI with points in $\tilde{\mathcal{B}}$ obtained by random sampling (R).

The average rate and power as a function of $\lambda \in [10^5, 10^{11}]$ are plotted in Fig. 2, with $\epsilon = \{0.1, 0.001\}$. As λ increases, both the average rate and power decrease, and the optimal policies are more conservative: given the current belief, the BS performs more directive DT over a smaller set $\hat{\mathcal{S}}$, using smaller values of P_{DT} and T_{DT}. Conversely, as λ decreases, both the average rate and power increse, and the optimal policies are more energy-demanding. Also, as λ decreases, the impact of ϵ on the performance of the optimal policies is slightly more evident: given the average power, we can achieve a larger average rate with a smaller ϵ, i.e., a larger $P_{\min}$, meaning that the optimal policies are more sensitive to the observation errors when performing BT than to the actual value of $P_{\min}$. Overall, the plot suggests that the performance of the optimal policies is quite robust to the observation errors when performing BT, whereas this is not true for any heuristic policy. The heuristic π_{H} works as follows: the BS

---

[1]Note that a gradient descent technique adjusts the parameter λ after each iteration in the direction that would reduce the error on that iteration the most. The target here depends on the parameter λ, but if we ignore that dependence when we take the derivative, then what we get is a semi-gradient update [20].

performs BT with (P_{BT}, T_{BT}) over $\hat{\mathcal{S}} = \{s\}$ (starting from $s = 1$). If the MU replies with ACK, then the BS performs DT with (P_{DT}, T_{DT}) over $\hat{\mathcal{S}}$, until the first NACK. At this point, the BS scans the two states $s - 1$ and $s + 1$ in two different micro time-slots. If the MU replies with ACK, then the BS repeats the scheme (starting from the new state). The heuristic π_H for different pairs (P_{DT}, T_{DT}) (black crosses) cannot achieve the performance of the optimal policies, which take advantage of the belief update mechanism and provide adaptive BT/DT procedures according to the current belief. The average rate and power for π_H are plotted in Fig. 2 for $\epsilon = 0.001$. The achievable rate drops significantly when considering $\epsilon = 0.1$ (not shown in Fig. 2), since π_H does not provide any countermeasure to the observation errors when performing BT. An upper bound on the average rate is given when performing DT over the system state (i.e., assuming that the BS knows the position of the MU), thus achieving the maximum transmission rate without wasting power. When the probability distributions collected in $\tilde{\mathcal{B}}$ are not very representative of the system dynamics history, then the performance of the optimal policies can greatly degrade, see Fig. 2, where the probability distributions are obtained by sampling random actions and observations, as suggested in [11], and compared to the ones obtained by Algorithm 2 with $W = S$. Here, the total number of beliefs collected in $\tilde{\mathcal{B}}$ is the same in the two cases. However, the performance of the optimal policies with beliefs obtained by Algorithm 2 can provide an improvement in the achievable rate of ~ 1 Gbit/s (1.15 Gbit/s gap using an average power of 35 dBm).

An example of the outcome of Algorithm 4 is given in Fig. 3, starting from parameters $\lambda = 0$ and $\alpha = 100$. Here, we set the constraint C (which corresponds to $V_c^*(b_0)$ for $\lambda = 10^5$ in Algorithm 3), and we achieve convergence of the rate to $\bar{R}$, with $\lambda_{opt} \simeq 10^5$. The average rate and power as a function of n are plotted in Fig. 3, with $\bar{R}/D$ and C/D.

V. CONCLUSIONS

In this paper, learning tools are investigated to optimize the trade-off between directive Data Transmission (DT) and directional Beam Training (BT) in a mm-wave vehicular network. In the proposed scenario, a Base Station (BS) has to face the issue of beam alignment/realignment towards a Mobile User (MU) moving along a road and, furthermore, has to decide the transmission parameters to use in the DT phase. This leads to a complex problem involving the estimation of the position of the MU within the road link, which must be leveraged in the design of BT/DT protocols. Numerical results reveal that adaptive BT/DT procedures do provide a major advantage over common-sense heuristic schemes, being able to maximize the transmission rate while showing robustness against BT errors, under a power constraint.

Fig. 3: Rate and power as a function of λ, which is properly tuned according to $\lambda_{n+1} = \max(0, \lambda_n + \alpha_n(V_{n+1}^c(b_0) - C))$.

REFERENCES

[1] 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.

[2] M. Giordani, A. Zanella, and M. Zorzi, "Millimeter wave communication in vehicular networks: Challenges and opportunities," in *Modern Circuits and Systems Technologies (MOCAST), 2017 6th International Conference on*, pp. 1–6, IEEE, 2017.
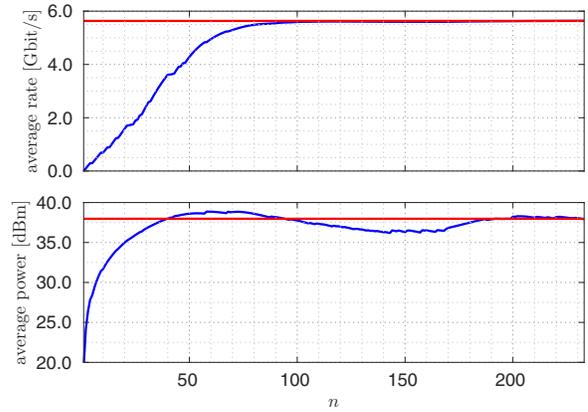
[3] V. Va, J. Choi, T. Shimizu, G. Bansal, and R. W. Heath, "Inverse multipath fingerprinting for millimeter wave v2i beam alignment," *IEEE Transactions on Vehicular Technology*, 2017.

[4] J. Palacios, D. De Donno, and J. Widmer, "Tracking mm-wave channel dynamics: Fast beam training strategies under mobility," in *INFOCOM 2017-IEEE Conference on Computer Communications, IEEE*, pp. 1–9, IEEE, 2017.

[5] M. Giordani, M. Mezzavilla, and M. Zorzi, "Initial access in 5g mmwave cellular networks," *IEEE Communications Magazine*, vol. 54, pp. 40–47, November 2016.

[6] V. Va, X. Zhang, and R. W. Heath, "Beam switching for millimeter wave communication to support high speed trains," in *Vehicular Technology Conference (VTC Fall), 2015 IEEE 82nd*, pp. 1–5, IEEE, 2015.

[7] V. Va, T. Shimizu, G. Bansal, and R. W. Heath, "Beam design for beam switching based millimeter wave vehicle-to-infrastructure communications," in *Communications (ICC), 2016 IEEE International Conference on*, pp. 1–6, IEEE, 2016.

[8] N. Michelusi and M. Hussain, "Optimal beam-sweeping and communication in mobile millimeter-wave networks," in *2018 IEEE International Conference on Communications (ICC)*, pp. 1–6, May 2018.

[9] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial Intelligence*, vol. 101, no. 1, pp. 99 – 134, 1998.

[10] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of artificial intelligence research*, vol. 4, pp. 237–285, 1996.

[11] M. T. J. Spaan and N. A. Vlassis, "Perseus: Randomized point-based value iteration for pomdps," *CoRR*, vol. abs/1109.2145, 2011.

[12] E. Altman, *Constrained Markov decision processes*, vol. 7. CRC Press, 1999.

[13] E. Altman and F. Spieksma, "The linear program approach in multi-chain markov decision processes revisited," *Zeitschrift für Operations Research*, vol. 42, no. 2, pp. 169–188, 1995.

[14] J. D. Isom, S. P. Meyn, and R. D. Braatz, "Piecewise linear dynamic programming for constrained pomdps.," in *AAAI*, pp. 291–296, 2008.

[15] D. Kim, J. Lee, K.-E. Kim, and P. Poupart, "Point-based value iteration for constrained pomdps," in *IJCAI*, pp. 1968–1974, 2011.

[16] P. Poupart, A. Malhotra, P. Pei, K.-E. Kim, B. Goh, and M. Bowling, "Approximate linear programming for constrained partially observable markov decision processes.," in *AAAI*, pp. 3342–3348, 2015.

[17] A. Undurti and J. P. How, "An online algorithm for constrained pomdps," in *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pp. 3966–3973, IEEE, 2010.

[18] S. M. Kay, *Fundamentals of statistical signal processing: Practical algorithm development*, vol. 3. Pearson Education, 2013.

[19] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[20] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*, vol. 1. MIT press Cambridge, 1998.