BEAM ALIGNMENT FOR MILLIMETER WAVE WIRELESS COMMUNICATIONS: A MULTISCALE APPROACH

by

Muddassar Hussain

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

School of Electrical and Computer Engineering

West Lafayette, Indiana

May 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Nicolò Michelusi, Chair

School of Electrical and Computer Engineering

Dr. David Love

School of Electrical and Computer Engineering

Dr. James Krogmeir

School of Electrical and Computer Engineering

Dr. Xiaojun Lin

School of Electrical and Computer Engineering

Approved by:

Dr. Dimitrios Peroulis

I dedicate this thesis to my parents and my adviser Dr. Nicolò Michelusi

ACKNOWLEDGMENTS

First, I would like to pay gratitude to my adviser Professor Nicolò Michelusi, a source of immense guidance, support, and inspiration. I greatly appreciate his mentorship, which helped me improve my research and presentation skills.

Second, I would like to thank my doctoral advisory committee, Professor David Love, Professor James Krogmeir, and Professor Xiaojun Lin, for their invaluable suggestions. I would also express my gratitude to Purdue University (Bilsland Dissertation Fellowship) and National Science Foundation (CNS-1642982), for funding my research. I am also grateful to my undergraduate adviser Professor Syed Ali Hassan, who guided and helped me pursue a research-oriented career.

Lastly, I would thank my parents for their constant encouragement, unwavering support, and prayers. I am grateful to my siblings, Faisal, Memona, and Fezza, for their support and prayers. My utmost gratitude goes to God Almighty, who made the arduous journey of my Ph.D. easier for me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Millimeter-wave communications use narrow beams to overcome the enormous signal attenuation. Such narrow-beam communication demands precise beam-alignment between transmitter and receiver and may entail huge overhead, especially in high mobility scenarios. Moreover, detection of the optimal beam is challenging in the presence of beam imperfections and system noise. This thesis addresses the challenges in the design of beam-training and data-communication by proposing various schemes that exploit different timescales. On a short timescale, we leverage the feedback from the receiver to efficiently perform beam-training and data-communication. To this end, we have worked in three different areas. In the first research direction, we design an optimal interactive beam-training and data-communication protocol, with the goal of minimizing power consumption under a minimum rate constraint. The optimality of a fixed-length beam-training phase followed by a data-communication phase is proved under the assumption of perfect binary feedback. In the second research direction, we propose a coded energy-efficient beam-training scheme, robust against the feedback/detection errors. In the third research direction, we investigate the design of the beam-training in the presence of uncertainty due to noise and beam imperfections. Based on the bounding of value-function, the second-best preference policy is proposed, which achieves a promising exploration-exploitation tradeoff. On the other hand, on longer timescales, we exploit the mobility and blockage dynamics and beam-training feedback to design throughput-efficient beam-training and data-communication. We propose a *point-based value iteration* (PBVI) algorithm to determine an approximately optimal policy. However, the design relies on the a-priori knowledge of the state dynamics, which may not be available in practice. To address this, we propose a dual timescale approach, where on the long timescale, a *recurrent deep variational autoencoder* (R-VAE) uses noisy beam-training observations to learn a probabilistic model of system dynamics; on the short timescale, an adaptive beam-training procedure is optimized using PBVI based on beam-training feedback and a probabilistic knowledge of the UE's position provided by the R-VAE. In turn, the observations collected during the beam-training procedure are used to refine the R-VAE via stochastic gradient descent in a continuous process of learning and adaptation.

1. INTRODUCTION

Mobile traffic has witnessed tremendous growth over the last decade, 18-folds over the past five years alone, and is expected to grow with a compound annual growth rate of 47% from 2016 to 2021 [1]. This rapid increase poses a severe burden to current systems operating below 6 GHz, due to limited bandwidth availability. Millimeter-wave (mm-wave) is emerging as a promising solution to enable multi-Gbps communication, thanks to abundant bandwidth availability [2]. However, high isotropic path loss and sensitivity to blockages pose challenges in supporting high capacity and mobility [3]. To overcome the path loss, mm-wave systems will thus leverage narrow beams by using large antenna arrays at both base stations (BSs) and user-ends (UEs).

Nonetheless, narrow transmission and reception beams are susceptible to frequent loss of alignment, due to mobility or blockage, which necessitate the use of beam-alignment protocols. Maintaining beam alignment between transmitter and receiver can be challenging, especially in mobile scenarios, and may entail significant overhead, thus potentially offsetting the benefits of mm-wave directionality. Therefore, it is imperative to design schemes to mitigate its overhead. This thesis addresses the challenges in the design of beam-alignment by proposing various schemes that exploit different timescales.

On short timescales, we leverage the feedback from the receiver to efficiently perform beam alignment and data communication. To this end, we have worked in three different areas: energy-efficient interactive beam-alignment [4] (Chapter 2), coded energy-efficient beam-alignment [5] (Chapter 3), second-best beam-alignment via Bayesian Multi-armed bandits [6] (Chapter 4).

In Chapter 2, we investigate the design of an optimal interactive beam-alignment and data communication protocol, with the goal of minimizing power consumption under a minimum rate constraint. The base station (BS) selects beam alignment or data communication and the beam parameters, based on feedback from the user equipment (UE). Based on the sectored antenna model and uniform prior on the angles of departure and arrival (AoD/AoA), the optimality of a *fixed-length* beam-training phase followed by a data-communication phase is demonstrated. Moreover, a *decoupled fractional* beam-alignment method is shown to be

optimal, which decouples over time the alignment of AoD and AoA, and iteratively scans a fraction of their region of uncertainty. A heuristic policy is proposed for non-uniform prior on AoD/AoA, with provable performance guarantees, and it is shown that the uniform prior is the worst-case scenario. The performance degradation due to detection errors is studied analytically and via simulation.

In Chapter 3, we investigate the design of a coded energy-efficient beam-alignment scheme, robust against detection errors. Specifically, the beam-alignment sequence is designed such that the error-free feedback sequences are generated from a codebook with the desired error correction capabilities. Therefore, in the presence of detection errors, the error-free feedback sequences can be recovered with high probability. The assignment of beams to codewords is designed to optimize energy efficiency, and a water-filling solution is proved.

In Chapter 4, a beam-alignment scheme is proposed based on Bayesian multi-armed bandits, with the goal to maximize the alignment probability and the data-communication throughput. A Bayesian approach is proposed, by considering the state as a posterior distribution over AoA and AoD given the history of feedback signaling and of beam pairs scanned by the base-station (BS) and the user-end (UE). A simplified sufficient statistic for optimal control is identified, in the form of preference of BS-UE beam pairs. By bounding a value function, the *second-best preference* policy is formulated, which strikes an optimal balance between exploration and exploitation by selecting the beam pair with the current *second-best* preference. Through Monte-Carlo simulation with analog beamforming, the superior performance of the *second-best preference* policy is demonstrated in comparison to existing schemes based on *first-best preference*, linear Thompson sampling, and upper confidence bounds.

On longer timescales, we exploit the mobility and blockage dynamics to design throughput efficient beam-alignment design. In Chapter 5 (previously published in [7]), we investigate the design of joint beam-alignment, data communication and handover. In the proposed scenario, two base stations use beam training to establish a mm-wave directive link towards a user end moving along a road. At each time, the serving BS decides to either perform beam training, data communication, or handover. Our goal is to maximize the average number of successfully transmitted bits subject to an average power constraint. The beam training, data communication, and handover strategies are jointly optimized by casting the

optimization problem as a partially observable Markov decision process, where the system state corresponds to the index of beam sectors where UE is located, the blockage variables, and the index of the serving BS. To address the high dimensionality of the problem, an approximate dynamic programming algorithm based on PERSEUS [8] is developed, where we optimize both the primal and dual function simultaneously. The numerical results show that the optimal policy based on the above optimization provides performance very close to the upper-bound. Motivated by the structure of the optimal policy, we propose two simple heuristic policies, namely finite-state-machine-based heuristic (FSM-HEU) and belief-based-heuristic (B-HEU) policies, which compared to the optimal design, incur lower computation cost and shows comparable performance. We compare the proposed policies to a baseline policy referred to as the conventional heuristic (C-HEU) policy.

In Chapter 6 of this dissertation, we propose a dual timescale beam-training and data-transmission approach: on a large timescale, a recurrent deep variational autoencoder (R-VAE) uses noisy beam-training observations to learn a probabilistic model of user mobility dynamics; on a short timescale, an adaptive beam-training procedure is optimized using *point-based value iteration* (PBVI) based on beam-training feedback and a probabilistic knowledge of the UE's position provided by the R-VAE. In turn, the observations collected during the beam-training procedure are used to refine the R-VAE via stochastic gradient descent in a continuous process of learning and adaptation.

The proposed beam-alignment schemes are outlined as follows.

1.1 Energy-Efficient Interactive Beam Alignment for Millimeter-Wave Networks

Millimeter-wave communications use narrow beams to overcome the huge path loss. This demands precise beam-alignment between transmitter and receiver and may entail huge overhead, especially in mobile environments. To address this challenge, in our previous work [9]–[12], we address the optimal design of beam-alignment protocols. In [9], we optimize the trade-off between data communication and beam-sweeping, under the assumption of an exhaustive search method, in a mobile scenario where the BS widens its beam to mitigate the uncertainty on the UE position. In [10], [11], we design a throughput-optimal beam-

alignment scheme for one and two UEs, respectively, and we prove the optimality of a *bisection search*. However, the model therein does not consider the energy cost of beam-training, which may be significant when targeting high detection accuracy. It is noteworthy that, if the energy consumption of beam-training is small, bisection search is the best policy since it is the fastest way to reduce the uncertainty region of the angles of arrival (AoA) and departure (AoD). For this reason, it has been employed in previous works related to multi-resolution codebook design, such as [13]. In [12], [14], we incorporate the energy cost of beam-training, and prove the optimality of a *fractional search* method. Yet, in [9]–[12], optimal design is carried out under restrictive assumptions that the UE receives isotropically, and that the duration of beam-training is fixed. In practice, the BS may switch to data transmission upon finding a strong beam, as in [15], and *both* BS and UE may use narrow beams to fully leverage the beamforming gain.

To the best of our knowledge, the optimization of *interactive* beam-alignment, *jointly* at both BS and UE, is still an open problem. Therefore, in Chapter 2 (previously published in [4]), we consider a more flexible model than our previous papers [9]–[12], by allowing dynamic switching between beam-training and data-communication and joint optimization over BS-UE beams, BS transmission power, and rate. *Indeed, we prove that a fixed-length beam-training scheme followed by data communication is optimal, and we prove the optimality of a decoupled fractional search method, which decouples over time the alignment of AoD and AoA and iteratively scans a fraction of their region of uncertainty.* Using Monte-Carlo simulation with analog beams, we demonstrate superior performance, with up to 4dB, 7.5dB, and 14dB power gains over the state-of-the-art bisection method [13], conventional exhaustive, and interactive exhaustive search policies, respectively. Compared to our recent paper [14], the system model adopted in Chapter 2 is more realistic since it captures the effects of fading and resulting outages, non-uniform priors on AoD/AoA, and detection errors. Additionally, the model in [14] is restricted to a two-phase protocol with deterministic beam-training duration. In Chapter 2, we show that this is indeed optimal. Beam-alignment has been a subject of intense research due to its importance in mm-wave communications. The research in this area can be categorized into beam-sweeping [9]–[13], [16]–[18], data-assisted schemes [19]–[22], and AoD/AoA estimation [23], [24]. The simplest and yet most popular beam-

sweeping scheme is *exhaustive* search [16], which sequentially scans through all possible BS-UE beam pairs and selects the one with maximum signal power. A version of this scheme has been adopted in existing mm-wave standards including IEEE 802.15.3c [25] and IEEE 802.11ad [26]. An interactive version of exhaustive search has been proposed in [15], wherein the beam-training phase is terminated once the power of the received beacon is above a certain threshold. The second popular scheme is *iterative search* [17], where scanning is first performed using wider beams followed by refinement using narrow beams. A variant of *iterative search* is studied in [27], where the beam sequence is chosen adaptively from a pre-designed multi-resolution codebook. However, this codebook is designed independently of the beam-alignment protocol, thereby potentially resulting in suboptimal design. In [18], the authors consider the design of a beamforming vector sequence based on a partially observable (PO-) Markov decision processes (MDPs). However, POMDPs are generally not amenable to closed-form solutions, and have high complexity. To reduce the computational overhead, the authors focus on a greedy algorithm, which yields a sub-optimal policy.

Data-aided schemes utilize the information from sensors to aid beam-alignment and reduce the beam-sweeping cost (e.g., from radar [19], lower frequencies [20], position information [21], [22]). AoD/AoA estimation schemes leverage the sparsity of mm-wave channels and include compressive sensing schemes [23] or approximate maximum likelihood estimators [24]. In [28], the authors compare different schemes and conclude that the performance of beam-sweeping is comparable with the best performing estimation schemes based on compressed sensing. Yet, beam-sweeping has the added advantage of low complexity over compressed sensing schemes, which often involve solving complex optimization problems and is more amenable to analytical insights on the beam-alignment process. For these reasons, in Chapter 2, we focus on beam-sweeping and derive insights on its optimal design.

All of the aforementioned schemes choose the beam-training beams from pre-designed codebooks, use heuristic protocols, or are not amenable to analytical insights. By choosing the beams from a restricted beam-space or a predetermined protocol, optimality may not be achieved. Moreover, all of these papers do not consider the energy and/or time overhead of beam-training as part of their design. In this work, we address these open challenges by

optimizing the beam-alignment protocol to maximize the communication performance.

Our Contributions: Our contributions are summarized as follows:

1. Based on a MDP formulation, under the *sectored* antenna model [29], uniform AoD/AoA prior, and small detection error assumptions, we prove the optimality of a *fixed-length* two-phase protocol, with a beam-training phase of fixed duration followed by a data communication phase. We provide an algorithm to compute the optimal duration.

2. We prove the optimality of a *decoupled fractional search* method, which scans a fixed fraction of the region of uncertainty of the AoD/AoA in each beam-training slot. Moreover, the beam refinements over the AoD and AoA dimensions are decoupled over time, thus proving the sub-optimality of *exhaustive search* methods.

3. Inspired by the decoupled fractional search method, we propose a heuristic scheme for the case of non-uniform prior on AoD/AoA with provable performance and prove that the uniform prior is indeed the worst-case scenario.

4. We analyze the effect of detection errors on the performance of the proposed protocol.

5. We evaluate its performance via simulation using analog beams, and demonstrate up to 4dB, 7.5dB, and 14dB power gains compared to the state-of-the-art bisection scheme [13], conventional and interactive exhaustive search policies, respectively. Remarkably, the sectored model provides valuable insights for beam-alignment design.

1.2 Coded Energy-Efficient Beam-Alignment

Existing beam alignment techniques such as [9], [10], [12], [14]–[16], [30], are designed based on the assumption that no detection errors occur in the beam-training. However, the performance may deteriorate due to mis-detection and false-alarm errors, causing a loss of alignment during the communication phase. Therefore, it is of great interest to design beam-alignment algorithms robust to detection errors and, at the same time, energy-efficient.

Motivated by these observations, in Chapter 3 (previously published in [5]), we consider the design of an energy-efficient beam-alignment protocol robust to detection errors. To do

so, we restrict the solution space for the beams such that the error-free feedback sequence can only be generated from a codebook with error correction capabilities. Thus, if detection errors occur, the error-free feedback sequence may still be recovered with high probability by leveraging the structure of the error correction code. We pose the beam-alignment problem as a convex optimization problem to minimize the average power consumption and provide its closed-form solution that resembles a "water-filling" over the beamwidths of the beam-training beam patterns. The numerical results depict the superior performance of the proposed coded technique, with up to 4dB and 8dB gains over exhaustive and uncoded beam-alignment schemes, respectively. Open- and closed-loop error control sounding schemes have been studied in [31], but with no consideration on energy-efficient design. *To the best of our knowledge, [5] (Chapter 3 is the first to propose a coded beam-alignment scheme, which is both energy-efficient and robust to detection errors.*)

In [32], beam-alignment is treated as a beam discovery problem in which locating beams with strong path reflectors is analogous to locating errors in a linear block code. Unlike [32], we use error correction to correct errors during the beam-training procedure, rather than to detect strong signal clusters. Unlike [10], [12], [14] which rely on continuous feedback from UEs to BS, we consider a scheme where the feedback is generated only at the end of the beam-training phase, which scales well to multiuser scenarios.

1.3 Second-best Beam-Alignment via Bayesian Multi-Armed Bandits

Noise and beam imperfections can cause the beam training feedback errors. In the presence of the feedback errors, the detection of the optimal beam becomes challenging. The case of erroneous or noisy feedback is considered in recent work [33], [34], and our work [5]. A coded beam-alignment scheme is proposed in [5] to correct these errors, but with no consideration of feedback to improve beam-selection. A multi-armed bandit (MAB) formulation based on upper confidence bound (UCB) is proposed in [33], by selecting the beam based on the empirical SNR distribution. A hierarchical beam-alignment scheme based on posterior matching is proposed in [34]: therein, a *first-best* policy is formulated, which selects the most likely beam pair based on the posterior distribution on the AoA-AoD pair.

However, as we will see numerically, both UCB and first-best policies are prone to errors due to under-exploration of the beam space.

In Chapter 4 (previously published in [6]), we propose a beam-alignment design with the goal to maximize the alignment probability and the average throughput during the data communication phase. We pose the problem as a MDP, where the beam pair is chosen based upon the *belief* over the AoA-AoD pair, given the history of scanned beams and the received signal power. We identify a simplified sufficient statistic in the form of preference of the AoA-AoD beam pairs. We derive lower and upper bounds to the value function, based on which we propose a heuristic policy that selects the beam pair with the *second-best* preference. We show numerically that this policy strikes a favorable trade-off between exploration and exploitation: instead of greedily choosing the beam corresponding to the most likely AoA-AoD pair (*first-best* [34]), it chooses the second most likely one, leading to better exploration; at the same time, it avoids wasting precious resources to scan unlikely beam pairs, leading to better exploitation than other MAB techniques, such as linear Thompson sampling (LTS) [35] and UCB [33]. The proposed *second-best* scheme is shown to outperform *first-best* [34], LTS-based [35] and UCB-based [33] schemes by up to 7%, 10% and 30% in alignment probability, respectively.

1.4 Mobility and Blockage aware Communications in Millimeter-Wave Vehicular Networks

Mobility can thus be a source of severe overhead and performance degradation. Nevertheless, mobility induces temporal correlation in the communication beams and in blockage events. In Chapter 5 (previously published in [7]), we design adaptive strategies for beam-training, data transmission and handover, that exploit these temporal correlations to reduce the beam-training overhead and optimally trade-off throughput and power consumption. Our design allows to: 1) predict future beam-pointing directions and narrow down the beam search procedure to few likely beams, thus avoiding the enormous cost of exhaustive search; 2) more efficiently detect blockage and perform handover in response to it; 3) dynamically adjust the duration of the data communication phase based on predicted beam coherence times. However, two key questions arise: *How do we leverage the system dynamics to opti-*

mize the communication performance? How much do we gain by doing so? To address these questions, in Chapter 5 we envision the use of adaptive communication strategies and their formulation via partially observable (PO) Markov decision processes (MDPs) to optimize the decision-making process under uncertainty in the state of the system [36].

In the proposed scenario, two base stations (BSs) on both sides of a road link serve a user equipment (UE) moving along it. At any time, the UE is associated with one of the two BSs (the serving BS). To enable directional data transmission (DT), the serving BS performs beam-training (BT); to compensate for blockage, it performs handover (HO) of the data traffic to the backup BS on the opposite side of the road link. The goal is to design the BT/DT/HO strategy so as to maximize the throughput delivered to the UE, subject to an average power constraint. Mobility induces dynamics in the communication beams and in blockage events; we show that these dynamics can be captured by a *probabilistic state transition model*, which can be learned from interactions with the UE. However, the system state is not directly observable due to noise, beam imperfections, and detection errors; we thus formulate the optimization of the decision-making process as a constrained POMDP, and develop an approximate *constrained point-based value iteration* (C-PBVI) method to meet the average power constraint requirement: compared with PERSEUS [37], originally proposed for unconstrained problems, C-PBVI allows to simultaneously optimize the primal and dual functions by decoupling the hyperplanes associated to reward and cost. We demonstrate its convergence numerically. Our numerical evaluations reveal a good match between the analysis based on a *sectored antenna model* with Markovian state transitions, and a more realistic scenario with analog beamforming and Gauss-Markov mobility, hence demonstrating the effectiveness of our proposed scenario in more realistic settings: simulations based on a 2D mobility model and 3D analog beamforming on both BSs and UE equipped with uniform planar arrays (UPA), demonstrate that C-PBVI performs near optimally, and outperforms a baseline scheme with periodic beam-training by up to 38% in spectral efficiency. Motivated by its structure, we design two heuristic policies with lower computational cost – belief-based and finite-state-machine-based heuristics – and show numerically that they incur a small 4% and 15% degradation in spectral efficiency compared to C-PBVI, respectively. Finally, we demonstrate numerically the effect of mobility and multiple users on the performance, based

on the statistical blockage model developed in [38]: the proposed low-complexity belief-based and finite-state-machine-based schemes achieve 50% and 25% higher spectral efficiency than the baseline scheme, respectively, demonstrating their robustness in mobile and dense user scenarios.

Related Work: Beam-training design for mm-wave systems has been an area of extensive research in the past decade; various approaches have been proposed, such as beam sweeping [39], estimation of angles of arrival (AoA) and of departure (AoD) [24], and data-assisted schemes [22]. Despite their simplicity, the overhead of these algorithms may offset the benefits of beamforming in highly mobile environments [40]. While wider beams require less beam-training, they result in a lower beamforming gain, hence a smaller achievable capacity. Contextual information, such as GPS readings of vehicles [22], may alleviate this overhead, but it does not eliminate the need for beam-training due to noise and inaccuracies in GPS acquisition. Thus, the design of schemes that alleviate the beam-training overhead is of great importance.

In most of the aforementioned works, a priori information on the vehicle's mobility as well as blockage dynamics is not leveraged in the design of communication protocols. In contrast, *we contend and demonstrate numerically that learning and exploiting such information via adaptive communications can greatly improve the performance of mm-wave networks* [41]. In our previous work [39], we bridged this gap by leveraging worst-case mobility information to design beam-sweeping and data communication schemes; in [42], we designed adaptive strategies for BT/DT that leverage a Markovian mobility model via POMDPs, but with no consideration of blockage (hence no handover).

A distinctive feature of the mm-wave channel is its highly dynamic link quality, due to the occurrence of blockages on very short time-scales [43]. In this respect, handover represents a fundamental functionality to preserve communication in the event of link obstruction; however, it is challenging to implement it in mm-wave networks, since the mm-wave link quality needs to be accurately tracked and blockages need to be quickly detected – a difficult task to accomplish using highly directional communications. Therefore, MDP-based handoff strategies proposed for sub-5GHz systems cannot be readily applied [44], [45]. In Chapter 5, we

develop feedback-based techniques to quickly detect blockages, and enable a fully-automatic and data-driven optimization of the handover strategy via POMDPs.

Recent work [4], [6], [34], [46], [47] that applies machine learning to mm-wave networks reveal a growing interest in the design of schemes that exploit side information to enhance the overall network performance. For example, [46] develops a coordinated beamforming technique using a combination of deep learning and ray-tracing, and demonstrates its ability to efficiently adapt to changing environments. More recent solutions are based on multi-armed bandit, by leveraging *contextual information* to reduce the training overhead as in [47], or the beam alignment feedback to improve the beam search as in [4], [6], [34]. However, no handover strategies are considered in these works, resulting in limited ability to combat blockage. In addition, these works neglect the impact of realistic mobility and blockage processes on the performance. Compared to this line of works, in Chapter 5 we design adaptive communication strategies that leverage learned statistical information on the mobility and blockage processes in the selection of BT/DT/HO actions, with the goal to optimize the average long-term communication performance of the system. Our proposed approach is in contrast to strategies that either use non-adaptive algorithms [46], lack a handover mechanism [4], [6], [34], [47], or assume a non realistic mobility pattern in their design.

Our Contributions:

- We define a POMDP framework to optimize the BT/DT/HO strategy in a mm-wave vehicular network, subject to 2D mobility of the UE and time-varying blockage, with the goal to maximize throughput subject to an average power constraint;
- We propose a novel feedback mechanism for BT, which reports the ID of the strongest BS-UE beam pair if the received power is above a threshold (a design parameter), otherwise it reports $\emptyset$ to indicate mis-alignment or blockage. We analyze its detection performance in closed form;
- To address the complexity of POMDPs, we design C-PBVI, a constrained point-based value iteration method. In order to incorporate the average power constraint, we extend PERSEUS [37], originally designed for unconstrained POMDPs, via a Lagrangian formulation, the separation of hyperplanes for reward-to-go and cost-to-go functions, and a

dual optimization step to solve the constrained problem. We demonstrate its convergence numerically;

- Inspired by the C-PBVI policy, we propose two heuristic schemes that trade complexity with sub-optimality, namely belief-based (B-HEU) and finite-state-machine-based (FSM-HEU) heuristic policies. We analyze the performance of FSM-HEU in closed form.

1.5 Learning and Adaptation in Millimeter-Wave Communications via Deep Variational Autoencoders and POMDPs

Maintaining beam alignment, especially in highly mobile V2X communication scenarios, is extremely challenging: traditional beam-alignment schemes such as the exhaustive search method [48] suffer from severe beam-training overhead, increased communication delay, and degraded throughput performance.

To achieve efficient design, adaptive beam-training schemes have been proposed in the literature [4], [7], [34], [41], [42], that leverage information on the mobility of the UE and beam-training feedback to minimize the beam-training overhead. In one of our recent works [7], we showed that statistical knowledge of the UE's mobility dynamics may be carefully exploited to reduce the beam-training overhead and achieve high spectral efficiency, even in highly mobile V2X scenarios. However, the design in [7] relies on a priori statistical knowledge of the mobility dynamics, which may not be available in practice, hence need to be estimated from noisy observations. Then, a key question arises: *How to jointly estimate mobility dynamics from noisy observations and leverage them to optimize the beam-training and data communication decisions?*

To address this challenge, in Chapter 6, we consider a mm-wave vehicular communication scenario, where a UE moves along a road according to an *unknown* mobility model and is served by a roadside BS. The UE and the BS are both equipped with large antenna arrays and use 3D beamforming to enable directional communication. The mobility of the UE and of the surrounding environment induce dynamics in the *strongest beam pair* that maximizes the beamforming gain; these unknown dynamics need to be learned to enable efficient beam-training. To this end, we propose a learning and adaptation framework that exploits two

timescales: on the long timescale (of the order of several hundreds of frames), the BS uses noisy signal quality measurements to learn the *strongest beam pair dynamics* induced by the UE's and environment's mobility, using a recurrent deep variational auto-encoder (R-VAE) [49]; the learned model is then used on the short timescale (one frame duration) to design adaptive beam-training schemes that leverage the probabilistic knowledge of the strongest beam pair provided by the R-VAE and beam-training feedback. In turn, beam-training observations are used to refine the R-VAE via stochastic gradient descent in a continuous process of learning and adaptation.

By approximating the beamforming gain via the sectored antenna model, we formulate the decision-making process over the short-timescale as a POMDP and propose a PBVI method to design an approximately optimal policy, which provides the rule to select the actions based on the belief (probability distribution over the optimal BS-UE beam pair given the history of actions taken so far and associated observations) and beam-training feedback. We compare the estimation performance of R-VAE with that based on the Baum-Welch algorithm [50], and a naive approach, which ignores the noise in the observations. Through numerical evaluations using 3D analog beamforming, we show that the R-VAE reduces the average Kullback-Leibler (KL) divergence between the ground-truth Markovian and the estimated mobility model by 92% and 86% with respect to the naive approach and the Baum-Welch algorithms, respectively. Moreover, when used in conjunction with the PBVI-based adaptive beam-training policy, the proposed dual timescale approach yields near-optimal spectral efficiency, comparable to a genie-aided scheme with knowledge of the ground-truth mobility model and noiseless feedback, and improves the spectral efficiency by 12.6% and 8% with respect to the naive approach and the Baum-Welch algorithms, respectively.

Finally, to trade computational complexity with accuracy, we propose a policy for the short timescale by reducing the POMDP to an MDP that operates under the assumption of error-free beam-training feedback. For example, the total time taken to optimize the policy and execute 1000 episodes is 4.7 times lower for the proposed MDP-based policy compared to the PBVI-based policy, while achieving spectral efficiency close to the latter in low feedback error regimes. These policies are compared to a policy that scans exhaustively over the

dominant beam pairs, demonstrating a spectral efficiency gain of 46% (PBVI-based) and 37% (MDP-based). Through Monte-Carlo simulation, we show a perfect match between the actual system using analog 3D beamforming and 2D UE's mobility and the abstracted analytical model based on a Markovian approximation of the mobility and a sectored model of the beamforming gain.

Related Work: The beam-alignment problem has been a topic of intensive research in the last decade, and can be categorized into beam sweeping [39], estimation of AoA and AoD [24], and contextual-information-aided schemes [19], [22], [51]. Despite their simplicity, these schemes do not incorporate mobility dynamics as part of their design, leading to a large beam-training overhead in high mobility [40]. Some recent papers use contextual information, such as GPS coordinates of the UE, [22], onboard sensors' data [19], sub-6HGz channel estimates [51] to reduce the beam-training overhead. In [22], [52], the BS uses a data-base of past measurements and the associated UE locations to predict the dominant beamforming directions via inverse fingerprinting. Similarly, [19] proposes a beam-alignment scheme using the onboard radar's signals. Additionally, in [51], the proposed method uses the sub-6GHz channel measurements to predict the mm-wave channel. In [52], a noisy tensor completion-based beam-training scheme is proposed, where the received power is predicted across a BS coverage area by using beam-training measurements on a subset of positions and beams. Although contextual information may reduce some beam-training overhead, beam-training is still required [22], [52], due to noise and inaccuracies in contextual information acquisition. Moreover, a UE may decide not to share contextual information due to privacy concerns. Therefore, contextual-information-agnostic efficient beam-training schemes are required for these scenarios.

Adaptive beam-training solutions, including machine learning-based, have been proposed in some recent works [4], [6], [34], [46], [47], [53], [54]. These works exploit side-information and/or beam-training feedback to reduce the beam-training overhead. Deep learning-based solutions have been proposed in [46], [53], [54]. For instance, [46] uses the received sounding signal from multiple surrounding BSs to predict the optimal beam via a deep learning framework. In [53], a convolutional neural network-based compressive sensing solution is proposed, trained based on simulated channels and then used to make beam predictions using only a

few measurements. Similarly, [54] proposes a deep-learning assisted beam-alignment, which predicts the optimal BS and beam, given the UE's position. Reinforcement learning-based beam-alignment schemes have been proposed in [4], [6], [34], [47]. In [47], the beam-alignment problem is posed as a contextual bandit problem, using the UE's location information as the context. In our previous works [4] and [6], we proposed to use the beam-training feedback to design optimal beam-training strategies under the assumption of error-free and erroneous beam-training feedback, respectively. In [34], a hierarchical search, exploiting the beam-training feedback is proposed. In the aforementioned works, the mobility dynamics of the UE are not leveraged in the beam-alignment protocol design, leading to a large beam-training overhead in high mobility [40].

Compared to the aforementioned works, in our recent work [7] we showed that by exploiting the mobility dynamics via a POMDP, the spectral efficiency of V2X communication could be greatly improved over conventional schemes, such as exhaustive search. Yet, [7] assumes a priori knowledge of the statistical mobility model of the UE, which needs to be learned in practice. Since the infinite-horizon POMDP in [7] depends on the unknown mobility dynamics and incurs a large optimization cost, the scheme therein is not amenable to simultaneous estimation of mobility and optimization of POMDP policy. In contrast to [7], herein, we decouple the beam-alignment design and mobility estimation by proposing a dual timescale approach in which the training of the mobility learning framework is carried on the long-timescale, interleaved with the execution of the policy in the short timescale: in the long timescale, a stochastic model of beam dynamics is learned, which provides side information (in the form of a prior belief) to optimize the beam-alignment policy in the short timescale; on the other hand, in the short timescale (one frame duration), the beam-alignment procedure is optimized using the prior belief provided by the mobility learning framework, agnostic to beam dynamics. Since learning of the mobility model is decoupled from the beam-alignment policy optimization, learning and adaptation can be done concurrently, in contrast to [7]. Moreover, by aiming to maximize frame throughput, the short-timescale policy optimization favors accurate detection of the optimal BS-UE's beam pair, hence improving the ability to predict optimal beam association for the next frames, and indirectly maximizing throughput in the long timescale.

Contributions: In a nutshell, the contributions of Chapter 6 are summarized as follows:

1. We propose a dual timescale approach in which the dynamics of the strongest beam pairs are learned over the long timescale, and then exploited over the short timescale to optimize the beam-training procedure;

2. We propose an R-VAE-based mobility learning framework to learn the dynamics of the strongest beam pairs over the long timescale, trained via stochastic gradient descent using beam-training observations;

3. We formulate a POMDP framework to optimize the decision-making process of beam-training and data transmission in the short-timescale, which uses the prior belief of the strongest beam pair provided by the R-VAE and beam-training feedback to maximize the average throughput. To solve the POMDP, we propose a linear time PBVI algorithm to find the approximately optimal policy;

4. For the special case of error-free feedback, we show that the POMDP can be reduced to a MDP with states as belief supports. Through structural properties of the MDP, we reveal that it is optimal to scan the most likely beams only, which enables a further reduction of state dimensionality. We propose a low-complexity value iteration algorithm that exploits the state-space reduction, and we demonstrate near-optimal performance in regimes with low feedback error rates.

1.6 Outline

The rest of this thesis is organized as follows. In Chapter 2, we present the energy-efficient interactive beam-alignment. In Chapter 3, we present the coded energy-efficient beam-alignment. In Chapter 4, we present the second-best beam-alignment via Bayesian multi-armed bandits. In Chapter 5, we present mobility and blockage aware communications in millimeter-wave vehicular networks. In Chapter 6, we present the recurrent variation autoencoder aided beam-training design. Finally, the thesis is concluded in Chapter 7.

2. ENERGY-EFFICIENT INTERACTIVE BEAM ALIGNMENT FOR MILLIMETER-WAVE NETWORKS

This chapter investigates the design of an optimal interactive beam-alignment and data communication protocol, with the goal of minimizing power consumption under a minimum rate constraint. The base-station selects beam-alignment or data communication and the beam parameters, based on feedback from the user-end. Based on the sectored antenna model and uniform prior on the angles of departure and arrival (AoD/AoA), the optimality of a *fixed-length* beam-alignment phase followed by a data-communication phase is demonstrated. Moreover, a *decoupled fractional* beam-alignment method is shown to be optimal, which decouples over time the alignment of AoD and AoA, and iteratively scans a fraction of their region of uncertainty. A heuristic policy is proposed for non-uniform prior on AoD/AoA, with provable performance guarantees, and it is shown that the uniform prior is the worst-case scenario. The performance degradation due to detection errors is studied analytically and via simulation. The numerical results with analog beams depict up to 4dB, 7.5dB, and 14dB gains over a state-of-the-art bisection method, conventional and interactive exhaustive search policies, respectively, and demonstrate that the sectored model provides valuable insights for beam-alignment design.

2.1 System Model

We consider a downlink scenario in a mm-wave cellular system with one base-station (BS) and one mobile user (UE) at distance d from the BS, both equipped with uniform linear arrays (ULAs) with $M_{\rm t}$ and $M_{\rm r}$ antennas, respectively, depicted in Fig. 2.1. Communication occurs over frames of fixed duration $T_{\rm fr}$, each composed of N slots indexed by $\mathcal{I}\equiv\{0,1,\ldots,N-1\}$ of duration $T=T_{\rm fr}/N$, each carrying S symbols of duration $T_{\rm sy}=T/S$. Let s be the transmitted symbol, with $\mathbb{E}[|s|^2]=1$. Then, the signal received at the UE is

$$y=\sqrt{P}\mathbf{c}_{\rm r}^H\mathbf{H}\mathbf{c}_{\rm t}s+\mathbf{c}_{\rm r}^H\mathbf{w}, \tag{2.1}$$

↑A version of this chapter was previously published by IEEE Transactions Wireless Communication [4][DOI: 10.1109/TWC.2018.2885041]

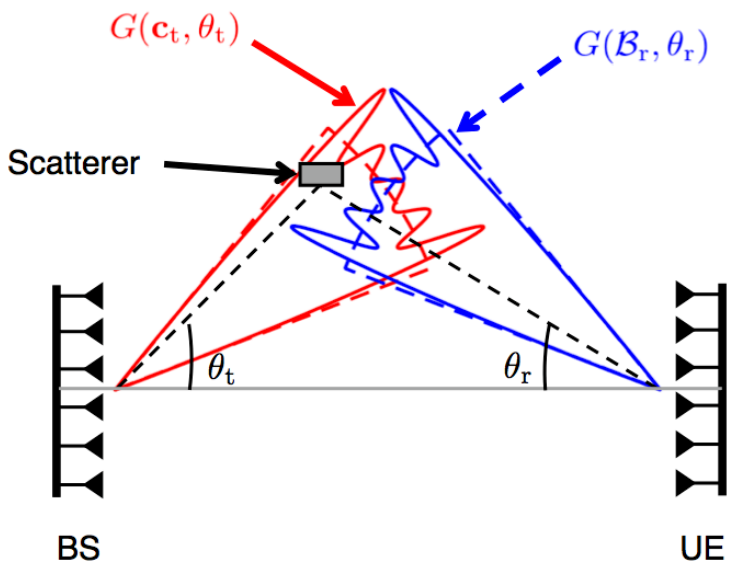

Figure 2.1. Actual beam pattern $G(\mathbf{c}_\mathrm{x}, \theta_\mathrm{x})$ generated using the algorithm in [27] with M_t=M_r=128 antennas. (solid lines) versus sectored model $G(\mathcal{B}_\mathrm{x}, \theta_\mathrm{x})$ (dashed lines) [29], on a linear scale. Sidelobes are not visible due to their small magnitude.

where P is the average transmit power of the BS; $\mathbf{H}\in\mathbb{C}^{M_\mathrm{r}\times M_\mathrm{t}}$ is the channel matrix; $\mathbf{c}_\mathrm{t}\in\mathbb{C}^{M_\mathrm{t}}$ is the BS beam-forming vector; $\mathbf{c}_\mathrm{r}\in\mathbb{C}^{M_\mathrm{r}}$ is the UE combining vector; $\mathbf{w} \sim \mathcal{CN}(\mathbf{0}, N_0 W_\mathrm{tot}\mathbf{I})$ is additive white Gaussian noise (AWGN). The symbols N_0 and W_tot denote the one-sided power spectral density of AWGN and the system bandwidth, respectively. By assuming analog beam-forming at both BS and UE, $\mathbf{c}_\mathrm{t}$ and $\mathbf{c}_\mathrm{r}$ satisfy the unit norm constraints $\|\mathbf{c}_\mathrm{t}\|_2^2$= $\|\mathbf{c}_\mathrm{r}\|_2^2$= 1. The channel matrix $\mathbf{H}$ follows the extended Saleh-Valenzuela geometric model [55],

$$\mathbf{H} = \sqrt{\frac{M_\mathrm{t} M_\mathrm{r}}{K}} \sum_{\ell=1}^{K} h_\ell \, \mathbf{d}_\mathrm{r}(\theta_{\mathrm{r},\ell}) \mathbf{d}_\mathrm{t}^H(\theta_{\mathrm{t},\ell}), \tag{2.2}$$

where $h_\ell \in \mathbb{C}$, $\theta_{\mathrm{t},\ell}$ and $\theta_{\mathrm{r},\ell}$ denote the small scale fading coefficient, AoD and AoA of the ℓ^{th} cluster, respectively. The terms $\mathbf{d}_\mathrm{r}(\theta_{\mathrm{r},\ell})\in\mathbb{C}^{M_\mathrm{r}}$ and $\mathbf{d}_\mathrm{t}(\theta_{\mathrm{t},\ell})\in\mathbb{C}^{M_\mathrm{t}}$ are the UE and BS array response vectors, respectively. For ULAs, $\theta_{\mathrm{t},\ell}$ (respectively, $\theta_{\mathrm{r},\ell}$) is the angle formed between

the outgoing (incoming) rays of the lth channel cluster and the perpendicular to the BS (UE) antenna array, as represented in Fig. 2.1, so that

$$\mathbf{d}_{\mathrm{x}}(\theta_{\mathrm{x}}) = \frac{1}{\sqrt{M_{\mathrm{x}}}} \left[1, e^{j\frac{2\pi d_{\mathrm{x}}}{\lambda}\sin\theta_{\mathrm{x}}}, \ldots, e^{j(M_{\mathrm{x}}-1)\frac{2\pi d_{\mathrm{x}}}{\lambda}\sin\theta_{\mathrm{x}}}\right]^{\top},$$

where $\mathrm{x} \in \{\mathrm{t}, \mathrm{r}\}$, d_{t} and d_{r} are the antenna spacing of the BS and UE arrays, respectively, λ is the wavelength of the carrier signal. In (2.2), $K \geq \mathrm{rank}(\mathbf{H})$ is the total number of clusters. Note that $\mathbf{H}$ has low-rank if $K \ll \min\{M_{\mathrm{t}}, M_{\mathrm{r}}\}$. In this chapter, we assume that there is a single dominant cluster (K=1). This assumption has been adopted in several previous works (e.g., see [56], [57]), and is motivated by channel measurements and modeling works such as [2], where it is shown that, in *dense urban environments*, with high probability the mm-wave channel exhibits only one or two clusters, with the dominant one containing most of the signal energy. While our analysis is based on a single cluster model, in Sec. 2.7 we demonstrate by simulation that the proposed scheme is robust also against multiple clusters. For the single cluster model, we obtain

$$\mathbf{H} = \sqrt{M_{\mathrm{t}} M_{\mathrm{r}}}\, h\, \mathbf{d}_{\mathrm{r}}(\theta_{\mathrm{r}})\mathbf{d}_{\mathrm{t}}^{H}(\theta_{\mathrm{t}}), \tag{2.3}$$

where $\mathbb{E}[|h|^2] = 1/\ell(d)$, $\ell(d)$ denotes the path loss between BS and UE as a function of distance d, and $\boldsymbol{\theta} = (\theta_{\mathrm{t}}, \theta_{\mathrm{r}})$ is the single-cluster AoD/AoA pair. We assume that $\boldsymbol{\theta}$ has prior joint distribution $f_0(\boldsymbol{\theta})$ with support $\mathrm{supp}(f_0)$=$\mathcal{U}_{\mathrm{t},0} \times \mathcal{U}_{\mathrm{r},0}$, which reflects the availability of prior AoD/AoA information acquired from previous beam-alignment phases, or based on geometric constraints (*e.g.*, presence of buildings blocking the signal in certain directions). We assume that h and $\boldsymbol{\theta}$ do not change over a frame, whose duration T_{fr} is chosen based upon the channel and beam coherence times T_c and T_b (time duration over which the AoD/AoA do not change appreciably) to satisfy this property. In [58], it has been reported that $T_c \ll T_b$. In the numerical values given below, $T_b \sim 100 T_c$. Therefore, by choosing $T_{\mathrm{fr}} \leq T_c$, we ensure that the variations in h and $\boldsymbol{\theta}$ over the frame duration T_{fr} are small and can be ignored. For example, using the relationships of T_c and T_b in [58], we obtain $T_c \simeq 10$[ms] and $T_b \simeq 1$[s] for a UE velocity of 100[km/h]. In our numerical evaluations, we will therefore use T_{fr}=10[ms]. It is

noteworthy that this assumption has also been used extensively in previous beam-alignment works, such as [23], [24], [56].

We assume that blockage occurs at longer time-scales than the frame duration, determined by the geometry of the environment and mobility of users, hence we neglect blockage dynamics within a frame duration [59]. By replacing (2.3) into (2.1), and defining the BS and UE beam-forming gains $G_{\mathrm{x}}(\mathbf{c}_{\mathrm{x}},\theta_{\mathrm{x}}) = M_{\mathrm{x}}|\mathbf{d}_{\mathrm{x}}^{H}(\theta_{\mathrm{x}})\mathbf{c}_{\mathrm{x}}|^2, \mathrm{x}\in\{\mathrm{t},\mathrm{r}\}$, we get

$$y = h\sqrt{PG_{\mathrm{t}}(\mathbf{c}_{\mathrm{t}},\theta_{\mathrm{t}})\cdot G_{\mathrm{r}}(\mathbf{c}_{\mathrm{r}},\theta_{\mathrm{r}})}e^{j\Psi(\boldsymbol{\theta})}s + \hat{w}, \tag{2.4}$$

where $\hat{w}\triangleq\mathbf{c}_{\mathrm{r}}^{H}\mathbf{w}\sim\mathcal{CN}(0,N_0W_{\mathrm{tot}})$ is the noise component and $\Psi(\boldsymbol{\theta}) = \angle\mathbf{d}_{\mathrm{t}}^{H}(\theta_{\mathrm{t}})\mathbf{c}_{\mathrm{t}} - \angle\mathbf{d}_{\mathrm{r}}^{H}(\theta_{\mathrm{r}})\mathbf{c}_{\mathrm{r}}$ is the phase.

In this chapter, we use the *sectored antenna* model [29] to approximate the BS and UE beam-forming gains [60], represented in Fig. 2.1. Under this model,

$$G_{\mathrm{x}}(\mathbf{c}_{\mathrm{x}},\theta_{\mathrm{x}}) \approx G_{\mathrm{x}}(\mathcal{B}_{\mathrm{x}},\theta_{\mathrm{x}}) = \frac{2\pi}{|\mathcal{B}_{\mathrm{x}}|}\,\chi_{\mathcal{B}_{\mathrm{x}}}(\theta_{\mathrm{x}}), \quad \mathrm{x}\in\{\mathrm{t},\mathrm{r}\}, \tag{2.5}$$

where $\mathcal{B}_{\mathrm{t}}\subseteq(-\pi/2,\pi/2]$ is the range of AoD covered by $\mathbf{c}_{\mathrm{t}}$, $\mathcal{B}_{\mathrm{r}}\subseteq(-\pi/2,\pi/2]$ is the range of AoA covered by $\mathbf{c}_{\mathrm{r}}$, $\chi_{\mathcal{A}}(\theta)$ is the indicator function of the event $\theta\in\mathcal{A}$, and $|\mathcal{A}| = \int_{\mathcal{A}}\mathrm{d}\theta$ is the measure of the set $\mathcal{A}$. Hereafter, the two sets $\mathcal{B}_{\mathrm{t}}$ and $\mathcal{B}_{\mathrm{r}}$ will be referred to as BS and UE beams, respectively. Additionally, we define $\mathcal{B}_k = \mathcal{B}_{\mathrm{t},k}\times\mathcal{B}_{\mathrm{r},k}$ as the 2-dimensional (2D) AoD/AoA support defined by the BS-UE beams. Note that the sectored model is used as an abstraction of the real model, which applies a precoding vector $\mathbf{c}_{\mathrm{t}}$ at the transmitter and a beamforming vector $\mathbf{c}_{\mathrm{r}}$ at the receiver. This abstraction, shown in Fig. 2.1, is adopted since direct optimization of $\mathbf{c}_{\mathrm{t}}$ and $\mathbf{c}_{\mathrm{r}}$ is not analytically tractable, due to the high dimensionality of the problem. In Sec. 2.7 we show via Monte-Carlo simulation that, by appropriate design of $\mathbf{c}_{\mathrm{t}}$ and $\mathbf{c}_{\mathrm{r}}$ to approximate the sectored model, our scheme attains near-optimal performance, and outperforms a state-of-the-art bisection search scheme [13]; thus, the sectored antenna

model provides a valuable abstraction for practical design. Following the sectored antenna model, we obtain the received signal by replacing $G_{\mathrm{x}}(\mathbf{c}_{\mathrm{x}}, \theta_{\mathrm{x}})$ with $G_{\mathrm{x}}(\mathcal{B}_{\mathrm{x}}, \theta_{\mathrm{x}})$ in (2.4), yielding

$$y = h\sqrt{PG_{\mathrm{t}}(\mathcal{B}_{\mathrm{t}}, \theta_{\mathrm{t}}) \cdot G_{\mathrm{r}}(\mathcal{B}_{\mathrm{r}}, \theta_{\mathrm{r}})}e^{j\Psi(\boldsymbol{\theta})}s + \hat{w}. \tag{2.6}$$

Although the analysis in this chapter is presented for ULAs (2D beamforming), the proposed scheme can be extended to the case of uniform planar arrays with 3D beamforming, by interpreting $\theta_{\mathrm{x}}, \mathrm{x} \in \{\mathrm{t}, \mathrm{r}\}$ as a vector denoting the azimuth and elevation pair in $(-\pi/2, \pi/2]^2$ and the beam $\mathcal{B}_{\mathrm{x}} \subseteq (-\pi/2, \pi/2]^2$. For notational convenience and ease of exposition, in this chapter we focus on the 2D beamforming case (also adopted in, e.g., [13], [23], [27], [28]).

The entire frame duration is split into two, possibly interleaved phases: a beam-alignment phase, whose goal is to detect the best beam to be used in the data communication phase. To this end, we partition the slots $\mathcal{I}$ in each frame into the indices in the set $\mathcal{I}_s$, reserved for beam-alignment, and those in the set $\mathcal{I}_d$, reserved for data communication, where $\mathcal{I}_s \cap \mathcal{I}_d = \emptyset$ and $\mathcal{I}_s \cup \mathcal{I}_d = \mathcal{I}$. The optimal frame partition and duration of beam-alignment are part of our design. In the sequel, we describe the operations performed in the beam-alignment and data communication slots, and characterize their energy consumption.

Beam-Alignment: At the beginning of each slot $k \in \mathcal{I}_s$, the BS sends a beacon signal $\mathbf{s}$ of duration $T_B < T$ using the transmit beam $\mathcal{B}_{\mathrm{t},k}$ with power P_k,[1] and the UE receives the signal using the receive beam $\mathcal{B}_{\mathrm{r},k}$. Note that $\mathcal{B}_k = \mathcal{B}_{\mathrm{t},k} \times \mathcal{B}_{\mathrm{r},k}$ and P_k are design parameters. If the UE detects the beacon (*i.e.*, the AoD/AoA $\boldsymbol{\theta}$ is in $\mathcal{B}_k$, or a false-alarm occurs, see [60]), then, in the remaining portion of the slot of duration $T - T_B$, it transmits an acknowledgment (ACK) packet to the BS, denoted as C_k=ACK. Otherwise (the UE does not detect the beacon due to either mis-alignment or misdetection error), it transmits C_k=NACK. We assume that the ACK/NACK signal C_k is received perfectly and within the end of the slot by the BS (for instance, by using a conventional microwave technology as a control channel [62]).

[1] ↑In practice, there are limits on how small the beacon duration can be made, due to peak power constraints [61], beacon synchronization errors [3], and auto-correlation properties of the beacon sequence [3].

As a result of (2.6), the UE attempts to detect the beam, and generates the ACK/NACK signal based on the following hypothesis testing problem,

$$\mathcal{H}_1 : \mathbf{y}_k = \sqrt{N_0 W_{\text{tot}} \nu_k} h e^{j\Psi_k(\boldsymbol{\theta})} \mathbf{s} + \hat{\mathbf{w}}_k, \qquad \text{(alignment, } \boldsymbol{\theta} \in \mathcal{B}_k\text{)} \tag{2.7}$$

$$\mathcal{H}_0 : \mathbf{y}_k = \hat{\mathbf{w}}_k, \qquad \text{(misalignment, } \boldsymbol{\theta} \notin \mathcal{B}_k\text{)} \tag{2.8}$$

where $\mathbf{y}_k$ is the received signal, $\mathbf{s}$ is the transmitted symbol sequence, $\hat{\mathbf{w}}_k \sim \mathcal{CN}(\mathbf{0}, N_0 W_{\text{tot}} \mathbf{I})$ is the AWGN vector, and ν_k is related to the beam-forming gain in slot k,

$$\nu_k = \frac{(2\pi)^2 P_k}{N_0 W_{\text{tot}} \, |\mathcal{B}_k|}. \tag{2.9}$$

The optimal detector depends on the availability of prior information on h. We assume that an estimate of the channel gain $\gamma = |h|^2$ is available at the BS and UE at the beginning of each frame, denoted as $\hat{\gamma} = |\hat{h}|^2$, where $\hat{h} = h + e$ and $e \sim \mathcal{CN}(0, \sigma_{\text{e}}^2)$ denotes the estimation noise. A Neyman-Pearson threshold detector is optimal in this case,

$$\frac{|\mathbf{s}^H \mathbf{y}_k|^2}{N_0 W_{\text{tot}} \|\mathbf{s}\|_2^2} \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\lessgtr}} \tau_{\text{th}}. \tag{2.10}$$

The detector's threshold τ_{th} and the transmission power P_k are designed based on the channel gain estimate $\hat{\gamma}$, so as to satisfy constraints on the false-alarm and misdetection probabilities, $p_{\text{fa}}, p_{\text{md}} \leq p_e$. We now compute these probabilities under the simplifying assumption that $\hat{h}$ and e are independent, so that $h|\hat{h} \sim \mathcal{CN}(\hat{h}, \sigma_{\text{e}}^2)$. Let $z_k \triangleq \frac{\mathbf{s}^H \mathbf{y}_k}{\sqrt{N_0 W_{\text{tot}}} \|\mathbf{s}\|_2}$, so that $|z_k|^2$ is the decision variable. We observe that

$$z_k = \begin{cases} \sqrt{\nu_k} h e^{j\Psi_k(\boldsymbol{\theta})} \|\mathbf{s}\|_2 + \frac{\mathbf{s}^H \hat{\mathbf{w}}_k}{\sqrt{N_0 W_{\text{tot}}} \|\mathbf{s}\|_2}, & \text{if } \mathcal{H}_1 \text{ is true;} \\ \frac{\mathbf{s}^H \hat{\mathbf{w}}_k}{\sqrt{N_0 W_{\text{tot}}} \|\mathbf{s}\|_2}, & \text{if } \mathcal{H}_0 \text{ is true.} \end{cases}$$

Since $\hat{h}$ and $\hat{\mathbf{w}}_k$ are independent and $h = \hat{h} - e$, we obtain

$$f(z_k|\hat{h}, \mathcal{H}_1, \boldsymbol{\theta}) = \mathcal{CN}\left(\sqrt{\nu_k}\hat{h}e^{j\Psi_k(\boldsymbol{\theta})}\|\mathbf{s}\|_2, 1+\nu_k\|\mathbf{s}\|_2^2\sigma_{\mathrm{e}}^2\right), \tag{2.11}$$

$$f(z_k|\hat{h}, \mathcal{H}_0) = \mathcal{CN}\left(0, 1\right), \tag{2.12}$$

so that $[|z_k|^2|\hat{h}, \mathcal{H}_0] \sim$ Exponential(1), and the false-alarm probability can be expressed as

$$p_{\mathrm{fa}}(\tau_{\mathrm{th}}) \triangleq \mathbb{P}\left(|z_k|^2 > \tau_{\mathrm{th}}|\hat{h}, \mathcal{H}_0\right) = \exp\left(-\tau_{\mathrm{th}}\right). \tag{2.13}$$

Similarly, the misdetection probability is found to be

$$p_{\mathrm{md}}(\nu_k, \tau_{\mathrm{th}}, \hat{\gamma}) \triangleq \mathbb{P}\left(|z_k|^2 < \tau_{\mathrm{th}}|\hat{h}, \mathcal{H}_1\right) \quad = 1 - Q_1\left(\sqrt{\frac{2\hat{\gamma}\nu_k\|\mathbf{s}\|_2^2}{1+\nu_k\|\mathbf{s}\|_2^2\sigma_{\mathrm{e}}^2}}, \sqrt{\frac{2\tau_{\mathrm{th}}}{1+\nu_k\|\mathbf{s}\|_2^2\sigma_{\mathrm{e}}^2}}\right), \tag{2.14}$$

where $Q_1(\cdot)$ is the first-order Marcum's Q function [63]. In fact, $z_k|(\hat{h}, \mathcal{H}_1)$ is complex Gaussian as in (2.11), so that, given $(\hat{\gamma}, \mathcal{H}_1)$, $\frac{2|z_k|^2}{1+\nu_k\|\mathbf{s}\|_2^2\sigma_{\mathrm{e}}^2}$ follows non-central chi-square distribution with 2 degrees of freedom and non-centrality parameter $\frac{2\nu_k\hat{\gamma}\|\mathbf{s}\|_2^2}{1+\nu_k\|\mathbf{s}\|_2^2\sigma_{\mathrm{e}}^2}$.

Herein, we design τ_{th} and P_k to achieve $p_{\mathrm{fa}}, p_{\mathrm{md}} \leq p_{\mathrm{e}}$. To satisfy $p_{\mathrm{fa}}(\tau_{\mathrm{th}}) \leq p_{\mathrm{e}}$ we need

$$\tau_{\mathrm{th}} \geq -\ln\left(p_{\mathrm{e}}\right). \tag{2.15}$$

Since $Q_1(a, b)$ is an increasing function of $a \geq 0$ and a decreasing function of $b \geq 0$, it follows that $p_{\mathrm{md}}(\nu_k, \tau_{\mathrm{th}})$ is a decreasing function of $\nu_k \geq 0$ and an increasing function of $\tau_{\mathrm{th}} \geq 0$. Then, to guarantee $p_{\mathrm{md}}(\nu_k, \tau_{\mathrm{th}}, \hat{\gamma}) \leq p_{\mathrm{e}}$, (2.15) should be satisfied with equality to attain the smallest p_{md}; additionally, there exists $\nu^* > 0$, determined as the unique solution of $p_{\mathrm{md}}(\nu^*, \tau_{\mathrm{th}}, \hat{\gamma}) = p_{\mathrm{e}}$ and independent of the beam shape $\mathcal{B}_k$, such that $p_{\mathrm{md}}(\nu_k, \tau_{\mathrm{th}}, \hat{\gamma}) \leq p_{\mathrm{e}}$ iff (if and only if) $\nu_k \geq \nu^*$.

Then, using (2.9) and letting $E_k \triangleq P_k T_{\rm sy} \|\mathbf{s}\|_2^2$ be the energy incurred for the transmission of the beacon $\mathbf{s}$ in slot k, E_k should satisfy

$$E_k \geq \phi_s(p_{\rm e})\,|\mathcal{B}_k|\,, \tag{2.16}$$

$$\text{where } \phi_s(p_{\rm e}) \triangleq N_0 W_{\rm tot} \nu^* T_{\rm sy} \|\mathbf{s}\|_2^2/(2\pi)^2 \tag{2.17}$$

is the energy/rad^2 required to achieve false-alarm and misdetection probabilities equal to $p_{\rm e}$.

Note that false-alarm and misdetection errors are deleterious to performance, since they result in mis-alignment and outages during data transmission. Therefore, they should be minimized. For this reason, in the first part of this chapter we assume that $p_{\rm e} \ll 1$, and neglect the impact of these errors on beam-alignment. Thus, we let $E_k \geq \phi_s\,|\mathcal{B}_k|$ be the energy required in each beam-alignment slot to guarantee detection with high probability, where ϕ_s is computed under some small $p_{\rm e} \ll 1$. We will consider the impact of these errors in Sec. 2.6.[2]

Data Communication: In the communication slots indexed by $k \in \mathcal{I}_d$, the BS uses $\mathcal{B}_{{\rm t},k}$, rate R_k, and transmit power P_k, while the UE processes the received signal using the beam $\mathcal{B}_{{\rm r},k}$. Therefore, letting $\gamma = |h|^2$ and ν_k as in (2.9). the instantaneous SNR can be expressed as

$$\mathsf{SNR}_k = \frac{\gamma P_k G_{\rm t}(\mathcal{B}_{{\rm t},k}, \theta_{\rm t}) G_{\rm r}(\mathcal{B}_{{\rm r},k}, \theta_{\rm r})}{N_0 W_{\rm tot}} = \nu_k \gamma \chi_{\mathcal{B}_{{\rm t},k}}(\theta_{\rm t}) \chi_{\mathcal{B}_{{\rm r},k}}(\theta_{\rm r}). \tag{2.18}$$

Outage occurs if $W_{\rm tot} \log_2(1+\mathsf{SNR}_k) < R_k$ due to either mis-alignment between transmitter and receiver, or low channel gain γ. The probability of this event, p_{out}, can be inferred from the posterior probability distribution of the AoD/AoA pair $\boldsymbol{\theta}$ and the channel gain γ,

[2] The design of beam-alignment schemes robust to errors when $p_{\rm e} \not\ll 1$ has been considered in [5]. Its analysis is outside the scope of this chapter.

given its estimate $\hat{\gamma}$, and the history of BS-UE beams and feedback until slot k, denoted as $\mathcal{H}^k \triangleq \{(\mathcal{B}_0, C_0), \ldots, (\mathcal{B}_{k-1}, C_{k-1})\}$. Thus, $p_{out} \triangleq \mathbb{P}(W_{\rm tot}\log_2(1+\mathsf{SNR}_k) < R_k|\hat{\gamma}, \mathcal{H}^k)$, yielding

$$\begin{aligned} p_{out} &\overset{(a)}{=} \mathbb{P}\left(\mathsf{SNR}_k < 2^{\frac{R_k}{W_{\rm tot}}} - 1 | \hat{\gamma}, \boldsymbol{\theta} \in \mathcal{B}_k\right) \mathbb{P}(\boldsymbol{\theta} \in \mathcal{B}_k|\mathcal{H}^k) + \mathbb{P}(\boldsymbol{\theta} \notin \mathcal{B}_k|\mathcal{H}^k) \\ &\overset{(b)}{=} 1 - \bar{F}_\gamma\left(\frac{2^{\frac{R_k}{W_{\rm tot}}} - 1}{\nu_k}\Big|\hat{\gamma}\right) \mathbb{P}(\boldsymbol{\theta} \in \mathcal{B}_k|\mathcal{H}^k), \end{aligned} \tag{2.19}$$

where (a) follows from the law of total probability and $\mathbb{P}(\boldsymbol{\theta}\in\mathcal{B}_k|\mathcal{H}^k)$ denotes the probability of correct beam-alignment; (b) follows by substituting $\bar{F}_\gamma(x|\hat{\gamma})\triangleq\mathbb{P}(\gamma\geq x|\hat{\gamma})$ into (a), given as

$$\bar{F}_\gamma(x|\hat{\gamma}) = Q_1\left(\sqrt{2\hat{\gamma}/\sigma_e^2}, \sqrt{2x/\sigma_e^2}\right). \tag{2.20}$$

Herein, we use the notion of ϵ-outage capacity to design R_k, defined as the largest transmission rate such that $p_{out}\leq\epsilon$, for a target outage probability $\epsilon<1$. This can be expressed as

$$\mathcal{C}_\epsilon(P_k, \mathcal{B}_k|\mathcal{H}^k, \hat{\gamma}) \triangleq W_{\rm tot}\log_2\left(1 + \nu_k \bar{F}_\gamma^{-1}\left(\frac{1-\epsilon}{\mathbb{P}(\boldsymbol{\theta}\in\mathcal{B}_k|\mathcal{H}^k)}\Big|\hat{\gamma}\right)\right), \tag{2.21}$$

where $\bar{F}_\gamma^{-1}(\cdot|\hat{\gamma})$ denotes the inverse posterior CCDF of γ, conditional on $\hat{\gamma}$. In other words, if $R_k\leq\mathcal{C}_\epsilon(P_k, \mathcal{B}_k|\mathcal{H}^k, \hat{\gamma})$, then the transmission is successful with probability at least $1-\epsilon$, and the average rate is at least $(1-\epsilon)R_k$. Note that, in order to achieve the target $p_{out}\leq\epsilon$, the probability of correct beam-alignment must satisfy $\mathbb{P}(\boldsymbol{\theta}\in\mathcal{B}_k|\mathcal{H}^k)\geq 1-\epsilon$. This can be achieved with a proper choice of $\mathcal{B}_k$, as discussed next.

Since the ACK/NACK feedback after data communication is generated by higher layers (*e.g.,* network or transport layer), we do not use it to improve beam-alignment. We define C_k=NULL, $\forall k\in\mathcal{I}_d$, to distinguish it from the ACK/NACK feedback signal in the beam-alignment slots.

2.2 Problem Formulation

In this section, we formulate the optimization problem, and characterize it as a Markov decision process (MDP). The goal is to minimize the power consumption at the BS over a frame duration, while achieving the quality of service (QoS) requirements of the UE (rate and delay). Therefore, the objective function of the following optimization problem captures the beam-alignment and data communication energy costs; the QoS requirements are specified in the constraints through a rate requirement $R_{\min}$ of the UE along with an outage probability of ϵ; additionally, the frame duration $T_{\rm fr}$ represents a delay guarantee on data transmission. The design variables in slot k are denoted by the 4-tuple $\mathbf{a}_k = (\xi_k, P_k, \mathcal{B}_k, R_k)$, where ξ_k corresponds to the decision of whether to perform beam-alignment (ξ_k=1) or data communication (ξ_k=0); we let R_k=0 for beam-alignment slots ($\xi_k = 1$). With this choice of $\mathbf{a}_k$, we aim to optimally select the beam-alignment slots $\mathcal{I}_s$ and data communication slots $\mathcal{I}_d$. If a slots is selected for beam-alignment ($\xi_k = 1$), we aim to optimize the associated power P_k and 2D beam $\mathcal{B}_k$. Likewise, if a slot is selected for data communication ($\xi_k = 0$), we aim to optimize the associated power P_k, data rate R_k, and 2D beam $\mathcal{B}_k$. Mathematically, the optimization problem is stated as

$$\mathrm{P}_1: \quad \bar{P} \triangleq \min_{\mathbf{a}_0,\ldots,\mathbf{a}_{N-1}} \quad \frac{1}{T_{\rm fr}} \mathbb{E}\left[\sum_{k=0}^{N-1} E_k \middle| f_0\right] \tag{2.22}$$

$$\text{s.t.} \quad \mathbf{a}_k = (\xi_k, P_k, \mathcal{B}_k, R_k), \forall k,$$

$$\mathcal{B}_k = \mathcal{B}_{\mathrm{t},k} \times \mathcal{B}_{\mathrm{r},k} \subseteq \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]^2, \forall k, \tag{2.23}$$

$$E_k \geq \phi_s |\mathcal{B}_k|, \ \forall k \in \mathcal{I}_s, \tag{2.24}$$

$$\frac{1}{N}\sum_k R_k \geq R_{\min}, \ R_k \leq \mathcal{C}_\epsilon(P_k, \mathcal{B}_k | \mathcal{H}^k, \hat{\gamma}), \forall k \in \mathcal{I}_d, \tag{2.25}$$

$$P_k = E_k / [\xi_k T_B + (1 - \xi_k) T], \ \forall k, \tag{2.26}$$

where f_0 in (2.22) denotes the prior belief over $\boldsymbol{\theta}$; (2.23) defines the 2D beam $\mathcal{B}_k$; (2.24) gives the energy consumption in the beam-alignment slots; (2.25) ensures the rate requirement $R_{\min}$ over the frame, and that R_k is within the ϵ-outage capacity, see (2.21); (2.26) gives the

relation between energy and power.[3] Since the cost is the average BS power consumption, the inequality constraints (2.24)-(2.25) must be tight, i.e., we replace them with

$$E_k = \xi_k \phi_s \left|\mathcal{B}_k\right| + (1-\xi_k)\frac{\psi_d(R_k)\left|\mathcal{B}_k\right|}{\bar{F}_\gamma^{-1}\left(\frac{1-\epsilon}{\mathbb{P}(\boldsymbol{\theta}\in\mathcal{B}_k|\mathcal{H}^k)}|\hat{\gamma}\right)}, \tag{2.27}$$

$$\frac{1}{N}\sum_k R_k = R_{\min}, \tag{2.28}$$

where (2.27) when ξ_k=0 is obtained by inverting (2.25) via (2.21) and (2.9) (with equality) to find P_k and $E_k = P_k T$, and we have defined the energy/rad^2 required to achieve the rate R

$$\psi_d(R) \triangleq (2\pi)^{-2} N_0 W_{\text{tot}} T (2^{\frac{R}{W_{\text{tot}}}} - 1).$$

Hereafter, we exclude P_k from the design space, since it is uniquely defined by the set of equality constraints (2.26)-(2.27). Thus, we simplify the design variable to $\mathbf{a}_k = (\xi_k, \mathcal{B}_k, R_k)$.

We pose P_1 as an MDP [64] over the time horizon $\mathcal{I}$. The state at the start of slot k is (f_k, D_k), where f_k is the probability distribution over the AoD/AoA pair $\boldsymbol{\theta}$, given the history $\mathcal{H}^k$ up to slot k, denoted as *belief*; D_k is the backlog (untransmitted data bits). Initially, f_0 is the prior belief and $D_0 \triangleq R_{\min} T_{\text{fr}}$. Given (f_k, D_k), the BS and UE select $\mathbf{a}_k = (\xi_k, \mathcal{B}_k, R_k)$.[4] Then, the UE generates the feedback signal: if ξ_k=0 (data communication), then C_k=NULL; if ξ_k=1 (beam-alignment), then C_k=ACK if $\boldsymbol{\theta}\in\mathcal{B}_k$, with probability

$$\mathbb{P}(C_k = \text{ACK}|f_k, \mathbf{a}_k) = \int_{\mathcal{B}_k} f_k(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{2.29}$$

and C_k=NACK otherwise. Upon receiving C_k, the new backlog in slot $k+1$ becomes[5]

$$D_{k+1} = \max\{D_k - R_k T, 0\}, \tag{2.30}$$

and the new belief f_{k+1} is computed via Bayes' rule, as given in the following lemma.

[3]↑Data communication takes the entire slot, whereas beam-alignment occurs over a portion $T_B < T$ of the slot to allow for the time to receive the ACK/NACK feedback from the receiver.
[4]↑Since feedback is error-free, both BS and UE have the same information to generate the action $\mathbf{a}_k$ and their beams.
[5]↑If $D_{k+1} \leq 0$, all bits have been transmitted.

Lemma 2.1. *Let f_0 be the prior belief on $\boldsymbol{\theta}$ with support* $\mathrm{supp}(f_0)=\mathcal{U}_0$. *Then,*

$$f_k(\boldsymbol{\theta}) = \frac{f_0(\boldsymbol{\theta})}{\int_{\mathcal{U}_k} f_0(\tilde{\boldsymbol{\theta}})\mathrm{d}\tilde{\boldsymbol{\theta}}}\chi_{\mathcal{U}_k}(\boldsymbol{\theta}), \tag{2.31}$$

where $\mathcal{U}_k \triangleq \mathrm{supp}(f_k)$ *is updated recursively as*

$$\mathcal{U}_{k+1} = \begin{cases} \mathcal{U}_k \cap \mathcal{B}_k, & k \in \mathcal{I}_\mathrm{s}, C_k = \mathrm{ACK} \\ \mathcal{U}_k \setminus \mathcal{B}_k, & k \in \mathcal{I}_\mathrm{s}, C_k = \mathrm{NACK} \\ \mathcal{U}_k, & k \in \mathcal{I}_\mathrm{d}. \end{cases} \tag{2.32}$$

Proof. The proof follows by induction using Bayes' rule. In fact, if C_k=ACK in a beam-alignment slot, then it can be inferred that $\boldsymbol{\theta}\in\mathcal{U}_k\cap\mathcal{B}_k$; otherwise ($C_k$=NACK) the UE lies outside $\mathcal{B}_k$, but within the support of f_k, i.e., $\boldsymbol{\theta}\in\mathcal{U}_k\setminus\mathcal{B}_k$. In the data communication slots, no feedback is generated, hence $f_{k+1}=f_k$ and $\mathcal{U}_{k+1}=\mathcal{U}_k$. A detailed proof is given in Appendix 8.A. □

Lemma 2.1 implies that $\mathcal{U}_k$ is a sufficient statistic for decision making in slot k, and is updated recursively via (2.32). Accordingly, the state space is defined as

$$\mathcal{S} \equiv \{(\mathcal{U}, D) : \mathcal{U} \subseteq \mathcal{U}_0, 0 \le D \le D_0\}. \tag{2.33}$$

Given the data backlog D_k=D, the action space is expressed as[6]

$$\begin{aligned} \mathcal{A}(D) \equiv &\left\{(0, \mathcal{B}, R) : \mathcal{B} \equiv \mathcal{B}_\mathrm{t} \times \mathcal{B}_\mathrm{r} \subseteq \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]^2, 0<R\le D/T\right\} \\ &\cup \left\{(1, \mathcal{B}, 0) : \mathcal{B} \equiv \mathcal{B}_\mathrm{t} \times \mathcal{B}_\mathrm{r} \subseteq \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]^2\right\}. \end{aligned} \tag{2.34}$$

Given $(\mathcal{U}_k, D_k) \in \mathcal{S}$, the action $\mathbf{a}_k\in\mathcal{A}(D_k)$ is chosen based on policy μ_k, which determines the BS-UE beam $\mathcal{B}_k$ and whether to perform beam-alignment (ξ_k=1, R_k=0) or data communication (ξ_k=0, R_k>0), with energy cost E_k given by (2.27). With this notation, we can express

[6]↑Note that, for a data communication action $(0, \mathcal{B}, R)$, we assume that $R > 0$; in fact, data communication with zero rate is equivalent to a beam-alignment action $(1, \emptyset, 0)$ with empty beam.

the problem P_1 as that of finding the policy μ^* which minimizes the power consumption under rate requirement and outage probability constraints,

$$\begin{aligned} \mathrm{P}_2: \quad & \bar{P} \triangleq \min_{\mu} \frac{1}{T_{\mathrm{fr}}} \mathbb{E}_{\mu}\left[\sum_{k=0}^{N-1} c(\mathbf{a}_k; \mathcal{U}_k, D_k) \middle| \mathcal{U}_0, D_0, f_0\right], \\ \text{s.t.} \quad & D_{k+1} = D_k - TR_k, \forall k \in \mathcal{I}, \;\; D_N = 0, \end{aligned} \tag{2.35}$$

where we have defined the cost per stage in state $(\mathcal{U}_k, D_k)$ under action $\mathbf{a}_k$ as

$$c(\mathbf{a}_k; \mathcal{U}_k, D_k) = \left[\xi_k \phi_s + \frac{(1-\xi_k)\psi_d(R_k)}{\bar{F}_{\gamma}^{-1}\left(\frac{1-\epsilon}{\mathbb{P}(\boldsymbol{\theta} \in \mathcal{B}_k | \mathcal{U}_k)} | \hat{\gamma}\right)}\right] |\mathcal{B}_k|\,, \tag{2.36}$$

and we used the sufficient statistic (Lemma 2.1) to express $\mathbb{P}(\boldsymbol{\theta} \in \mathcal{B}_k | \mathcal{H}^k) = \mathbb{P}(\boldsymbol{\theta} \in \mathcal{B}_k | \mathcal{U}_k)$ in (2.27). P_2 can be solved via dynamic programming (DP): the value function in state $(\mathcal{U}_k, D_k)$ under action $\mathbf{a}_k \in \mathcal{A}(D_k)$, $V_k(\mathbf{a}_k; \mathcal{U}_k, D_k)$, and the optimal value function, $V_k^*(\mathcal{U}_k, D_k)$, are expressed as

$$\begin{aligned} V_k(\mathbf{a}_k; \mathcal{U}_k, D_k) &= c(\mathbf{a}_k; \mathcal{U}_k, D_k) + \mathbb{E}\left[V_{k+1}^*(\mathcal{U}_{k+1}, D_{k+1}) \middle| \mathcal{U}_k, D_k; \mathbf{a}_k\right], \\ V_k^*(\mathcal{U}_k, D_k) &= \min_{\mathbf{a}_k \in \mathcal{A}(D_k)} V_k(\mathbf{a}_k; \mathcal{U}_k, D_k), \end{aligned} \tag{2.37}$$

where the minimum is attained by the optimal policy. To enforce $D_N = 0$, we initialize it as

$$V_N^*(\mathcal{U}_N, D_N) = \begin{cases} 0, & D_N = 0 \\ \infty, & D_N > 0. \end{cases} \tag{2.38}$$

Further analysis is not doable for a generic prior f_0. To unveil structural properties, we proceed as follows:

1. We optimize over the extended action space

$$\mathcal{A}_{\mathrm{ext}}(D) \equiv \left\{(0, \mathcal{B}, R) : \mathcal{B} \subseteq \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]^2, 0 < R \leq D/T\right\} \cup \left\{(1, \mathcal{B}, 0) : \mathcal{B} \subseteq \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]^2\right\}, \tag{2.39}$$

obtained by removing the "rectangular beam" constraint $\mathcal{B}\equiv\mathcal{B}_{\rm t}\times\mathcal{B}_{\rm r}$ in (2.34). Thus, $\mathcal{B}\in\mathcal{A}_{\rm ext}(D)$ can be any subset of $[-\pi/2,\pi/2]^2$, not restricted to a "rectangular" shape $\mathcal{B}\equiv\mathcal{B}_{{\rm t},k}\times\mathcal{B}_{{\rm r},k}$. By optimizing over an extended action space, a lower bound to the value function is obtained, denoted as $\hat{V}_k^*(\mathcal{U}_k,D_k)\leq V_k^*(\mathcal{U}_k,D_k)$, possibly not achievable by a "rectangular" beam.

2. In Sec. 2.3, we find structural properties under such extended action space, for the case of a uniform belief f_0. In this setting, we prove the optimality of a *fractional search* method, which selects $\mathcal{B}_k$ as $\mathcal{B}_k\subseteq\mathcal{U}$ with $|\mathcal{B}_k|=\rho_k|\mathcal{U}_k|$ (beam-alignment) or $|\mathcal{B}_k|=\vartheta|\mathcal{U}_k|$ (data communication), for appropriate *fractional parameters* ρ_k and ϑ; additionally, we prove the optimality of a *deterministic* duration of the beam-alignment phase (Theorems 2.1 and 2.3).

3. In Sec. 2.4, we prove that such lower bound is indeed achievable by a *decoupled fractional search* method, which decouples the BS and UE beam-alignment over time using rectangular beams, hence it is optimal.

4. In Sec. 2.5, we use these results to design a heuristic policy with performance guarantees for the case of non-uniform prior f_0, and show that the uniform prior is the worst case.

2.3 Uniform Prior

We denote the beam $\mathcal{B}$ taking value from the extended action space $\mathcal{A}_{\rm ext}(D)$ as "2D beam", to distinguish it from $\mathcal{B}\in\mathcal{A}(D)$, that obeys a "rectangular" constraint. Additionally, since the goal is to minimize the energy consumption, we restrict $\mathcal{B}\subseteq\mathcal{U}$ during data communication and $\mathcal{B}\subset\mathcal{U}$ during beam-alignment, yielding the following extended action space in state $(\mathcal{U},D)$:[7]

$$\mathcal{A}_{\rm ext}(\mathcal{U},D)\equiv\{(0,\mathcal{B},R):\mathcal{B}\subseteq\mathcal{U},0<R\leq D/T\}\cup\{(1,\mathcal{B},0):\mathcal{B}\subset\mathcal{U}\}\,. \tag{2.40}$$

[7]↑In fact, the AoD/AoA lie within the belief support $\mathcal{U}_k$; projecting a "2D beam" outside of $\mathcal{U}_k$ is suboptimal, since it yields an unnecessary energy cost. Additionally, choosing $\mathcal{B}_k=\mathcal{U}_k$ during beam-alignment is suboptimal, since it triggers an ACK with probability one, which is uninformative; we thus restrict $\mathcal{B}_k\subset\mathcal{U}_k$. A formal proof is provided in Appendix 8.B.

In this section, we consider the independent uniform prior on $\boldsymbol{\theta} = (\theta_{\mathrm{t}}, \theta_{\mathrm{r}})$, i.e.,

$$f_0(\boldsymbol{\theta}) = f_{\mathrm{r},0}(\theta_{\mathrm{r}}) \cdot f_{\mathrm{t},0}(\theta_{\mathrm{t}}), \;\; f_{\mathrm{x},0}(\theta_{\mathrm{x}}) = \frac{\chi_{\mathcal{U}_{\mathrm{x},0}}(\theta_{\mathrm{x}})}{|\mathcal{U}_{\mathrm{x},0}|}. \tag{2.41}$$

From Lemma 2.1, it directly follows that f_k is *uniform* in its support $\mathcal{U}_k$, and the state transition probabilities from state $(\mathcal{U}_k, D_k)$ under the beam-alignment action $(1, \mathcal{B}_k, 0) \in \mathcal{A}_{\mathrm{ext}}(\mathcal{U}, D)$, given in (2.29) for the general case, can be specialized as $D_{k+1} = D_k$ and

$$\mathcal{U}_{k+1} = \begin{cases} \mathcal{B}_k, & \text{w.p. } \frac{|\mathcal{B}_k|}{|\mathcal{U}_k|}, \\ \mathcal{U}_k \setminus \mathcal{B}_k, & \text{w.p. } 1 - \frac{|\mathcal{B}_k|}{|\mathcal{U}_k|}, \end{cases} \tag{2.42}$$

where "w.p." abbreviates "with probability". On the other hand, under the data communication action $(0, \mathcal{B}_k, R_k)$, the new state becomes $\mathcal{U}_{k+1} = \mathcal{U}_k$, and $D_{k+1} = D_k - R_k T$.

In order to determine the optimal policy with extended action set, we proceed as follows:

1. In Sec. 2.3.1, we find the structure of the optimal data communication beam, as a function of the transmit rate R_k and support $\mathcal{U}_k$, and investigate its energy cost;

2. Next, in Sec. 2.3.2, we prove that it is suboptimal to perform beam-alignment *after* data communication within the frame. Instead, it is convenient to narrow down the beam as much as possible via beam-alignment, to achieve the most energy-efficient data communication;

3. Finally, in Sec. 2.3.3, we investigate the structure of the value function, to prove the optimality of a *fixed-length* beam-alignment and of a *fractional-search method*.

2.3.1 Optimal data communication beam

In the following theorem, we find the optimal 2D beam for data communication.

Theorem 2.1. *In any communication slot $k \in \mathcal{I}_d$, the 2D beam $\mathcal{B}_k$ is optimal iff*

$$\mathcal{B}_k \subseteq \mathcal{U}_k \quad |\mathcal{B}_k| = \vartheta \, |\mathcal{U}_k| \,, \tag{2.43}$$

where $\vartheta = (1-\epsilon)/q^*$, *with* $q^* = \arg\max_{q\in[1-\epsilon,1]} q\bar{F}_{\gamma}^{-1}(q|\hat{\gamma})$.

Proof. The proof is provided in Appendix 8.C. □

The significance of this result is that the optimal beam in the data communication phase is a *fraction* ϑ of the region of uncertainty $\mathcal{U}_k$, with ϑ reflecting the desired outage constraint. By substituting (2.43) into (2.36), and letting

$$\phi_d(R,\epsilon) \triangleq \frac{\psi_d(R)(1-\epsilon)}{q^*\bar{F}_{\gamma}^{-1}(q^*|\hat{\gamma})} \tag{2.44}$$

be the energy/rad^2 to achieve transmission rate R with outage probability ϵ, the cost per stage of a data communication action with beam given by Theorem 2.1 can be expressed as

$$c(\mathbf{a}_k;\mathcal{U}_k,D_k) = \phi_d(R_k,\epsilon)\,|\mathcal{U}_k|\,. \tag{2.45}$$

2.3.2 Beam-alignment before data communication is optimal

In Theorem 2.2, we prove that it is suboptimal to precede data communication to beam-alignment. Instead, it is more energy efficient to narrow down the beam as much as possible via beam-alignment, before switching to data communication.

Theorem 2.2. *Let* μ *be a policy and* $\{(\mathcal{U}_k,D_k), k\in\mathcal{I}\}$ *be a realization of the state process under* μ *such that* $\exists j : \xi_j(\mathcal{U}_j,D_j)=0$ *and* $\xi_{j+1}(\mathcal{U}_{j+1},D_{j+1})=1$ *(beam-alignment is followed by data communication, for some slot* j*). Then,* μ *is suboptimal.*

Proof. The theorem is proved in two parts using contradiction. The first part deals with the case when a data communication slot is followed by a beam-alignment slot having non-zero beam-width. The second part deals with the case when a data communication slot is followed by a beam-alignment slot having zero beam-width. Let μ be a policy such that, for some state $(\mathcal{U}_j,D_j)$ and slot index j, $\mu_j(\mathcal{U}_j,D_j)=(0,\mathcal{B}_j,R_j)$, satisfying the conditions of Theorem 2.1 (data communication action); thus, the state at $j+1$ is $(\mathcal{U}_{j+1},D_{j+1})=(\mathcal{U}_j,D_j-TR_j)$. Further, assume that, in this state, $\mu_{j+1}(\mathcal{U}_j,D_j-TR_j)=(1,\mathcal{B}_{j+1},0)$ (beam-alignment), with $\mathcal{B}_{j+1}\subset\mathcal{U}_j$ (strict subset, see (2.40)), so that the state in slot $j+2$ is either $(\mathcal{B}_{j+1},D_j-TR_j)$

with probability $|\mathcal{B}_{j+1}|/|\mathcal{U}_j|$ (ACK), or $(\mathcal{U}_j \setminus \mathcal{B}_{j+1}, D_j - TR_j)$ otherwise (NACK). This policy follows beam-alignment to data communication, and we want to prove that it is suboptimal. We use (2.37) to get the cost-to-go function in slot j under policy μ as

$$\begin{aligned} V_j^{\mu}(\mathcal{U}_j, D_j) &= \phi_d(R_j, \epsilon)\,|\mathcal{U}_j| + V_{j+1}^{\mu}(\mathcal{U}_j, D_j - TR_j) \\ &= \phi_d(R_j, \epsilon)\,|\mathcal{U}_j| + \phi_s\,|\mathcal{B}_{j+1}| + \frac{|\mathcal{B}_{j+1}|}{|\mathcal{U}_j|} V_{j+2}^{\mu}(\mathcal{B}_{j+1}, D_j - TR_j) \\ &\quad + \frac{|\mathcal{U}_j \setminus \mathcal{B}_{j+1}|}{|\mathcal{U}_j|} V_{j+2}^{\mu}(\mathcal{U}_j \setminus \mathcal{B}_{j+1}, D_j - TR_j). \end{aligned} \tag{2.46}$$

We consider the two cases $|\mathcal{B}_{j+1}|>0$ and $|\mathcal{B}_{j+1}|=0$ separately. In both cases, we will construct a new policy $\tilde{\mu}$ and compare the cost-to-go function at j under the two policies μ and $\tilde{\mu}$.

$\underline{|\mathcal{B}_{j+1}|> 0}$: We define $\tilde{\mu}$ as being equal to μ except for the following: $\tilde{\mu}_j(\mathcal{U}_j, D_j)=(1, \mathcal{B}_{j+1}, 0)$, so that $\tilde{\mu}$ executes the beam-alignment action in slot j, instead of $j+1$. It follows that

$$V_j^{\tilde{\mu}}(\mathcal{U}_j, D_j) = \phi_s\,|\mathcal{B}_{j+1}| + \frac{|\mathcal{B}_{j+1}|}{|\mathcal{U}_j|} V_{j+1}^{\tilde{\mu}}(\mathcal{B}_{j+1}, D_j) + \frac{|\mathcal{U}_j \setminus \mathcal{B}_{j+1}|}{|\mathcal{U}_j|} V_{j+1}^{\tilde{\mu}}(\mathcal{U}_j \setminus \mathcal{B}_{j+1}, D_j). \tag{2.47}$$

Furthermore, we design $\tilde{\mu}$ such that $\tilde{\mu}_{j+1}(\mathcal{B}_{j+1}, D_j) = (0, \tilde{\mathcal{B}}'_{j+1}, R_j)$ and $\tilde{\mu}_{j+1}(\mathcal{U}_j \setminus \mathcal{B}_{j+1}, D_j) = (0, \tilde{\mathcal{B}}''_{j+1}, R_j)$, so that $\tilde{\mu}$ executes the data communication action in slot $j+1$, instead of j, with beams $\tilde{\mathcal{B}}'_{j+1}$ and $\tilde{\mathcal{B}}''_{j+1}$ satisfying the conditions of Theorem 2.1. It follows that the system moves from state $(\mathcal{B}_{j+1}, D_j)$ to $(\mathcal{B}_{j+1}, D_j - TR_j)$, and from $(\mathcal{U}_j \setminus \mathcal{B}_{j+1}, D_j)$ to $(\mathcal{U}_j \setminus \mathcal{B}_{j+1}, D_j - TR_j)$ under policy $\tilde{\mu}$, yielding

$$\begin{aligned} V_{j+1}^{\tilde{\mu}}(\mathcal{B}_{j+1}, D_j) &\overset{(a)}{=} \phi_d(R_j, \epsilon)\,|\mathcal{B}_{j+1}| + V_{j+2}^{\tilde{\mu}}(\mathcal{B}_{j+1}, D_j - TR_j), \\ V_{j+1}^{\tilde{\mu}}(\mathcal{U}_j \setminus \mathcal{B}_{j+1}, D_j) &\overset{(b)}{=} \phi_d(R_j, \epsilon)\,|\mathcal{U}_j \setminus \mathcal{B}_{j+1}| + V_{j+2}^{\tilde{\mu}}(\mathcal{U}_j \setminus \mathcal{B}_{j+1}, D_j - TR_j). \end{aligned} \tag{2.48}$$

By substituting (2.48)-(a),(b) into (2.47), and using the fact that $\tilde{\mu}_k$ and μ_k are identical for $k \geq j+2$ (hence $V_{j+2}^{\tilde{\mu}} = V_{j+2}^{\mu}$), it follows that

$$V_j^{\tilde{\mu}}(\mathcal{U}_j, D_j) - V_j^{\mu}(\mathcal{U}_j, D_j) \overset{(a)}{=} -\phi_d(R_j, \epsilon) 2 \frac{|\mathcal{B}_{j+1}|\,|\mathcal{U}_j \setminus \mathcal{B}_{j+1}|}{|\mathcal{U}_j|} \overset{(b)}{<} 0, \tag{2.49}$$

where (a) follows from $|\mathcal{U}_j \setminus \mathcal{B}_{j+1}| = |\mathcal{U}_j| - |\mathcal{B}_{j+1}|$; (b) follows from $|\mathcal{B}_{j+1}| > 0$ and $\mathcal{B}_{j+1} \subset \mathcal{U}_j$.

$\underline{|\mathcal{B}_{j+1}| = 0}$: In this case, we design $\tilde{\mu}$ equal to μ except for the following: $\tilde{\mu}_j(\mathcal{U}_j, D_j) = (0, \tilde{\mathcal{B}}'_j, R_j/2)$, with $\tilde{\mathcal{B}}'_j$ satisfying the conditions of Theorem 2.1, so that state $(\mathcal{U}_j, D_j)$ transitions to state $(\mathcal{U}_j, D_j - TR_j/2)$. Moreover $\tilde{\mu}_{j+1}(\mathcal{U}_j, D_j - TR_j/2) = (0, \tilde{\mathcal{B}}''_j, R_j/2)$, with $\tilde{\mathcal{B}}''_j$ satisfying the conditions of Theorem 2.1, so that the system moves to state $(\mathcal{U}_j, D_j - TR_j)$ in slot $j+2$. Under this new policy, the BS performs data communication in both slots, with rate $R_j/2$. Thus, the cost-to-go function under $\tilde{\mu}$ in slot j is given as

$$\begin{aligned} V_j^{\tilde{\mu}}(\mathcal{U}_j, D_j) &= \phi_d\left(\frac{R_j}{2}, \epsilon\right)|\mathcal{U}_j| + V_{j+1}^{\hat{\mu}}\left(\mathcal{U}_j, D_j - T\frac{R_j}{2}\right) \\ &= 2\phi_d\left(\frac{R_j}{2}, \epsilon\right)|\mathcal{U}_j| + V_{j+2}^{\hat{\mu}}\left(\mathcal{U}_j, D_j - TR_j\right). \end{aligned} \tag{2.50}$$

By comparing (2.50) and (2.46) and using the fact that μ and $\tilde{\mu}$ are identical for $k \geq j+2$, we get

$$V_j^{\tilde{\mu}}(\mathcal{U}_j, D_j) - V_j^{\mu}(\mathcal{U}_j, D_j) \overset{(a)}{=} \left[2\phi_d\left(\frac{R_j}{2}, \epsilon\right) - \phi_d\left(R_j, \epsilon\right)\right]|\mathcal{U}_j| \overset{(b)}{<} 0, \tag{2.51}$$

where (a) follows from $|\mathcal{B}_{j+1}|=0$; (b) follows from the strict convexity of $\phi_d(R, \epsilon)$ over $R>0$, implying that $2\phi_d\left(\frac{R_j}{2}, \epsilon\right) < \phi_d(R_j, \epsilon)$. (2.49) and (2.51) imply that μ does not satisfy Bellman's optimality equation, hence it is suboptimal, yielding a contradiction. The theorem is proved. □

From Theorem 2.2, we infer that:

Corollary 2.1. *Under an optimal policy μ^*, the frame can be split into a beam-alignment phase, followed by a data communication phase until the end of the frame. The duration $L^* \in \mathcal{I}$ of beam-alignment is, possibly, a random variable, function of the realization of the beam-alignment process.*

To capture this phase transition, we introduce the state variable $\nabla \in \{\text{BA}, \text{DC}\}$, denoting that the system is operating in the beam-alignment phase (∇=BA) or switched to data communication (∇=DC). The extended state is denoted as $(\mathcal{U}_k, D_k, \nabla_k)$, with the following

DP updates. If ∇_k=DC, then the system remains in the data communication phase until the end of the frame, and ∇_j=DC, $\forall j \geq k$, yielding

$$\hat{V}_k^*(\mathcal{U}_k, D_k, \mathrm{DC}) = \min_{0<R\leq D_k/T} \Big\{ \phi_d(R,\epsilon)\,|\mathcal{U}_k| + \hat{V}_{k+1}^*(\mathcal{U}_{k+1}, D_k - TR, \mathrm{DC}) \Big\}. \tag{2.52}$$

Using the convexity of $\phi_d(R,\epsilon)$ with respect to R, it is straightforward to prove the following.

Lemma 2.2. $\hat{V}_k^*(\mathcal{U}_k, D_k, \mathrm{DC}) = (N-k)\phi_d\left(\frac{D_k}{T(N-k)}, \epsilon\right)|\mathcal{U}_k|\,.$

That is, it is optimal to transmit with constant rate $\frac{D_k}{T(N-k)}$ in the remaining $(N-k)$ slots until the end of the frame. On the other hand, if $\nabla_k = \mathrm{BA}$, then $\nabla_j = \mathrm{BA}, \forall j \leq k$ and $D_k = D_0$, since no data has been transmitted yet. Then,

$$\begin{aligned}\hat{V}_k^*(\mathcal{U}_k, D_0, \mathrm{BA}) = \min\Bigg\{ & (N-k)\phi_d\left(\frac{NR_{\min}}{N-k}, \epsilon\right)|\mathcal{U}_k|\,, \\ & \min_{\mathcal{B}_k\subset\mathcal{U}_k} \phi_s\,|\mathcal{B}_k| + \frac{|\mathcal{B}_k|}{|\mathcal{U}_k|}\hat{V}_{k+1}^*(\mathcal{B}_k, D_0, \mathrm{BA}) + \left(1 - \frac{|\mathcal{B}_k|}{|\mathcal{U}_k|}\right)\hat{V}_{k+1}^*(\mathcal{U}_k \setminus \mathcal{B}_k, D_0, \mathrm{BA}) \Bigg\},\end{aligned} \tag{2.53}$$

where the outer minimization reflects an optimization over the actions "switch to data communication in slot k with rate $R_k = \frac{NR_{\min}}{N-k}$," or "perform beam-alignment." The inner minimization represents an optimization over the 2D beam $\mathcal{B}_k$ used for beam-alignment.

2.3.3 Optimality of deterministic beam-alignment duration with fractional-search method

It is important to observe that the proposed protocol is *interactive*, so that the duration of the beam-alignment phase, $L^* \in \mathcal{I}$, is possibly a random variable, function of the realization of the beam-alignment process. For example, if it occurs that the AoD/AoA is identified with high accuracy, the BS may decide to switch to data communication to achieve energy-efficient transmissions until the end of the frame. Although it may seem intuitive that L^* should indeed be random, in this section we will show that, instead, L^* is *deterministic*. Additionally, we prove the optimality of a *fractional search method*, which dictates the optimal beam design.

To unveil these structural properties, we define $v_k^*(\mathcal{U}_k) \triangleq \frac{\hat{V}_k^*(\mathcal{U}_k, D_0, \mathrm{BA})}{|\mathcal{U}_k|}$. Then, (2.53) yields

$$v_k^*(\mathcal{U}_k) = \min\Big\{(N-k)\phi_d\left(\frac{NR_{\min}}{N-k}, \epsilon\right), \min_{\rho\in[0,1)} \phi_s\rho + \rho^2 v_{k+1}^*(\mathcal{B}_{\mathrm{t},k}) + (1-\rho)^2 v_{k+1}^*(\mathcal{U}_k \setminus \mathcal{B}_k)\Big\}, \tag{2.54}$$

where $v_N^*(\mathcal{U}_N)=\infty$ and we used ρ in place of $\frac{|\mathcal{B}_k|}{|\mathcal{U}_k|}$, with $\rho<1$ since $\mathcal{B}_k \subset \mathcal{U}_k$. Using this fact, we find that $v_{N-1}^*(\mathcal{U}_{N-1})=\phi_d\left(NR_{\min}, \epsilon\right)$ is *independent* of $\mathcal{U}_{N-1}$. By induction on k, it is then straightforward to see that $v_k^*(\mathcal{U}_k)$ is *independent* of $\mathcal{U}_k, \forall k$. We thus let $v_k^* \triangleq v_k^*(\mathcal{U}_k), \forall \mathcal{U}_k$ to capture this independence, which is then defined recursively as

$$v_k^* = \min\Big\{(N-k)\phi_d\left(\frac{NR_{\min}}{N-k}, \epsilon\right), \min_{\rho\in[0,1)} \phi_s\rho + \left[\rho^2 + (1-\rho)^2\right] v_{k+1}^*\Big\}. \tag{2.55}$$

The value of ρ achieving the minimum in (2.55) is $\rho_k = \frac{|\mathcal{B}_k|}{|\mathcal{U}_k|} = \frac{1}{2}\left(1 - \frac{\phi_s}{2v_{k+1}^*}\right)^+$, yielding

$$v_k^* = \min\Bigg\{\underbrace{(N-k)\phi_d\left(\frac{NR_{\min}}{N-k}, \epsilon\right)}_{\Gamma_k \text{ (data communication)}}, \underbrace{v_{k+1}^* - \frac{[(2v_{k+1}^* - \phi_s)^+]^2}{8v_{k+1}^*}}_{\Lambda_k \text{ (beam-alignment)}}\Bigg\}.$$

From this decomposition, we infer important properties:

1. Given v_k^*, the original value function is obtained as $\hat{V}_k^*(\mathcal{U}_k, D_0, \mathrm{BA}) = v_k^* |\mathcal{U}_k|$. If, at time k, $\Gamma_k < \Lambda_k$, then it is optimal to switch to data communication in the remaining $N-k$ slots, with constant rate $\frac{NR_{\min}}{N-k}$.

2. Otherwise, it is optimal to perform beam-alignment, with beam $\mathcal{B}_k \subset \mathcal{U}_k$, $|\mathcal{B}_k| = \rho_k |\mathcal{U}_k|$.

3. Finally, since the time to switch to data communication is solely based on $\{v_k^*\}$, but not on $\mathcal{U}_k$, it follows that *fixed-length* beam-alignment is optimal, with duration

$$L^* = \min\{k : \Gamma_k < \Lambda_k\}. \tag{2.56}$$

These structural results are detailed in the following theorem.

Theorem 2.3. *Let*

$$L_{\min}= \underset{L\in\{0,\dots,N-1\}}{\arg\min} \left\{ L : (N-L)\phi_d\left(\frac{NR_{\min}}{N-L},\epsilon\right) > \frac{\phi_s}{2} \right\} \tag{2.57}$$

and, for $L_{\min} \le L < N$,

$$\begin{cases} v_L^{(L)} = (N-L)\phi_d\left(\frac{NR_{\min}}{N-L},\epsilon\right), \\ v_k^{(L)} = v_{k+1}^{(L)} - \frac{(2v_{k+1}^{(L)}-\phi_s)^2}{8v_{k+1}^{(L)}}, \ k < L. \end{cases} \tag{2.58}$$

Then, the beam-alignment phase has deterministic *duration*

$$L^* = \arg \min_{L\in\{0\}\cup\{L_{\min},\dots,N-1\}} v_0^{(L)}. \tag{2.59}$$

For $0 \le k < L^*$ *(beam-alignment phase),* $\mathcal{B}_k$ *is optimal iff*

$$\mathcal{B}_k \subset \mathcal{U}_k, \ \ |\mathcal{B}_k| = \rho_k |\mathcal{U}_k|, \tag{2.60}$$

where ρ_k *is the* fractional search *parameter, defined as*

$$\begin{cases} \rho_{L^*-1} = \frac{1}{2} - \frac{\phi_s}{4(N-L^*)\phi_d\left(\frac{NR_{\min}}{N-L^*},\epsilon\right)}, \\ \rho_k = \frac{1-\rho_{k+1}}{1-2\rho_{k+1}^2}\rho_{k+1}, \ k < L^*-1. \end{cases} \tag{2.61}$$

Moreover, $\rho_k \in (0,1/2)$*, strictly increasing in* k*. For* $k \ge L^*$*, the data communication phase occurs with rate* $\frac{NR_{\min}}{N-L^*}$*, and 2D beam given by Theorem 2.1.*

Proof. Since the optimal duration of the beam-alignment phase is deterministic, as previously discussed, we consider a fixed beam-alignment duration L, and then optimize over L to

achieve minimum energy consumption. Let $L \in \mathcal{I}$. Then, the DP updates are obtained by adapting (2.55) to this case (so that the outer minimization disappears for $k < L$), yielding

$$\begin{cases} v_L^{(L)} = (N-L)\phi_d\left(\frac{NR_{\min}}{N-L}, \epsilon\right), \\ v_k^{(L)} = g_k(\rho_k), \ k < L, \ \text{where} \\ g_k(\rho) \triangleq \phi_s \rho + \left[\rho^2 + (1-\rho)^2\right] v_{k+1}^{(L)}, \\ \rho_k = \arg\min_{\rho\in[0,1]} g_k(\rho) = \frac{1}{2}\left(1 - \frac{\phi_s}{2v_{k+1}^{(L)}}\right)^+. \end{cases} \tag{2.62}$$

Since the goal is to minimize the energy consumption, the optimal L is obtained by solving $L^* = \arg\min_L v_0^{(L)}$. We now prove that $0 < L < L_{\min}$ is suboptimal, so that this optimization can be restricted to $L \in \{0\} \cup \{L_{\min}, \cdots, N-1\}$, as in (2.59). Let $0 < L < L_{\min}$, so that $v_L^{(L)} \leq \phi_s/2$, as can be seen from the definition of $L_{\min}$ in (2.57). Note that $v_k^{(L)}$ is a non-decreasing function of k. In fact, $v_k^{(L)} \leq g_k(0) = v_{k+1}^{(L)}$. Then, it follows that $v_k^{(L)} \leq \phi_s/2, \forall k$, hence $\rho_k = 0, \forall k$, yielding $v_0^{(L)} = v_L^{(L)}$ by induction. However, $v_L^{(L)}$ is an increasing function of L (it is more energy efficient to spread transmissions over a longer interval), hence $v_0^{(L)} > v_0^{(0)}$ and such L is suboptimal. This proves that any $0 < L < L_{\min}$ is suboptimal.

We now prove the updates for $L \geq L_{\min}$, i.e., $v_L^{(L)} > \phi_s/2$. By induction, we have that $v_k^{(L)} > \phi_s/2, \forall k$. In fact, this condition trivially holds for $k = L$, by hypothesis. Now, assume $v_{k+1}^{(L)} > \phi_s/2$ for some $k < L$. Then, $v_k^{(L)} = \min_{\rho\in[0,1]} g_k(\rho)$, minimized at $\rho_k = \frac{1}{2}\left(1 - \frac{\phi_s}{2v_{k+1}^{(L)}}\right)$, so that $v_k^{(L)} = g_k(\rho_k)$, yielding (2.58). This recursion is an increasing function of $v_{k+1}^{(L)}$, yielding $v_k^{(L)} > \phi_s/2$, thus proving the induction. It follows that $\rho_k = \frac{1}{2}\left(1 - \frac{\phi_s}{2v_{k+1}^{(L)}}\right), \forall k$, yielding the recursion given by (2.58). The fractional search parameter ρ_k is finally obtained by substituting $v_{k+1}^{(L)} = \frac{\phi_s}{2(1-2\rho_k)}$ into the recursion (2.58) to find a recursive expression of ρ_k from ρ_{k+1}, yielding (2.61). These fractional values are used to obtain $\mathcal{B}_k$ in (2.60).

To conclude, we show by induction that $\rho_k \in (0, 1/2)$, strictly increasing in k. This is true for $k = L-1$ since $\rho_{L-1} \in (0, 1/2)$. Assume that $\rho_{k+1} \in (0, 1/2)$, for some $k \leq L-2$. Then, by inspection of (2.61), it follows that $0 < \rho_k < \rho_{k+1} < 1/2$. The theorem is thus proved. □

2.4 Decoupled BS and UE Beam-Alignment

In the previous section, we proved the optimality of a fractional search method, based on an extended action space that uses the 2D beam $\mathcal{B}_k \in [-\pi/2, \pi/2]^2$, which may take any shape. However, actual beams should satisfy the rectangular constraint $\mathcal{B}_k = \mathcal{B}_{\mathrm{t},k} \times \mathcal{B}_{\mathrm{r},k}$, and therefore, it is not immediate to see that the optimal scheme outlined in Theorem 2.3 is attainable in practice. Indeed, in this section we prove that there exists a feasible beam design attaining optimality. The proposed beam design decouples over time the beam-alignment of the AoD at the BS (*BS beam-alignment*) and of the AoA at the UE (*UE beam-alignment*). To explain this approach, we define the support of the marginal belief with respect to $\theta_{\mathrm{x}}, \mathrm{x} \in \{\mathrm{t}, \mathrm{r}\}$ as $\mathcal{U}_{\mathrm{x},k} \equiv \mathrm{supp}(f_{\mathrm{x},k})$. In *BS beam-alignment*, indicated with $\beta_k = 1$, the 2D beam is chosen as $\mathcal{B}_k = \mathcal{B}_{\mathrm{t},k} \times \mathcal{U}_{\mathrm{r},k}$, where $\mathcal{B}_{\mathrm{t},k} \subset \mathcal{U}_{\mathrm{t},k}$, so that the BS can better estimate the support of the AoD, whereas the UE receives over the entire support of the AoA. On the other hand, in *UE beam-alignment*, indicated with $\beta_k = 2$, the 2D beam is chosen as $\mathcal{B}_k = \mathcal{U}_{\mathrm{t},k} \times \mathcal{B}_{\mathrm{r},k}$, where $\mathcal{B}_{\mathrm{r},k} \subset \mathcal{U}_{\mathrm{r},k}$, so that the UE can better estimate the support of the AoA, whereas the BS transmits over the entire support of the AoD. We now define a policy μ that uses this principle, and then prove its optimality.

Definition 2.1 (Decoupled fractional search policy)**.** *Let L^*, ϑ, $\{\rho_k : k = 0, \dots, L^* - 1\}$ as in Theorems 2.1, 2.3. In slots $k = L^*, \dots, N$, data communication occurs with rate $R_k = \frac{N R_{\min}}{N - L^*}$ and beams*

$$\mathcal{B}_{\mathrm{t},k} \subseteq \mathcal{U}_{\mathrm{t},k}, \;\; \mathcal{B}_{\mathrm{r},k} \subseteq \mathcal{U}_{\mathrm{r},k}, \;\; |\mathcal{B}_{\mathrm{t},k}| \, |\mathcal{B}_{\mathrm{r},k}| = \vartheta \, |\mathcal{U}_{\mathrm{t},k}| \, |\mathcal{U}_{\mathrm{r},k}| \, . \tag{2.63}$$

In slots $k = 0, 1, \dots, L^*$, $\beta_k \in \{1, 2\}$ is chosen arbitrarily and beam-alignment occurs with beams

$$\begin{cases} \mathcal{B}_{\mathrm{t},k} \subset \mathcal{U}_{\mathrm{t},k}, \;\; \mathcal{B}_{\mathrm{r},k} = \mathcal{U}_{\mathrm{r},k}, \;\; |\mathcal{B}_{\mathrm{t},k}| = \rho_k \, |\mathcal{U}_{\mathrm{t},k}| \, , \text{ if } \beta_k = 1 \\ \mathcal{B}_{\mathrm{t},k} = \mathcal{U}_{\mathrm{t},k}, \;\; \mathcal{B}_{\mathrm{r},k} \subset \mathcal{U}_{\mathrm{r},k}, \;\; |\mathcal{B}_{\mathrm{r},k}| = \rho_k \, |\mathcal{U}_{\mathrm{r},k}| \, , \text{ if } \beta_k = 2. \end{cases} \tag{2.64}$$

Theorem 2.4. *The decoupled fractional search policy is optimal, with minimum power consumption*

$$\bar{P}_{\mathrm{u}} = \frac{v_0^{(L^*)}}{T_{\mathrm{fr}}} \left|\mathcal{U}_0\right|. \tag{2.65}$$

Proof. The proof is provided in Appendix 8.E. □

The intuition behind this result is that, by decoupling the beam-alignment of the AoD and AoA over time, the proposed method maintains a rectangular support $\mathcal{U}_k = \mathcal{U}_{\mathrm{t},k} \times \mathcal{U}_{\mathrm{r},k}$, so that no loss of optimality is incurred by using a rectangular beam $\mathcal{B}_k = \mathcal{B}_{\mathrm{t},k} \times \mathcal{B}_{\mathrm{r},k}$. Additionally, we can infer that the *exhaustive search* method is suboptimal, since it searches over the AoD/AoA space in an exhaustive manner, rather than by decoupling this search over time.

2.5 Non-Uniform Prior

In this section, we investigate the case of non-uniform prior f_0. We use the previous analysis to design a heuristic scheme with performance guarantees. We consider the decoupled fractional search policy (Definition 2.1), with the following additional constraints: in the beam-alignment phase $k < L^*$, if $\beta_k^* = 1$ (BS beam-alignment), then

$$\mathcal{B}_{\mathrm{t},k}^* = \arg \max_{\mathcal{B}_{\mathrm{t},k} \subset \mathcal{U}_{\mathrm{t},k}} \int_{\mathcal{B}_{\mathrm{t},k}} f_{\mathrm{t},k}(\theta_{\mathrm{t}}) \mathrm{d}\theta_{\mathrm{t}}, \text{ s.t. } \left|\mathcal{B}_{\mathrm{t},k}\right| = \rho_k \left|\mathcal{U}_{\mathrm{t},k}\right|; \tag{2.66}$$

if $\beta_k^* = 2$ (UE beam-alignment), then

$$\mathcal{B}_{\mathrm{r},k}^* = \arg \max_{\mathcal{B}_{\mathrm{r},k} \subset \mathcal{U}_{\mathrm{r},k}} \int_{\mathcal{B}_{\mathrm{r},k}} f_{\mathrm{r},k}(\theta_{\mathrm{r}}) \mathrm{d}\theta_{\mathrm{r}}, \text{ s.t. } \left|\mathcal{B}_{\mathrm{r},k}\right| = \rho_k \left|\mathcal{U}_{\mathrm{r},k}\right|. \tag{2.67}$$

Hence, the probability of ACK can be bounded as

$$\left.\begin{array}{ll} \text{Case } \beta_k^* = 1: & \int_{\mathcal{B}_{\mathrm{t},k}^*} f_{\mathrm{t},k}(\theta_{\mathrm{t}}) \mathrm{d}\theta_{\mathrm{t}} \geq \dfrac{\left|\mathcal{B}_{\mathrm{t},k}^*\right|}{\left|\mathcal{U}_{\mathrm{t},k}\right|} \\ \text{Case } \beta_k^* = 2: & \int_{\mathcal{B}_{\mathrm{r},k}^*} f_{\mathrm{r},k}(\theta_{\mathrm{r}}) \mathrm{d}\theta_{\mathrm{r}} \geq \dfrac{\left|\mathcal{B}_{\mathrm{r},k}^*\right|}{\left|\mathcal{U}_{\mathrm{r},k}\right|} \end{array}\right\} = \rho_k. \tag{2.68}$$

In other words, such choice of the BS-UE beam maximizes the probability of successful beam-detection, so that the resulting probability of ACK is at least as good as in the uniform case.

Similarly, in the data communication phase $k \geq L^*$, the BS transmits with rate $R_k = \frac{NR_{\min}}{N-L^*}$, and the beams are chosen as in Definition 2.1, with the additional constraint

$$(\mathcal{B}^*_{\mathrm{t},k}, \mathcal{B}^*_{\mathrm{r},k}) = \arg \max_{\mathcal{B}_{\mathrm{t},k}\times\mathcal{B}_{\mathrm{r},k}\subseteq\mathcal{U}_k} \int_{\mathcal{B}_{\mathrm{t},k}\times\mathcal{B}_{\mathrm{r},k}} f_k(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \quad \text{s.t.} \quad |\mathcal{B}_{\mathrm{t},k}|\,|\mathcal{B}_{\mathrm{t},k}| = \vartheta|\mathcal{U}_{\mathrm{t},k}|\,|\mathcal{U}_{\mathrm{r},k}|. \tag{2.69}$$

Under this choice, the energy consumption per data communication slot is obtained from (2.36),

$$E_k = \psi_d(R_k)\frac{|\mathcal{B}_k|}{\bar{F}_\gamma^{-1}\left(\frac{1-\epsilon}{\mathbb{P}(\boldsymbol{\theta}\in\mathcal{B}_k|\mathcal{U}_k)}\right)} \tag{2.70}$$

$$\overset{(a)}{\leq} \psi_d(R_k)\frac{|\mathcal{B}_k|}{\bar{F}_\gamma^{-1}\left(\frac{(1-\epsilon)|\mathcal{U}_k|}{|\mathcal{B}_k|}\right)} \overset{(b)}{=} \phi_d(R_k,\epsilon)\,|\mathcal{U}_k|\,, \tag{2.71}$$

where (a) follows from $\mathbb{P}(\boldsymbol{\theta}\in\mathcal{B}_k|\mathcal{U}_k)\geq|\mathcal{B}_k|/|\mathcal{U}_k|$, and (b) from $|\mathcal{B}_{\mathrm{t},k}|\,|\mathcal{B}_{\mathrm{r},k}| = \vartheta\,|\mathcal{U}_{\mathrm{t},k}|\,|\mathcal{U}_{\mathrm{r},k}|$, and from (2.44) with $\vartheta = (1-\epsilon)/q^*$ (Theorem 2.1). This result implies that data communication is more energy efficient than in the uniform case, see (2.45). These observations suggest that the uniform prior yields the worst performance, as confirmed by the following theorem.

Theorem 2.5. *The minimum power consumption for the non-uniform prior is upper bounded by $\bar{P}_{\mathrm{nu}} \leq \bar{P}_{\mathrm{u}}$, with equality when f_0 is uniform.*

Proof. We denote the value function of the non-uniform case under such policy as $V_{\mathrm{nu},k}(\mathcal{U}_k, D_k)$. Additionally, we let $\bar{P}_{\mathrm{nu}}$ be the corresponding minimum power consumption, solution of problem P_2 in (2.35), to distinguish it from the minimum power consumption in the uniform case, given by (2.65). For $k = L^*$ (data communication begins), (2.70) implies that

$$V_{\mathrm{nu},k}(\mathcal{U}_k, D_0) \leq (N-L^*)\phi_d\left(\frac{NR_{\min}}{N-L^*},\epsilon\right)|\mathcal{U}_k|\,. \tag{2.72}$$

For $k < L^*$ (beam-alignment phase), it can be expressed as

$$V_{\mathrm{nu},k}(\mathcal{U}_k, D_0)=\phi_s \left|\mathcal{B}_k^*\right| + \int_{\mathcal{B}_k^*} f_k(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} V_{\mathrm{nu},k+1}(\mathcal{B}_k^*, D_0) + \left(1- \int_{\mathcal{B}_k^*} f_k(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\right) V_{\mathrm{nu},k+1}(\mathcal{U}_k \setminus \mathcal{B}_k^*, D_0), \tag{2.73}$$

where $\mathcal{B}_k^*$ is given by (2.66) or (2.67). The minimum power consumption is given by $\bar{P}_{\mathrm{nu}}=V_{\mathrm{nu},0}(\mathcal{U}_0, D_0)/T_{\mathrm{fr}}$, so that $\bar{P}_{\mathrm{nu}} \leq \bar{P}_{\mathrm{u}}$ is equivalent to $V_{\mathrm{nu},k}(\mathcal{U}_k, D_k) \leq v_k^{(L^*)} \left|\mathcal{U}_k\right|$ when $k=0$. We prove this inequality for general k by induction. The induction hypothesis holds for $k=L^*$, see (2.72) with $v_{L^*}^{(L^*)}$ given in (2.58). Assume it holds for $k+1$, where $k \leq L^*-1$. Then, (2.73) can be expressed as

$$\begin{aligned} V_{\mathrm{nu},k}(\mathcal{U}_k, D_0) &\leq \phi_s \left|\mathcal{B}_k^*\right| + \int_{\mathcal{B}_k^*} f_k(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} v_{k+1}^{(L^*)} \left|\mathcal{B}_k^*\right| + \left(1- \int_{\mathcal{B}_k^*} f_k(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\right) v_{k+1}^{(L^*)} \left|\mathcal{U}_k \setminus \mathcal{B}_k^*\right| \\ &\overset{(a)}{=} \left[\phi_s\rho_k + v_{k+1}^{(L^*)}\left(1-2\rho_k+2\rho_k^2\right)\right]\left|\mathcal{U}_k\right| - \left(\int_{\mathcal{B}_k^*} f_k(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} - \rho_k\right) v_{k+1}^{(L^*)} \left|\mathcal{U}_k\right| \left(1-2\rho_k\right), \end{aligned}$$

where (a) follows from (2.66)-(2.67) and $\left|\mathcal{U}_k \setminus \mathcal{B}_k^*\right| = \left|\mathcal{U}_k\right| - \left|\mathcal{B}_k^*\right|$. Finally, the bound (2.68) yields

$$V_{\mathrm{nu},k}(\mathcal{U}_k, D_0) \leq \left[\phi_s\rho_k + v_{k+1}^{(L^*)}\left(1-2\rho_k+2\rho_k^2\right)\right]\left|\mathcal{U}_k\right| = v_k^{(L^*)} \left|\mathcal{U}_k\right|,$$

where the last equality is obtained by using the recursion (2.58) and the fact that $\rho_k = \frac{1}{2} - \frac{\phi_s}{4v_{k+1}^{(L^*)}}$ (see proof of Theorem 2.3). This proves the induction step. Clearly, equality is attained in the uniform case. The theorem is thus proved. □

This result is in line with the fact that one can leverage the structure of the joint distribution over $\boldsymbol{\theta}$ to improve the beam-alignment algorithm. However, for the first time to the best of our knowledge, this result provides a heuristic scheme with provable performance guarantees.

2.6 Impact of False-alarm and Misdetection

In this section, we analyze the impact of false-alarm and misdetection on the performance of the decoupled fractional search policy (Definition 2.1). For simplicity, we focus only on the uniform prior case. Under false-alarm and misdetection, the MDP introduced in Sec. 2.2 does not follow the Markov property. To overcome this problem, we augment it with the state variable $e_k \in \{0,1\}$, with $e_k = 0$ iff no errors have been introduced up to slot k. Note that, if errors have been introduced ($e_k = 1$), then necessarily $\boldsymbol{\theta} \notin \mathcal{U}_k$, so that we can write $e_k = 1 - \chi(\boldsymbol{\theta} \in \mathcal{U}_k)$. It should be noted that e_k is not observable in reality and is considered for the purpose of analysis only (indeed, the policy under analysis does not use such information). We thus define the state as $(\mathcal{U}_k, e_k)$,[8] and study the transition probabilities during the beam-alignment phase $k < L^*$. From state $(\mathcal{U}_k, 0)$ (no errors have been introduced), the transitions are

$$(\mathcal{U}_{k+1}, e_{k+1}) = \begin{cases} (\mathcal{B}_k, 0), & \text{w.p. } \rho_k(1-p_{\text{md}}) \\ (\mathcal{B}_k, 1), & \text{w.p. } (1-\rho_k)\, p_{\text{fa}} \\ (\mathcal{U}_k \setminus \mathcal{B}_k, 0), & \text{w.p. } (1-\rho_k)\,(1-p_{\text{fa}}) \\ (\mathcal{U}_k \setminus \mathcal{B}_k, 1), & \text{w.p. } \rho_k\, p_{\text{md}}, \end{cases} \tag{2.74}$$

where p_{fa} and p_{md} denote the false-alarm and misdetection probabilities, respectively. In fact, if no errors occur, then $\boldsymbol{\theta} \in \mathcal{B}_k$ with probability $\frac{|\mathcal{B}_k|}{|\mathcal{U}_k|} = \rho_k$ and $\boldsymbol{\theta} \notin \mathcal{B}_k$ otherwise, yielding the first and third cases; if a false-alarm or misdetection error is introduced, then the BS infers incorrectly that $\boldsymbol{\theta} \in \mathcal{B}_k$ (second case) or $\boldsymbol{\theta} \notin \mathcal{B}_k$ (fourth case), respectively, and the new state becomes $e_{k+1} = 1$. Once errors have been introduced (state $(\mathcal{U}_k, 1)$), it follows that $\boldsymbol{\theta} \notin \mathcal{B}_k$, so that $\mathcal{U}_{k+1} = \mathcal{B}_k$ iff a false-alarm error occurs, and the transitions are

$$(\mathcal{U}_{k+1}, e_{k+1}) = \begin{cases} (\mathcal{B}_k, 1), & \text{w.p. } p_{\text{fa}} \\ (\mathcal{U}_k \setminus \mathcal{B}_k, 1), & \text{w.p. } 1 - p_{\text{fa}}. \end{cases} \tag{2.75}$$

[8] The backlog D_k is removed from the state space, since no data is transmitted during the beam-alignment phase.

The average throughput and power are given by

$$\bar{T}_{\rm err} = \mathbb{E}\left[(1-e_{L^*})(1-\epsilon)R_{\min}|\mathcal{U}_0, e_0=0\right],$$
$$\bar{P}_{\rm err} = \frac{1}{T_{\rm fr}}\mathbb{E}\left[\phi_s \sum_{k=0}^{L^*-1}\rho_k\left|\mathcal{U}_k\right| + (N-L^*)\phi_d\left(\frac{NR_{\min}}{N-L^*},\epsilon\right)\left|\mathcal{U}_{L^*}\right|\Big|\mathcal{U}_0, e_0=0\right]. \tag{2.76}$$

In fact, a rate equal to $R_{\min}$ is sustained if: (1) no outage occurs in the data communication phase, with probability $1-\epsilon$; (2) no errors occur during the beam-alignment phase, $e_{L^*}=0$.

The analysis of the underlying Markov chain $\{(\mathcal{U}_k, e_k),\, k\geq 0\}$ yields the following theorem.

Theorem 2.6. *Under the decoupled fractional search policy,*

$$\bar{T}_{\rm err} = (1-\epsilon)R_{\min}\prod_{k=0}^{L^*-1}\Big[(1-\rho_k)(1-p_{\rm fa}) + \rho_k(1-p_{\rm md})\Big], \tag{2.77}$$
$$\bar{P}_{\rm err} = \bar{P}_{\rm u} + \frac{h_0+u_0}{T_{\rm fr}}\left|\mathcal{U}_0\right|, \tag{2.78}$$

where $\bar{P}_{\rm u}$ in (2.65) is the error-free case, and we have defined $h_{L^*}=u_{L^*}=0$ and, for $k<L^*$,

$$h_k = \phi_s\frac{\rho_k - p_{\rm fa}}{2} + \left[\rho_k p_{\rm fa} + (1-\rho_k)(1-p_{\rm fa})\right]h_{k+1}, \tag{2.79}$$
$$u_k = \Big[\rho_k^2(1-p_{\rm md}) + (1-\rho_k)^2(1-p_{\rm fa})\Big]u_{k+1} - (1-p_{\rm fa}-p_{\rm md})\rho_k\left[\frac{\phi_s}{2} + h_{k+1}(1-2\rho_k)\right]. \tag{2.80}$$

Proof. The proof is provided in Appendix 8.E. □

2.7 Numerical Results

In this section, we demonstrate the performance of the proposed *decoupled fractional search* (DFS) scheme and compare it with the *bisection search* algorithm developed in [13] and two variants of *exhaustive search*. In the bisection algorithm [13] (BiS), in each beam-alignment slot the uncertainty region is divided into two regions of equal width, scanned in sequence by the BS by transmitting beacons corresponding to each region. Then, the UE compares the signal power (the strongest indicating alignment) and transmits the feedback to the BS. Since in each beam-alignment slot two sectors are scanned (each of duration T_B), the total duration of the beam-alignment phase is $(2T_B+T_F)L$ [s], where T_F is the feedback

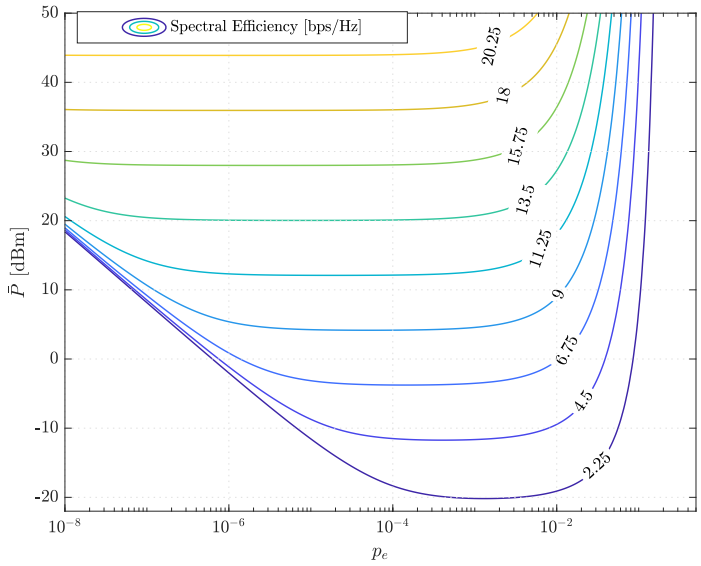

Figure 2.2. Spectral efficiency versus beam-alignment error probability p_e for DFS.

time. In *conventional exhaustive search* (CES), the BS-UE scan exhaustively the entire beam space. In the BS beam-alignment sub-phase, the BS searches over $N_B^{(BS)}$ beams covering the AoD space, while the UE receives isotropically; in the second UE beam-alignment sub-phase, the BS transmits using the best beam found in the first sub-phase, whereas the UE searches exhaustively over $N_B^{(UE)}$ beams covering the AoA space. Since the UE reports the best beam at the end of each sub-phase, the total duration of the beam-alignment phase is $[N_B^{(BS)} + N_B^{(UE)}]T_B + 2T_F$. On the other hand, in the *interactive exhaustive search* (IES) method, the UE reports the feedback at the end of each beam-alignment slot, and each beam-alignment sub-phase terminates upon receiving an ACK from the UE. Since the BS awaits for feedback at the end of each beam, the duration of the beam-alignment phase is $(T_B + T_F)[\hat{N}_B^{(BS)} + \hat{N}_B^{(UE)}]$, where $\hat{N}_B \leq N_B$ is the number of beams scanned until receiving an ACK; assuming the AoD/AoA is uniformly distributed over the beam space, the expected duration of the beam-alignment phase is then $\frac{1}{2}(T_B + T_F)\left[N_B^{(BS)} + N_B^{(UE)} + 2\right]$.

We use the following parameters: [carrier frequency]= 30GHz, d = 10m, [path loss exponent]= 2, T_{fr}=20ms, T_B=50μs, T_F=50μs, $|\mathcal{U}_0|$=$[\pi]^2$, N_0= $-$ 173dBm, W_{tot}=500MHz, M_{t}=M_{r}=128.

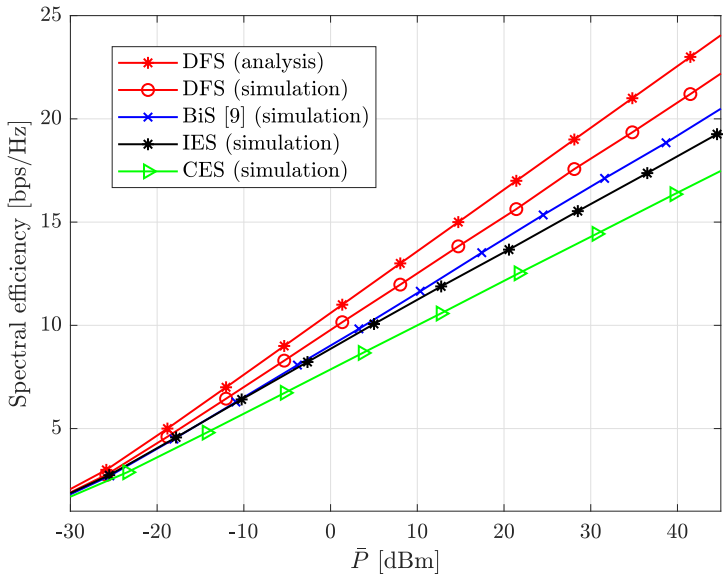

Figure 2.3. Spectral efficiency versus average power consumption.

In Fig. 2.2, we depict the average power vs the probability of false-alarm and misdetection p_e for different values of the spectral efficiency using expressions (2.77) and (2.78). We use $\epsilon = 0.01$, and consider Rayleigh fading with no CSI at BS, corresponding to $h{\sim}\mathcal{CN}(0, 1/\ell(d))$ with $\hat{h}{=}0$ and $\sigma_e^2{=}1/\ell(d)$. We restrict the optimization of L over $L{\in}\{0, \ldots, L_{\max}\}$, to capture a maximum resolution constraint for the antenna array, where we chose $L_{\max} = 14$. From the figure, we observe that, for a given p_e, as the spectral efficiency increases so does the average power consumption due to increase in the energy cost of data communication. Moreover, the figure reveals that, for a given value of spectral efficiency, there exists an optimal range of p_e, where power consumption is minimized. The performance degrades for p_e above the optimal range due to false-alarm and misdetection errors during beam-alignment, causing outage in data communication; similarly, it degrades for p_e below the optimal range due to an increased power consumption of beam-alignment.

In Fig. 2.3, we plot the results of a Monte-Carlo simulation with analog beams generated using the algorithm in [27]. In this case, we obtain $\phi_s{=}-94$dBm with $p_{\text{fa}}{=}p_{\text{md}}{=}10^{-5}$. For BiS and DFS we set $L_{\max} = 10$ to capture a maximum resolution constraint for the antenna array; for the exhaustive search methods, we choose $N_B^{(BS)} = N_B^{(UE)}{=}32$. The performance gap between the analytical and the simulation-based curves for DFS is attributed to the fact

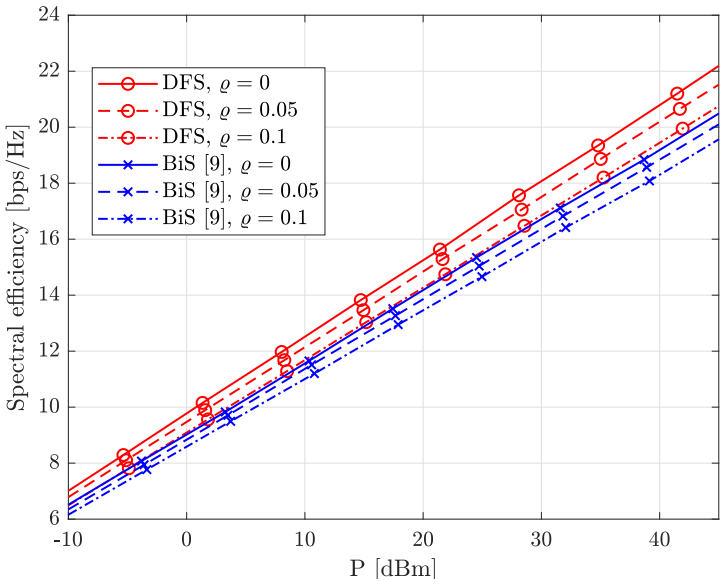

Figure 2.4. Performance degradation with multi-cluster channel ($K = 2$).

that the beams used in the simulation have non-zero side-lobe gain and non-uniform main-lobe gain, as opposed to the "sectored" beams used in the analytical model. This results in false-alarm, misdetection errors, and leakage, which lead to some performance degradation. *However, the simulation is in line with the analytical curve, and exhibits superior performance compared to the other schemes, thus demonstrating that the analysis using the sectored gain model provides useful insights for practical design.* For instance, to achieve a spectral efficiency of 15bps/Hz, BiS [13] requires 4dB more average power than DFS, mainly due to the time and energy overhead of scanning two sectors in each beam-alignment slot, whereas IES and CES require 7.5dB and 14dB more power, respectively. The performance degradation of IES and CES is due to the exhaustive search of the best sector, which demands a huge time overhead. Indeed, IES outperforms CES since it stops beam-alignment once a strong beam is detected.

So far in our analysis, we assumed a channel with a single cluster of rays, see (2.3). In Fig. 2.4, we depict the performance of DFS and BiS [13] in a multi-cluster channel ($K = 2$ in (2.2)), with the weakest cluster having a fraction ϱ of the total energy, $0 \leq \varrho \leq 0.1$. It can be seen that the performance of both DFS and BiS degrade as ϱ increases, since a portion of the energy is lost in the weaker cluster, and the algorithms may misdetect the weaker

cluster instead of the strongest one. For example, for spectral efficiency of 15bps/Hz, both schemes exhibit $\sim$ 2dB and $\sim$ 5dB performance loss at $\varrho = 5\%$ and $\varrho = 10\%$, respectively, compared to $\varrho = 0$ (single cluster). However, DFS consistently outperforms BiS, with a gain of $\sim$ 3.5dB. This evaluation demonstrates the robustness of the proposed algorithm in multi-cluster scenarios.

3. CODED ENERGY-EFFICIENT BEAM-ALIGNMENT

s This chapter proposes a coded energy-efficient beam-alignment scheme, robust against detection errors. Specifically, the beam-alignment sequence is designed such that the error-free feedback sequences are generated from a codebook with the desired error correction capabilities. Therefore, in the presence of detection errors, the error-free feedback sequences can be recovered with high probability. The assignment of beams to codewords is designed to optimize energy efficiency, and a water-filling solution is proved. The numerical results with analog beams depict up to 4dB and 8dB gains over exhaustive and uncoded beam-alignment schemes, respectively.

3.1 System Model

We consider a mm-wave cellular network with a single base-station (BS) and M user-ends (UEs) denoted as $\mathrm{UE}_i, i = 1, 2, \ldots, M$, in a downlink scenario. UE_i is at distance $d_i \leq d_{\max}$ from BS, where $d_{\max} > 0$ is the coverage radius of the BS. We assume that there is a single strongest path between the BS and each UE_i, whose angle of departure (AoD) and angle of arrival (AoA) are denoted by $\theta_{\mathrm{t},i} \sim \mathcal{U}[-\pi/2, \pi/2]$ and $\theta_{\mathrm{r},i} \sim \mathcal{U}[-\pi/2, \pi/2]$, respectively. $\mathcal{U}[a, b]$ denotes the uniform distribution over the interval $[a, b]$. We use the *sectored antenna* model to approximate the beam patterns of the BS and UEs [29]. Under such model, the beamforming gain is characterized by the angular support of the BS and UE beams, denoted as $\mathcal{B}_{\mathrm{t},k} \subseteq [-\pi/2, \pi/2]$ and $\mathcal{B}_{\mathrm{r},k} \subseteq [-\pi/2, \pi/2]$, respectively, and is given by

$$G(\mathcal{B}_k, \boldsymbol{\theta}_i) = \frac{(2\pi)^2}{|\mathcal{B}_k|} \chi(\boldsymbol{\theta}_i \in \mathcal{B}_k), \tag{3.1}$$

where $\mathcal{B}_k \equiv \mathcal{B}_{\mathrm{t},k} \times \mathcal{B}_{\mathrm{r},k}$ and $\boldsymbol{\theta}_i \triangleq (\theta_{\mathrm{t},i}, \theta_{\mathrm{r},i})$; $\chi(\boldsymbol{\theta} \in \mathcal{A})$ is the indicator function of the set $\mathcal{A}$, and $|\mathcal{A}| \triangleq \int_{\mathcal{A}} \mathrm{d}\boldsymbol{\theta}$ is its Lebesgue measure. In other words, if the AoD/AoA $\boldsymbol{\theta}$ lies in the beam

↑A version of this chapter was previously published by Allerton 2018 [5][DOI:10.1109/ALLERTON.2018.8635944]

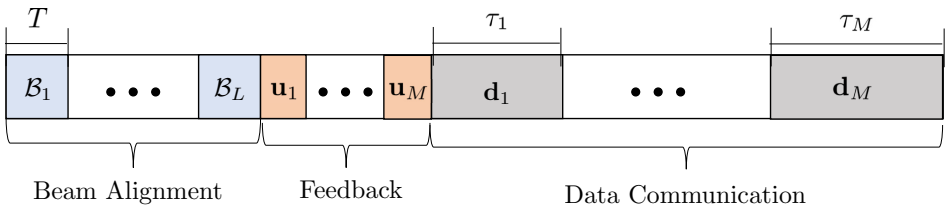

Figure 3.1. Timing Diagram.

support $\mathcal{B}_k$ of the BS and UE, then the signal is received with gain $\frac{(2\pi)^2}{|\mathcal{B}_k|}$; otherwise, only noise is received. The received signal at UE$_i$ can thus be expressed as

$$\boldsymbol{y}_k^{(i)} = h_k^{(i)} \sqrt{P_k G(\mathcal{B}_k, \boldsymbol{\theta}_i)} \boldsymbol{s}_k + \boldsymbol{n}_k^{(i)}, \tag{3.2}$$

where k is the slot index, $\boldsymbol{s}_k$ is the transmitted sequence, P_k is the transmission power of the BS, $h_k^{(i)}$ is the complex channel gain between the BS and UE$_i$, and $\boldsymbol{n}_{k,i} \sim \mathcal{CN}(\mathbf{0}, N_0 W_{\text{tot}} \boldsymbol{I})$ is complex additive white Gaussian noise (AWGN). The quantity N_0 denotes the one-sided power spectral density of the AWGN channel and W_{tot} is the system bandwidth. We assume Rayleigh fading channels $h_k^{(i)} \sim \mathcal{CN}(0, 1/\ell(d_i)), \forall i, k$, independent across UEs and i.i.d over slots, where $\ell(d_i)$ is the path loss between the BS and UE$_i$.

We consider a time-slotted system where the frame duration $T_{\text{fr}}[s]$ is divided into three phases: beam-alignment, feedback and data communication, of duration T_s, T_{fb} and T_d, respectively, with T_s+T_{fb}+T_d=T_{fr}, as depicted in Fig. 3.1. Data transmission is orthogonalized across users according to a TDMA strategy. We now describe these phases in more detail.

Beam-Alignment Protocol: The beam-alignment phase, of duration T_s, is divided into L slots, each of duration $T = T_s/L$, indexed by the set $\mathcal{I}_s = \{1, \ldots, L\}$. In each beam-alignment slot, the BS sends a pilot sequence $\boldsymbol{s}_k$ using the sequence of beams $\{\mathcal{B}_{\text{t},k}, k{=}1, \ldots, L\}$. Simultaneously, each UE receives using the sequence of beams $\{\mathcal{B}_{\text{r},k}, k{=}1, \ldots, L\}$. In each

beam-alignment slot, UE$_i$ tests whether $\boldsymbol{\theta}_i \in \mathcal{B}_k$ (alignment) or $\boldsymbol{\theta}_i \notin \mathcal{B}_k$ (mis-alignment). This can be expressed as the following hypothesis testing problem:

$$\begin{aligned} \mathcal{H}_1: \ \boldsymbol{y}_k^{(i)} &= h_k^{(i)} \sqrt{\frac{(2\pi)^2 P_k}{|\mathcal{B}_k|}} \boldsymbol{s}_k + \boldsymbol{n}_k^{(i)}, \ \text{(alignment)}, \\ \mathcal{H}_0: \ \boldsymbol{y}_k^{(i)} &= \boldsymbol{n}_k^{(i)}, \ \text{(mis-alignment)}. \end{aligned} \tag{3.3}$$

Under no CSI ($h_k^{(i)}$ unknown), the optimal Neyman-Pearson detector for the above binary problem is the threshold detector

$$\frac{|\boldsymbol{s}_k^H \boldsymbol{y}_k^{(i)}|^2}{N_0 W_{\text{tot}} \|\boldsymbol{s}_k\|_2^2} \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\lessgtr}} \tau_{\text{th}}. \tag{3.4}$$

If UE$_i$ infers that $\mathcal{H}_1$ is true, then it generates $u_k^{(i)}=1$, otherwise $u_k^{(i)}=0$. Each UE generates its detection sequence $\boldsymbol{u}_i \triangleq (u_1^{(i)}, u_2^{(i)}, \ldots, u_L^{(i)}) \in \{0,1\}^L$ with the above detector. This is used to infer the AoD/AoA $\boldsymbol{\theta}_i$, and to design the beams for the data communication phase, as detailed below.

$$\text{Let } \boldsymbol{c}_i \triangleq (c_1^{(i)}, c_2^{(i)}, \ldots, c_L^{(i)}) \text{ with } c_k^{(i)} = \chi(\boldsymbol{\theta}_i \in \mathcal{B}_k) \tag{3.5}$$

denote the error-free detection sequence. The detected $\boldsymbol{u}_i$ may incur mis-detection ($u_k^{(i)}=0$ but $c_k^{(i)}=1$) or false-alarm errors ($u_k^{(i)}=1$ but $c_k^{(i)}=0$) , with probabilities (these can be obtained from the signal model (3.3))

$$p_{\text{md},i} = 1 - \exp\left(-\frac{\tau_{\text{th}} |\mathcal{B}_k| N_0 W_{\text{tot}} \ell(d_i)}{|\mathcal{B}_k| N_0 W_{\text{tot}} \ell(d_i) + P_k (2\pi)^2 \|\boldsymbol{s}_k\|_2^2} \right), \tag{3.6}$$

$$p_{\text{fa},i} = \exp(-\tau_{\text{th}}). \tag{3.7}$$

The BS transmission power P_k and detector threshold τ_{th} are designed to guarantee maximum error probabilities $p_{\text{md},i}, p_{\text{fa},i} \leq p_e$ across users (this can be achieved via appropriate beam design, see [60]), which yields

$$\begin{aligned} \tau_{\text{th}} &= -\ln(p_e), \\ P_k &\geq \frac{N_0 W_{\text{tot}} \ell(d_i)}{(2\pi)^2 \|\boldsymbol{s}_k\|_2^2} \left[\frac{\ln(p_e)}{\ln(1-p_e)} - 1\right] |\mathcal{B}_k|, \ \forall i \in \{1, \ldots, M\}. \end{aligned} \tag{3.8}$$

Equivalently, we can express the energy $E_k \triangleq T_{\text{sy}} P_k \|\boldsymbol{s}_k\|^2$ as

$$E_k \geq \phi_s \, |\mathcal{B}_k| \tag{3.9}$$

where T_{sy} is the symbol duration; ϕ_s is the energy/rad^2 to guarantee the required detection performance among all UEs,

$$\phi_s \triangleq \frac{N_0 W_{\text{tot}} T_{\text{sy}}}{(2\pi)^2} \left[\frac{\ln(p_e)}{\ln(1-p_e)} - 1\right] \cdot \ell(d_{\max}). \tag{3.10}$$

In the rest of the chapter, we enforce equality in (3.9) for the purpose of energy-efficient beam-alignment design, and assume that $p_{\text{md},i} = p_{\text{fa},i} = p_e, \forall i$. Note that this is the worst-case scenario; in fact, in practice, an UE closer to the BS may experience a lower mis-detection probability $p_{\text{md},i} < p_e$ as a result of $\ell(d_i) < \ell(d_{\max})$, see (3.6).

With this notation, we write the detection sequence as

$$\boldsymbol{u}_i \triangleq \boldsymbol{c}_i \oplus \boldsymbol{e}_i, \tag{3.11}$$

where $\oplus$ denotes entry-wise modulo 2 addition, and $\boldsymbol{e}_i \in \{0,1\}^L$ is the beam-alignment error sequence of UE$_i$. Due to the i.i.d. Rayleigh fading assumption and to the fact that false-alarm and misdetection errors occur with probability $p_{\text{md},i} = p_{\text{fa},i} = p_e$, independently

across slots, it follows that $\boldsymbol{e}_i$ is independent of $\boldsymbol{c}_i$, and that errors are i.i.d. across UEs and slots, with probability mass function (pmf)

$$p(\boldsymbol{e}_i) = p_e^{W(\boldsymbol{e}_i)}(1-p_e)^{L-W(\boldsymbol{e}_i)}, \tag{3.12}$$

where $W(d) \triangleq \sum_{k=1}^{L} d_k$ is the Hamming weight of $\boldsymbol{d} \in \{0,1\}^L$.

We now design a coded beam-alignment strategy, robust to detection errors. If UE_i was provided with the error-free detection sequence $\boldsymbol{c}_i$, it could infer the support of $\boldsymbol{\theta}_i$ relative to the beam sequence $\{\mathcal{B}_k, k=1,\ldots,L\}$ to be

$$\boldsymbol{\theta}_i \in \mathcal{U}_{\boldsymbol{c}_i} \triangleq \cap_{k=1}^{L} \mathcal{B}_k^{c_k^{(i)}}, \tag{3.13}$$

where we have defined

$$\mathcal{B}_k^c = \begin{cases} \mathcal{B}_k & c=1, \\ [-\frac{\pi}{2}, \frac{\pi}{2}]^2 \setminus \mathcal{B}_k & c=0. \end{cases} \tag{3.14}$$

In fact, $c_k^{(i)}=1 \Leftrightarrow \boldsymbol{\theta}_i \in \mathcal{B}_k$ and $c_k^{(i)}=0 \Leftrightarrow \boldsymbol{\theta}_i \in [-\frac{\pi}{2}, \frac{\pi}{2}]^2 \setminus \mathcal{B}_k$, yielding (3.13) when considering the entire sequence $\boldsymbol{c}_i$. We let $\mathcal{C}$ be the set of all possible error-free detection sequences with non-empty beam support, i.e.

$$\mathcal{C} \triangleq \{\boldsymbol{c} \in \{0,1\}^L : \mathcal{U}_{\boldsymbol{c}} \neq \emptyset\}, \tag{3.15}$$

and $\mathcal{G}$ be the corresponding beam-support,

$$\mathcal{G} \triangleq \{\mathcal{U}_{\boldsymbol{c}} : \boldsymbol{c} \in \mathcal{C}\}. \tag{3.16}$$

Note that $(\mathcal{C}, \mathcal{G})$ are uniquely defined by the beam sequence $\{\mathcal{B}_k, k=1,\ldots,L\}$. Likewise, $\{\mathcal{B}_k, k=1,\ldots,L\}$ is uniquely defined by a specific choice of $(\mathcal{C}, \mathcal{G})$, as can be seen by letting

$$\mathcal{B}_k \equiv \cup_{\boldsymbol{c} \in \mathcal{C}: c_k=1} \mathcal{U}_{\boldsymbol{c}}, \ \mathcal{U}_{\boldsymbol{c}} \in \mathcal{G}. \tag{3.17}$$

Therefore, the problem of finding the optimal beam sequence, $\{\mathcal{B}_k, k = 1, \ldots, L\}$ is equivalent to that of finding the sets $\mathcal{C}$ and $\mathcal{G}$. However, a joint optimization over $\mathcal{C}$ and $\mathcal{G}$ is intractable due to the combinatorial nature of the problem and lack of convexity. Therefore, we resort to selecting $\mathcal{C}$ and $\mathcal{G}$ independently, where $\mathcal{C}$ is chosen from a predefined codebook with the desired error correction capability and $\mathcal{G}$ is designed to optimize energy efficiency.

Error Correction and Scheduling : One way to choose $\mathcal{C}$ would be as all possible binary sequences of length L, $\mathcal{C}\equiv\{0,1\}^L$. However, a single error during the beam-alignment phase would result in an incorrect selection of the communication beam. For instance, in the case L=3, if the error-free codeword is $\boldsymbol{c}_i$=[1, 1, 1] (and thus $\boldsymbol{\theta}_i\in\mathcal{U}_{[1,1,1]}$) but UE$_i$ detects $\boldsymbol{u}_i$=[1, 0, 1], then it will incorrectly infer that $\boldsymbol{\theta}_i\in\mathcal{U}_{[1,0,1]}$, resulting in outage in the data communication phase.

In order to compensate for detection errors, we endow $\mathcal{C}$ with error correction capabilities up to ε errors, e.g., using Hamming codes. Therefore, at the end of the beam-alignment phase, each UE applies the decoding function $f : \{0,1\}^L\to\mathcal{C}$ to the detection sequence $\boldsymbol{u}_i$. In this chapter, we use the minimum Hamming distance criterion to design $f(\cdot)$, i.e.,

$$f(\boldsymbol{u}) \triangleq \arg\min_{\boldsymbol{c}\in\mathcal{C}} \|\boldsymbol{u} - \boldsymbol{c}\|_2^2. \tag{3.18}$$

After decoding, each UE feeds back to the BS the ID of its corrected sequence $\hat{\boldsymbol{c}}_i \triangleq f(\boldsymbol{u}_i),\ \forall i \in \{1, 2, \ldots, M\}$. We assume that the feedback signals are received without errors at the BS, which thus infers that

$$\boldsymbol{\theta}_i \in \mathcal{U}_{f(\boldsymbol{u}_i)}, \tag{3.19}$$

where $\mathcal{U}_{\boldsymbol{d}}$ is defined in (3.13). Given $f(\boldsymbol{u}_i)$, the BS allocates the communication resources $(\tau_i, \mathcal{B}_{\mathrm{d},i}, P_i, R_i)$ to UE$_i$ during the data communication phase, denoting the allocated time, BS transmission power and rate, and communication beam. In this chapter, we assume a

TDMA strategy, i.e., $\tau_i = T_d/M$, $\forall i \in \{1, 2, \ldots, M\}$. The beam pair $\mathcal{B}_{\mathrm{d},i} \equiv \mathcal{B}^{\mathrm{d}}_{\mathrm{t},i} \times \mathcal{B}^{\mathrm{d}}_{\mathrm{r},i}$ is chosen as

$$\mathcal{B}_{\mathrm{d},i} \equiv \mathcal{B}_{\mathrm{d}}(\boldsymbol{u}_i) = \mathcal{U}_{f(\boldsymbol{u}_i)}. \tag{3.20}$$

Note that, due to the error correction capability endowed in the design of $\mathcal{C}$, if less than (or equal to) ε errors have been introduced in the beam-alignment phase, then $f(\boldsymbol{u}_i) = \boldsymbol{c}_i$, and thus correct alignment is achieved in the data communication phase ($\mathcal{B}_{\mathrm{d},i} \equiv \mathcal{U}_{\boldsymbol{c}_i}$); otherwise, if $f(\boldsymbol{u}_i) \neq \boldsymbol{c}_i$, then the data communication beam is not aligned with the AoD/AoA, and outage occurs ($\mathcal{B}_{\mathrm{d},i} \cap \mathcal{U}_{\boldsymbol{c}_i} \equiv \emptyset$). The resulting mis-alignment probability of UE$_i$ can then be bounded as

$$p_{\mathrm{ma},i}(\mathcal{B}_{\mathrm{d}}) \leq \mathbb{P}(W(\boldsymbol{e}_i) > \varepsilon) = \sum_{l=\varepsilon+1}^{L} \begin{pmatrix} L \\ l \end{pmatrix} p_e^l (1-p_e)^{L-l}, \tag{3.21}$$

as per the error model (3.12). Note that this is a function of ϕ_s via (3.10), duration L of beam-alignment and number of correctable errors ε (i.e., choice of the error correction codebook $\mathcal{C}$). However, it is independent of the beam-alignment sequence $\mathcal{B}_k, k \in \mathcal{I}_s$. Therefore, the optimization over ϕ_s, L, $\mathcal{C}$ and $\mathcal{B}_k, k \in \mathcal{I}_s$ can be decoupled: ϕ_s, L, $\mathcal{C}$ can be chosen to achieve a target mis-alignment performance $p_{\mathrm{ma},i} \leq p^{\max}_{\mathrm{ma}}, \forall i$, whereas $\mathcal{B}_k, k \in \mathcal{I}_s$ is optimized to achieve energy-efficient design. This optimization is developed in the next section.

Data Communication: In the data communication phase, the BS transmits to UE$_i$ in the assigned TDMA slot using power P_i and rate R_i. These are designed to satisfy a maximum outage probability $p_{\mathrm{out}}(P_i, R_i) \leq \rho$, with no CSI at the transmitter ($h_k^{(i)}$ unknown at BS), and a minimum rate constraint $R_{\min,i}$ of UE$_i$ over the frame. In case of mis-alignment, data communication is in outage, see (3.21). We now consider the case of alignment, i.e., $f(\boldsymbol{u}_i) = \boldsymbol{c}_i$ and $\boldsymbol{\theta}_i \in \mathcal{B}_{\mathrm{d}}(\boldsymbol{u}_i)$. In this case, the instantaneous signal-to-noise ratio (SNR) during the data communication slots associated with UE$_i$ is

$$\mathsf{SNR}_k^{(i)} = \frac{(2\pi)^2 \gamma_k^{(i)} P_i}{N_0 W_{\mathrm{tot}} |\mathcal{B}_{\mathrm{d}}(\boldsymbol{u}_i)|}, \tag{3.22}$$

where $\gamma_k^{(i)} \triangleq |h_k^{(i)}|^2$. The outage probability is then given by

$$\begin{aligned} p_{\text{out}}(P_i, R_i) &= \mathbb{P}(W_{\text{tot}} \log_2(1 + \mathsf{SNR}_k^{(i)}) \leq R_i | \boldsymbol{u}_i) \\ &= 1 - \exp\left(-(2^{\frac{R_i}{W_{\text{tot}}}} - 1) \frac{\ell(d_i) N_0 W_{\text{tot}}}{P_i (2\pi)^2} |B_{\text{d}}(\boldsymbol{u}_i)| \right). \end{aligned}$$

To meet the minimum rate constraint of UE_i over the frame, we enforce $R_i = \frac{T_{\text{fr}}}{\tau_i} R_{\min,i}$. To enforce $p_{\text{out}}(P_i, R_i) \leq \rho$,[1] we find the power P_i and the energy $E_i \triangleq P_i \tau_i$ as

$$E_i = \phi_{\text{d},i} |\mathcal{B}_{\text{d}}(\boldsymbol{u}_i)|, \tag{3.23}$$

where $\phi_{\text{d},i}$ is the minimum energy/rad^2 required to meet the rate requirement of UE_i with outage probability ρ, given by

$$\phi_{\text{d},i} \triangleq \frac{\tau_i \ell(d_i) N_0 W_{\text{tot}} \left[2^{\frac{T_{\text{fr}} R_{\min,i}}{\tau_i W_{\text{tot}}}} - 1 \right]}{(2\pi)^2 \ln(1/(1-\rho))}. \tag{3.24}$$

3.2 Optimization Problem

The optimum beam-alignment design seeks to minimize the average power consumption $\bar{P}_{\text{avg}}(\mathcal{B})$ of the BS, over the beam-sequence $\mathcal{B}=\{\mathcal{B}_k, k \in \mathcal{I}_s\}$ in the beam-alignment phase, i.e.,

$$\boldsymbol{P1}: \ \mathcal{B}^* = \arg\min_{\mathcal{B}} \bar{P}_{\text{avg}}(\mathcal{B}), \tag{3.25}$$

where, using (3.9) and (3.23), $\bar{P}_{\text{avg}}(\mathcal{B})$ is given by

$$\bar{P}_{\text{avg}}(\mathcal{B}) = \frac{1}{T_{\text{fr}}} \mathbb{E}\left[\sum_{k=1}^{L} \phi_s \, |\mathcal{B}_k| + \sum_{i=1}^{M} \phi_{d,i} \, |\mathcal{B}_{\text{d}}(\boldsymbol{u}_i)| \right], \tag{3.26}$$

with $\mathcal{B}_{\text{d}}(\boldsymbol{u}_i)$ given by (3.20). The expectation is over the detected and error-free sequences $\{(\boldsymbol{u}_i, \boldsymbol{c}_i), i = 1, \ldots, M\}$.

[1] ↑Note that the overall outage probability including mis-alignment is given by $p_{\text{ma}}^{\max} + (1 - p_{\text{ma}}^{\max})\rho$.

Using (3.17) and (3.20), we can express the beam-alignment and data communication beams as

$$\mathcal{B}_k = \cup_{\boldsymbol{d}\in\mathcal{C}:d_k=1}\mathcal{U}_{\boldsymbol{d}}, \quad \mathcal{B}_{\rm d}(\boldsymbol{u}_i) = \mathcal{U}_{f(\boldsymbol{u}_i)}. \tag{3.27}$$

In fact, $\mathcal{U}_{\hat{\boldsymbol{c}}_i}$ represents the estimated support of the AoD/AoA of UE$_i$, when it detects the error corrected sequence $\hat{\boldsymbol{c}}_i{=}f(\boldsymbol{u}_i)$. Note that $\{\mathcal{U}_{\boldsymbol{d}} : \boldsymbol{d}{\in}\mathcal{C}\}$ forms a partition of the AoD/AoA space $[-\pi/2,\pi/2]^2$. In fact, using (3.13), the fact that $\cap_{k=1}^L \mathcal{B}_k^{d_k} \equiv \emptyset, \forall \boldsymbol{d} \notin \mathcal{C}$, and the set definition (3.14), we can show that

$$\cup_{\boldsymbol{d}\in\mathcal{C}}\,\mathcal{U}_{\boldsymbol{d}} \equiv \cup_{\boldsymbol{d}\in\{0,1\}^L}\cap_{k=1}^L\,\mathcal{B}_k^{d_k} = [-\pi/2,\pi/2]^2, \tag{3.28}$$

$$\mathcal{U}_{\boldsymbol{d}_1}\cap\mathcal{U}_{\boldsymbol{d}_2} \equiv \cap_{k=1}^L[\mathcal{B}_k^{d_{1,k}}\cap\mathcal{B}_k^{d_{2,k}}] \equiv \emptyset, \quad \forall \boldsymbol{d}_1 \neq \boldsymbol{d}_2. \tag{3.29}$$

Therefore, letting $\omega_{\boldsymbol{d}} \triangleq |\mathcal{U}_{\boldsymbol{d}}|$ be the beamwidth of the sector $\mathcal{U}_{\boldsymbol{d}}$ and using (3.27), we can rewrite the average power as

$$\begin{aligned}
\bar{P}_{\rm avg}(\omega){=}&\frac{1}{T_{\rm fr}}\,\mathbb{E}\bigg[\sum_{k=1}^{L}\phi_s\,|\cup_{\boldsymbol{d}\in\mathcal{C}:d_k=1}\mathcal{U}_{\boldsymbol{d}}| \\
&+\sum_{i=1}^{M}\phi_{d,i}\Big\{\,|\mathcal{U}_{\boldsymbol{c}_i}|\,\chi(W(\boldsymbol{e}_i){\leq}\varepsilon)+\Big|\mathcal{U}_{f(\boldsymbol{c}_i\oplus\boldsymbol{e}_i)}\Big|\,\chi(W(\boldsymbol{e}_i){>}\varepsilon)\Big\}\bigg] \\
\overset{(a)}{=}&\frac{1}{T_{\rm fr}}\,\mathbb{E}\bigg[\sum_{k=1}^{L}\phi_s\sum_{\boldsymbol{d}\in\mathcal{C}:d_k=1}\omega_{\boldsymbol{d}} \\
&+\sum_{i=1}^{M}\phi_{d,i}\Big\{\omega_{\boldsymbol{c}_i}\chi(W(\boldsymbol{e}_i)\leq\varepsilon)+\omega_{f(\boldsymbol{c}_i\oplus\boldsymbol{e}_i)}\chi(W(\boldsymbol{e}_i)>\varepsilon)\Big\}\bigg],
\end{aligned}$$

where in (a) we used the facts that $\{\mathcal{U}_{\boldsymbol{c}}{:}\boldsymbol{c}{\in}\{0,1\}^L\}$ is a partition of $[-\pi/2,\pi/2)^2$ and that, if fewer than ε errors occur in the beam-alignment phase, then the support of $\boldsymbol{\theta}_i$ is detected correctly. Note that, since the AoD/AoA pair $\boldsymbol{\theta}_i$ is uniformly distributed in the space $[-\pi/2,\pi/2)^2$, the probability of occurrence of the error-free sequence $\boldsymbol{c}_i{=}\boldsymbol{x}$ is

$$\mathbb{P}(\boldsymbol{c}_i=\boldsymbol{x}) = \mathbb{P}(\boldsymbol{\theta}_i\in\cap_{k=1}^L\mathcal{B}_k^{x_k}) = \mathbb{P}(\boldsymbol{\theta}_i\in\mathcal{U}_{\boldsymbol{x}}) = \frac{\omega_{\boldsymbol{x}}}{\pi^2}, \tag{3.30}$$

while the error sequence $\boldsymbol{e}_i, \forall \boldsymbol{e}_i \in \{0,1\}^L$ follows the pmf $p(\boldsymbol{e}_i)$ given in (3.12). This leads to

$$\bar{P}_{\text{avg}}(\omega) = \frac{1}{T_{\text{fr}}}\Bigg[\phi_s \sum_{\boldsymbol{d}\in\mathcal{C}} W(\boldsymbol{d})\omega_{\boldsymbol{d}} + \frac{M\bar{\phi}_d}{\pi^2}\sum_{\boldsymbol{c}\in\mathcal{C}}\Big\{\omega_{\boldsymbol{c}}^2\,\mathbb{P}(W(\boldsymbol{e})\le\varepsilon) + \sum_{\boldsymbol{e}\in\{0,1\}^L: W(\boldsymbol{e})>\epsilon} \omega_{f(\boldsymbol{c}\oplus\boldsymbol{e})}\omega_{\boldsymbol{c}} p(\boldsymbol{e})\Big\}\Bigg], \tag{3.31}$$

where we used the fact that

$$\sum_{k=1}^{L}\sum_{\boldsymbol{d}\in\mathcal{C}:d_k=1}\omega_{\boldsymbol{d}} = \sum_{\boldsymbol{d}\in\mathcal{C}}\sum_{k=1}^{L}\chi(d_k=1)\omega_{\boldsymbol{d}} = \sum_{\boldsymbol{d}\in\mathcal{C}} W(\boldsymbol{d})\omega_{\boldsymbol{d}},$$

and we have defined $\bar{\phi}_d \triangleq \frac{1}{M}\sum_{i=1}^{M}\phi_{d,i}$. Thus, the optimization problem **P1** can be restated as that of optimizing the "beamwidths" $\omega_{\boldsymbol{d}}, \boldsymbol{d}\in\mathcal{C}$. The sequence of beams with desired beamwidth solution of this optimization problem can then be obtained via (3.27), where $|\mathcal{U}_{\boldsymbol{d}}| = \omega_{\boldsymbol{d}}$. Note that $\omega_{\boldsymbol{d}} \triangleq |\mathcal{U}_{\boldsymbol{d}}|$ needs to satisfy the constraint $\sum_{\boldsymbol{d}\in\mathcal{C}}\omega_{\boldsymbol{d}} = \pi^2$, since $\{\mathcal{U}_{\boldsymbol{d}}, \boldsymbol{d}\in\mathcal{C}\}$ is a partition of $[-\pi/2, \pi/2]^2$.

However, it can be shown that the cost function $\bar{P}_{\text{avg}}(\omega)$ is non-convex with respect to ω, due to the quadratic terms $\omega_{f(\boldsymbol{c}\oplus\boldsymbol{e})}\omega_{\boldsymbol{c}}$ appearing in (3.31). In order to overcome this limitation, we propose to upper bound (3.31) by a convex function. To determine this upper bound, note that the partition constraint $\sum_{\boldsymbol{d}\in\mathcal{C}}\omega_{\boldsymbol{d}} = (2\pi)^2$ and $\omega_{\boldsymbol{d}}\ge 0, \forall\boldsymbol{d}\in\mathcal{C}$ imply that $\omega_{f(\boldsymbol{c}\oplus\boldsymbol{e})} \le \pi^2$. Thus, we upper bound (3.31) as

$$\bar{P}_{\text{avg}}(\omega) \le \frac{1}{T_{\text{fr}}}\Bigg[\phi_s \sum_{\boldsymbol{d}\in\mathcal{C}} W(\boldsymbol{d})\omega_{\boldsymbol{d}} + \frac{M\bar{\phi}_d}{\pi^2}\sum_{\boldsymbol{c}\in\mathcal{C}}\Big\{\mathbb{P}(W(\boldsymbol{e})\le\varepsilon)(\omega_{\boldsymbol{c}}^2 - \pi^2\omega_{\boldsymbol{c}}) + \pi^2\omega_{\boldsymbol{c}}\Big\}\Bigg] \triangleq \hat{P}_{\text{avg}}(\omega). \tag{3.32}$$

Note that, if the probability of incurring more than ε errors is made sufficiently small (by appropriately choosing the error correction code $\mathcal{C}$), say $\mathbb{P}(W(\boldsymbol{e})>\varepsilon)\le\delta \ll 1$, then we can bound the gap $\hat{P}_{\text{avg}}(\omega) - \bar{P}_{\text{avg}}(\omega)$ by

$$0 \le \hat{P}_{\text{avg}}(\omega) - \bar{P}_{\text{avg}}(\omega) \le \frac{M\bar{\phi}_d\pi^2}{T_{\text{fr}}}\delta, \tag{3.33}$$

Thus, we consider the minimization of the upper bound $\hat{P}_{\text{avg}}(\omega)$ instead of the original function $\bar{P}_{\text{avg}}(\omega)$, yielding the optimization problem

$$\boldsymbol{P2}: \ \omega^* = \arg\min_{\omega \geq 0} \hat{P}_{\text{avg}}(\omega) \text{ s.t. } \sum_{\boldsymbol{d} \in \mathcal{C}} \omega_{\boldsymbol{d}} = \pi^2, \tag{3.34}$$

We now study the optimization problem $\boldsymbol{P2}$. Note that this is a convex quadratic problem with respect to $\omega_{\boldsymbol{c}} : \boldsymbol{d} \in \mathcal{C}$. The dual function associated with $\boldsymbol{P2}$ is given by

$$g(\mu) = \min_{\omega \geq 0} \hat{P}_{\text{avg}}(\omega) - \mu \left(\sum_{\boldsymbol{d} \in \mathcal{C}} \omega_{\boldsymbol{d}} - \pi^2 \right),$$

whose minimizer yields the "water-filling" solution

$$\omega^*_{\boldsymbol{d}} = \frac{\pi^2 \phi_s}{2\mathbb{P}(W(\boldsymbol{e}) \leq \varepsilon) M \bar{\phi}_d} \left[\lambda - W(\boldsymbol{d}) \right]^+. \tag{3.35}$$

The dual variable λ is chosen so as to satisfy the constraint

$$\sum_{\boldsymbol{d} \in \mathcal{C}} \omega^*_{\boldsymbol{d}} = \pi^2, \tag{3.36}$$

or equivalently, as the unique solver of

$$h(\lambda) = \frac{\phi_s}{2\mathbb{P}(W(\boldsymbol{e}) \leq \varepsilon) M \bar{\phi}_d} \sum_{w=0}^{L} n_w \left[\lambda - w \right]^+ = 1, \tag{3.37}$$

where $n_w \triangleq \sum_{\boldsymbol{c} \in \mathcal{C}} \chi\left(W(\boldsymbol{c}) = w\right)$ is the number of codewords in the codebook $\mathcal{C}$ with Hamming weight equal to w.

The optimal dual variable λ^* can be found using the bisection method over the interval $[\lambda_{\min}, \lambda_{\max}]$. In fact, $h(\lambda)$ is a non-decreasing function of $\lambda > 0$, with $h(0) = 0$ and, using the fact that $[\lambda - w]^+ \leq \lambda$, we find that

$$h(\lambda) \leq \frac{\phi_s}{2\mathbb{P}(W(\boldsymbol{e}) \leq \varepsilon) M \bar{\phi}_d} \lambda |\mathcal{C}|,$$

where $|\mathcal{C}|$ is the cardinality of $\mathcal{C}$, hence $\lambda^* \geq \frac{2M\bar{\phi}_d P(W(\boldsymbol{e})\leq\varepsilon)}{|\mathcal{C}|\phi_s}$. Moreover, by denoting $\overline{W} \triangleq \frac{1}{|\mathcal{C}|}\sum_{w=0}^{L} n_w w$ as the average weight of the codewords in $\mathcal{C}$, we observe that

$$\sum_{w=0}^{L} n_w \left[\lambda - w\right] = \left[\lambda - \overline{W}\right]|\mathcal{C}| \leq \frac{2M\bar{\phi}_d P(W(\boldsymbol{e}) \leq \varepsilon)}{\phi_s} h(\lambda),$$

thus implying the following upper and lower bounds to λ^*,

$$\lambda_{\min} \triangleq \frac{2M\bar{\phi}_d P(W(\boldsymbol{e}) \leq \varepsilon)}{|\mathcal{C}|\phi_s} \leq \lambda^* \leq \lambda_{\min} + \overline{W} \triangleq \lambda_{\max}.$$

3.3 Numerical Results

In this section, we compare the performance of the proposed scheme with other schemes. We use Monte-Carlo simulation with 10^5 iterations for each simulation point. The common simulation parameters used are as follows: T_{fr}=20ms, T=10μs, [Number of BS antennas]=64, [Number of UE antennas] = 1, [BS-UE separation]=10m, N_0= $-$ 173dBm, W_{tot}=500MHz, [carrier frequency]=30GHz, ϕ_s=6dBm, and ρ=10^{-3}. Moreover, we use the beamforming algorithm in [65] to generate the beamforming codebook. With these values, we have observed numerically that the probability of detection errors is in the range $p_e \in [0.1, 0.3]$, due not only to noise and the Rayleigh fading channel, but also to sidelobes, which are not accounted for in the hypothesis testing problem (3.3). Thus, we set p_e=0.3 to capture this more realistic scenario.

In Fig. 3.2, we depict the spectral efficiency (Throughput/W_{tot}) versus the average power consumption. The curves correspond to three different choices of the codebook $\mathcal{C}$: the Hamming codebook $\mathcal{C}$=(7, 4), representing the proposed coded energy-efficient scheme, with error correction capability up to ε=1 errors; $\mathcal{C}$=$\{[\boldsymbol{I}]_{:,i}, i$=$1, \ldots, L\}$, representing the exhaustive search scheme, where $[\boldsymbol{I}]_{:,i}$ denotes the ith column of the $L \times L$ identity matrix $\boldsymbol{I}$; and $\mathcal{C} = \{0, 1\}^L$, representing the scheme with no error correction capabilities (uncoded). We use L=16 for the exhaustive search scheme, and L=7 for the coded and uncoded schemes. In the figure, we observe that the proposed scheme using (7,4) Hamming codebook outperforms the other two schemes, thanks to its error correction capabilities, with a performance gain up to

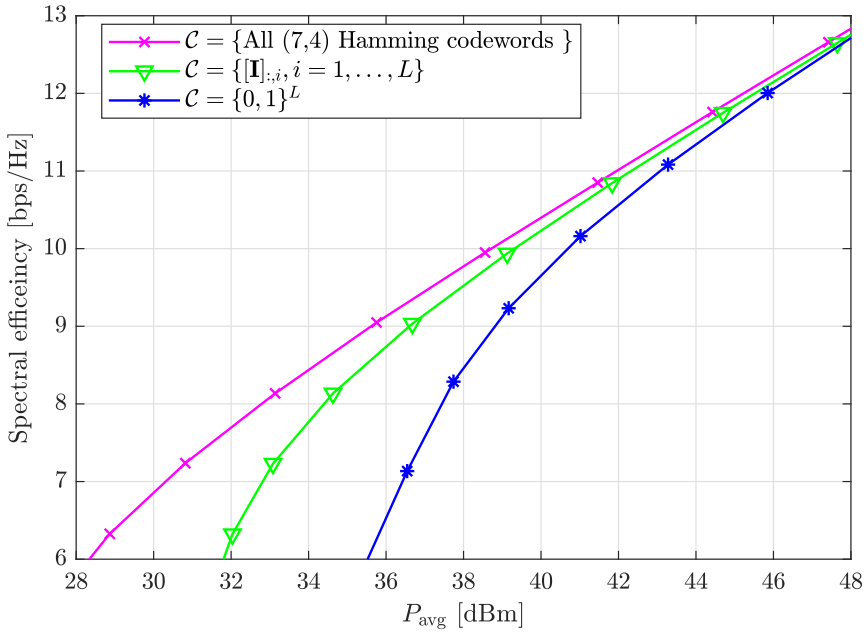

Figure 3.2. Spectral Efficiency versus average power consumption.

4dB over exhaustive and 8dB over the uncoded scheme. Surprisingly, the exhaustive scheme exhibits superior performance compared to the uncoded scheme, despite its more significant time overhead (L=16 vs L=7). This can be attributed to the fact that the codewords in the exhaustive codebook exhibit a minimum Hamming distance of 2, whereas the uncoded codebook exhibits minimum Hamming distance equal to 1, and is thus more susceptible to detection errors during the beam-alignment phase.

4. SECOND-BEST BEAM-ALIGNMENT VIA BAYESIAN MULTI-ARMED BANDITS

In this chapter, a beam-alignment scheme is proposed based on Bayesian multi-armed bandits, with the goal to maximize the alignment probability and the data-communication throughput. A Bayesian approach is proposed, by considering the state as a posterior distribution over angles of arrival (AoA) and of departure (AoD), given the history of feedback signaling and of beam pairs scanned by the base-station (BS) and the user-end (UE). A simplified sufficient statistic for optimal control is identified, in the form of preference of BS-UE beam pairs. By bounding a value function, the *second-best preference* policy is formulated, which strikes an optimal balance between exploration and exploitation by selecting the beam pair with the current *second-best* preference. Through Monte-Carlo simulation with analog beamforming, the superior performance of the *second-best preference* policy is demonstrated in comparison to existing schemes based on *first-best preference*, linear Thompson sampling, and upper confidence bounds, with up to 7%, 10% and 30% improvements in alignment probability, respectively.

4.1 System Model

We consider a downlink scenario with one BS and one UE, as depicted in Fig. 4.1. Time is divided into frames of duration $T_{\rm fr}{=}T_{\rm s}N$, each with N slots of duration $T_{\rm s}$. The frame is partitioned into two phases: a *beam-alignment phase* of duration $LT_{\rm s}$ ($L{<}N$ slots), followed by a *downlink data communication phase*, of duration $(N{-}L)T_{\rm s}$. Each beam-alignment slot is further partitioned into a *pilot transmission phase*, of duration $T_{\rm pt}$, followed by a *feedback phase*, of duration $T_{\rm fb}$, with $T_{\rm s}{=}T_{\rm pt}{+}T_{\rm fb}$. These are detailed next.

The BS and UE are equipped with uniform linear arrays (ULAs) with $M_{\rm t}$ and $M_{\rm r}$ antenna elements, respectively, and use analog beamforming. The signal received at the UE is

$$\boldsymbol{z}_k = \sqrt{P_{{\rm tx},k}}\boldsymbol{u}_k^H \boldsymbol{H}_k \boldsymbol{v}_k \boldsymbol{s} + \boldsymbol{w}_k,\ \forall k \in \{0,1,\ldots,N-1\}, \tag{4.1}$$

↑A version of this chapter was previously published by IEEE Globecom 2019 [6][DOI:10.1109/GLOBECOM38437.2019.9013578]

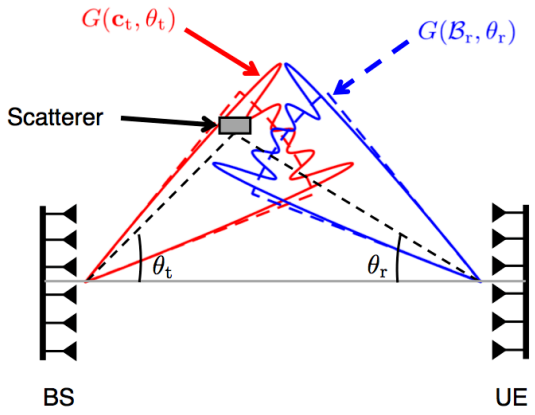

Figure 4.1. System model; $M_\mathrm{t} = M_\mathrm{r} = 128$; beamforming algorithm in [27].

where $P_{\mathrm{tx,k}}$ is the average transmit power of the BS; $\mathbf{s} \in \mathbb{C}^S$ is the transmitted signal with S symbols with $\mathbb{E}\left[\|\boldsymbol{s}\|_2^2\right] = S$; $\boldsymbol{H}_k \in \mathbb{C}^{M_\mathrm{r} \times M_\mathrm{t}}$ is the channel matrix; $\boldsymbol{v}_k \in \mathbb{C}^{M_\mathrm{t}}$ is the BS beamforming vector with $\|\boldsymbol{v}_k\|_2^2 = 1$; $\boldsymbol{u}_k \in \mathbb{C}^{M_\mathrm{r}}$ is the UE combining vector with $\|\boldsymbol{u}_k\|_2^2 = 1$; $\boldsymbol{w}_k \sim \mathcal{CN}(\mathbf{0}, N_0 W_\mathrm{tot} \boldsymbol{I})$ is additive white Gaussian noise (AWGN), with one-sided power spectral density N_0 and system bandwidth W_tot.

Channel Model: We use the extended Saleh-Valenzuela geometric model with a single-cluster [55], as adopted in several previous works (e.g., see [4], [56], [57]). In fact, typical mm-wave channels have been shown to exhibit one dominant cluster containing most of the signal energy [66]. The single-cluster channel is modeled as

$$\boldsymbol{H}_k = \alpha_k \boldsymbol{a}_\mathrm{r}(\theta_{\mathrm{r},k}) \boldsymbol{a}_\mathrm{t}^H(\theta_{\mathrm{t},k}), \tag{4.2}$$

where $\theta_k \triangleq (\theta_{\mathrm{r},k}, \theta_{\mathrm{t},k}) \in [-\frac{\pi}{2}, \frac{\pi}{2}]^2$ is the angle of arrival (AoA) and angle of departure (AoD) pair associated to the dominant cluster, with complex fading gain α_k; $\boldsymbol{a}_\mathrm{r}$ and $\boldsymbol{a}_\mathrm{t}$ are the UE and BS array response vectors, respectively, defined as

$$\boldsymbol{a}_\mathrm{x}(\theta_\mathrm{x}) = \frac{1}{\sqrt{M_\mathrm{x}}} \left[1, e^{j\frac{2\pi d_\mathrm{x}}{\lambda}\psi_\mathrm{x}}, \cdots, e^{j(M_\mathrm{x}-1)\frac{2\pi d_\mathrm{x}}{\lambda}\psi_\mathrm{x}}\right]^\top, \mathrm{x} \in \{\mathrm{t}, \mathrm{r}\},$$

where $\psi_{\rm x}=\sin\theta_{\rm x}$, $d_{\rm x}$ is the antenna spacing, $\lambda=c/f_c$ is the wavelength at carrier frequency f_c, c denotes the speed of light. We assume that during the duration of one frame $T_{\rm fr}$, θ_k remains unchanged, $\theta_k=\theta$, and α_k are i.i.d. Reyleigh fading in each slot with distribution $\alpha_k\sim\mathcal{CN}(0,\ell(d)^{-1})$, where $\ell(d)$ is the path loss at distance d from the BS. In fact, the AoA-AoD pair change much slower than the channel gain [58].

Codebook structure: In slot k, the BS uses the beamforming vector $\boldsymbol{v}_k\in\mathcal{V}$ and the UE uses the combining vector $\boldsymbol{u}_k\in\mathcal{U}$, from the codebooks $\mathcal{V}$ and $\mathcal{U}$, respectively. We assume a sectored model [4], in which the AoA and AoD spaces are partitioned into sectors of equal beamwidth (as shown in Fig. 4.1 for the case of four sectors, this model approximates well analog beamforming). Accordingly, let $\mathcal{B}_{\rm r}(\boldsymbol{u})\subseteq[-\frac{\pi}{2},\frac{\pi}{2}]$ and $\mathcal{B}_{\rm t}(\boldsymbol{v})\subseteq[-\frac{\pi}{2},\frac{\pi}{2}]$ denote the AoA and AoD supports of the UE combiner and BS beamformer vectors $\boldsymbol{u}\in\mathcal{U}$ and $\boldsymbol{v}\in\mathcal{V}$, respectively, with equal beamwidth $|\mathcal{B}_{\rm r}(\boldsymbol{u})|=\frac{\pi}{|\mathcal{U}|},\forall\boldsymbol{u}\in\mathcal{U}$ and $|\mathcal{B}_{\rm t}(\boldsymbol{v})|=\frac{\pi}{|\mathcal{V}|},\forall\boldsymbol{v}\in\mathcal{V}$, where $|\mathcal{B}|$ denotes the measure $|\mathcal{B}|\triangleq\int_{\mathcal{B}}\mathrm{d}x$. We define $\mathcal{B}(\boldsymbol{u},\boldsymbol{v})\triangleq\mathcal{B}_{\rm r}(\boldsymbol{u})\times\mathcal{B}_{\rm t}(\boldsymbol{v})$ as the joint AoA-AoD support of $(\boldsymbol{u},\boldsymbol{v})$. We assume that the angular supports are mutually orthogonal and form a partition of the entire AoA-AoD space $[-\frac{\pi}{2},\frac{\pi}{2}]^2$, i.e., $\mathcal{B}(\boldsymbol{u},\boldsymbol{v})\cap\mathcal{B}(\tilde{\boldsymbol{u}},\tilde{\boldsymbol{v}})=\emptyset,\forall(\boldsymbol{u},\boldsymbol{v})\neq(\tilde{\boldsymbol{u}},\tilde{\boldsymbol{v}})$ and $\cup_{\boldsymbol{u}\in\mathcal{U}}\mathcal{B}_{\rm r}(\boldsymbol{u})=\cup_{\boldsymbol{v}\in\mathcal{V}}\mathcal{B}_{\rm t}(\boldsymbol{v})=[-\frac{\pi}{2},\frac{\pi}{2}]$. Let $(\boldsymbol{u}^{(i)},\boldsymbol{v}^{(i)}),i\in\mathcal{I}\triangleq\{1,2,\ldots,|\mathcal{U}||\mathcal{V}|\}$ be any ordering of combining and beamforming vectors, and $\mathcal{B}^{(i)}\triangleq\mathcal{B}(\boldsymbol{u}^{(i)},\boldsymbol{v}^{(i)})$ be their support. Let $A_k\in\mathcal{I}$ be the beam index of the combining and beamforming vectors scanned in slot k, so that $(\boldsymbol{u}_k,\boldsymbol{v}_k)=(\boldsymbol{u}^{(A_k)},\boldsymbol{v}^{(A_k)})$. Let X be a discrete random variable denoting the index of the support that the AoA-AoD pair θ of the channel belongs to, so that $\theta\in\mathcal{B}^{(X)}$. Then, from (4.1)-(4.2), the received signal can be expressed as[1]

$$\boldsymbol{z}_k\approx\sqrt{P_{\rm tx,k}}\alpha_k\Big[(\sqrt{G}-\sqrt{g})\delta[A_k,X]+\sqrt{g}\Big]\boldsymbol{s}+\boldsymbol{w}_k, \tag{4.3}$$

where $\delta[\,\cdot\,]$ is the Kronecker's delta function, equal to 1 if alignment is achieved ($A_k=X$), equal to 0 otherwise ($A_k\neq X$); G and g are, respectively, the main and side lobe gains of the sectored model, expressed as

$$G=\min_{(\theta_{\rm r},\theta_{\rm t})\in\mathcal{B}(\boldsymbol{u}^{(i)},\boldsymbol{v}^{(i)})}|\boldsymbol{a}_{\rm r}(\theta_{\rm r})^H\boldsymbol{u}^{(i)}|^2|\boldsymbol{a}_{\rm t}^H(\theta_{\rm t})\boldsymbol{v}^{(i)}|^2,\ \forall i,$$

[1]↑The phase of $\boldsymbol{u}_k^H\boldsymbol{a}_{\rm r}(\theta_{\rm r})\boldsymbol{a}_{\rm t}^H(\theta_{\rm t})\boldsymbol{v}_k$ is incorporated into α_k.

$$g = \max_{(\theta_{\rm r},\theta_{\rm t})\notin\mathcal{B}(\boldsymbol{u}^{(i)},\boldsymbol{v}^{(i)})} |\boldsymbol{a}_{\rm r}(\theta_{\rm r})^H \boldsymbol{u}^{(i)}|^2 |\boldsymbol{a}_{\rm t}^H(\theta_{\rm t})\boldsymbol{v}^{(i)}|^2, \ \forall i.$$

In the following, we describe the beam-alignment and data communication procedures.

Beam-Alignment: In each slot k of the beam-alignment phase, the BS transmits a pilot sequence $\boldsymbol{s}$ using the beam index A_k, with transmit power $P_{{\rm tx},k}{=}P_{\rm ba}$. Upon receiving $\boldsymbol{z}_k$ (based on the combining vector with index A_k), the UE uses a matched filter to compute the signal strength and sends the *normalized received power* feedback signal Y_k back to the BS, of the form

$$Y_k = \frac{|\boldsymbol{s}^H \boldsymbol{z}_k|^2}{\|\boldsymbol{s}\|^2 N_0 W_{\rm tot}(1+\Lambda g)}, \tag{4.4}$$

where $\Lambda \triangleq \frac{P_{\rm ba}\|\boldsymbol{s}\|^2}{N_0 W_{\rm tot}\ell(d)}$ is the pre-beamforming receive SNR during beam-alignment. Then, the probability density function (pdf) of Y_k conditional on $(X, A_k){=}(x, a_{\rm s})$ is given by

$$f(Y_k{=}y|X{=}x; A_k{=}a_{\rm s}){=}\left[\nu e^{-\nu y}\right]^{\delta[a_{\rm s},x]} [e^{-y}]^{1-\delta[a_{\rm s},x]}, \tag{4.5}$$

where $1/\nu$ is the mean signal power in case of alignment, with

$$\nu \triangleq \frac{1+g\Lambda}{1+G\Lambda}. \tag{4.6}$$

The BS uses a Bayesian approach to select A_k: starting from $\mathcal{H}_0{\triangleq}\emptyset$ and given the history of feedback and scanned beam indices $\mathcal{H}_k{\triangleq}\{(A_j, Y_j)\}_{j=0}^{k-1}$, the next beam index A_k is selected. This procedure continues until the end of the beam-alignment phase.

Data communication: Upon completion of the beam-alignment phase, given the history of feedback and actions $\mathcal{H}_L$, the BS selects the data communication parameters: beam index for data communication $A_{\rm d}{\in}\mathcal{I}$, transmission power $P_{\rm d}{\in}[0, P_{\max}]$, and data rate $R_{\rm d}{\geq}0$. These parameters are used until the end of the data communication phase.

Let $b_0[x]$ be the *prior belief* over $X{=}x$ (or equivalently over $\theta{\in}\mathcal{B}^{(x)}$) available at the beginning of the beam-alignment phase. We define the expected rate during the communication phase (normalized by the frame duration), as

$$\begin{aligned}&\bar{R}(A_{\mathrm{d}}, P_{\mathrm{d}}, R_{\mathrm{d}}|b_0, \mathcal{H}_L)\\ &\quad\triangleq \frac{T_{\mathrm{fr}} - LT_{\mathrm{s}}}{T_{\mathrm{fr}}}\, \mathbb{P}\left(X = A_{\mathrm{d}}\middle|b_0, \mathcal{H}_L\right) \hat{R}(R_{\mathrm{d}}, P_{\mathrm{d}}),\end{aligned} \tag{4.7}$$

where we have defined

$$\hat{R}(R_{\mathrm{d}}, P_{\mathrm{d}}){\triangleq}R_{\mathrm{d}}\mathbb{P}\left[R_{\mathrm{d}}{\leq}W_{\mathrm{tot}} \log_2\left(1 + \frac{|\alpha_k|^2 P_{\mathrm{d}} G}{N_0 W_{\mathrm{tot}}}\right)\right]. \tag{4.8}$$

The probability term in (4.7) is the probability of achieving correct alignment, given the prior b_0 and the history $\mathcal{H}_L$ of feedback and actions during the beam-alignment phase, whereas the probability term in (4.8) denotes the probability of non-outage with respect to the realization of the fading process (i.i.d. over time), given that correct alignment has been achieved (we assume that mis-alignment yields outage with probability one, since $g \ll G$).

4.2 Problem Formulation and Solution

We now formulate the beam-alignment and data communication problem in the context of a decision process. We define a policy μ, part of our design, which operates as follows. At time k during beam-alignment, given the history of feedback and actions $\mathcal{H}_k$, the BS selects the *beam-alignment action* $A_k{=}a_{\mathrm{s}}{\in}\mathcal{I}$ with probability $\mu_k(a_{\mathrm{s}}|\mathcal{H}_k)$; given $\mathcal{H}_L$, the BS selects the data communication parameters as $(A_{\mathrm{d}}, P_{\mathrm{d}}, R_{\mathrm{d}}) = \mu_{\mathrm{d}}(\mathcal{H}_L)$. The goal is to design μ so as to maximize the expected communication rate, i.e.,

$$\textbf{P0:}\ \max_{\mu} \mathbb{E}_{\mu}\left[\bar{R}(A_{\mathrm{d}}, P_{\mathrm{d}}, R_{\mathrm{d}}|b_0, \mathcal{H}_L)|b_0\right],$$

where the expectation $\mathbb{E}_\mu$ is conditional on the prior belief b_0 and on the policy μ being executed during beam-alignment and data communication. Note that, using (4.7), we can rewrite the optimization problem as

$$\textbf{P1:}\ \max_{\mu} \mathbb{E}_\mu \left[\mathbb{P}\left(X = A_{\mathrm{d}} \middle| b_0, \mathcal{H}_L\right) \middle| b_0 \right] \frac{T_{\mathrm{fr}} - LT_{\mathrm{s}}}{T_{\mathrm{fr}}} \max_{R_d \geq 0, 0 \leq P_d \leq P_{\max}} \hat{R}(R_{\mathrm{d}}, P_{\mathrm{d}}),$$

i.e., the problem can be decomposed into the following two independent problems: 1) find the optimal rate and power (R_d^*, P_d^*) that maximize the expected rate in the communication phase, conditional on correct alignment being achieved ($X{=}A_{\mathrm{d}}$); 2) find the optimal beam-alignment policy and the beam index for communication A_{d} so as to maximize the probability of correct alignment. The first problem can be solved efficiently by maximizing (4.8). In the sequel, we consider the latter problem.

Let $b_k[x] \triangleq \mathbb{P}(X = x | \mathcal{H}_k, b_0)$ be the belief over $X{=}x$ given the history of actions and feedback and prior belief b_0. It serves as a sufficient statistic for optimal control for problem **P1**. In the following lemma, we present an equivalent simplified sufficient statistic along with its dynamics.

Lemma 4.1. *Let $m_0[x] \triangleq \ln b_0[x]$ denote the prior preference of $X = x$. Given the action and feedback pair (A_k, Y_k), the belief at $k+1$ is updated as*

$$b_{k+1}[x] = \frac{\exp\{m_{k+1}[x]\}}{\sum_{l \in \mathcal{I}} \exp\{m_{k+1}[l]\}}, \tag{4.9}$$

where

$$m_{k+1}[x] = m_k[x] + J(Y_k)\delta[A_k, x],\ \forall x \in \mathcal{I}, \tag{4.10}$$

and we have defined

$$J(y) \triangleq (1-\nu)y + \ln \nu. \tag{4.11}$$

Proof. Given the belief b_k and $(A_k, Y_k) = (a_s, y)$, we have

$$\begin{aligned} b_{k+1}[x] &\overset{(a)}{=} \mathbb{P}(X = x|\mathcal{H}_{k+1}) \\ &\overset{(b)}{\propto} f(Y_k = y|X = x, A_k = a_s, \mathcal{H}_k)\ \mathbb{P}(X = x|A_k = a_s, \mathcal{H}_k) \\ &\overset{(c)}{=} f(Y_k = y|X = x, A_k = a_s) b_k[x] \\ &\overset{(d)}{=} [\nu \exp\{-\nu y\}]^{\delta[a_s,x]}\ [\exp\{-y\}]^{1-\delta[a_s,x]} b_k[x] \\ &\overset{(e)}{=} \exp\{-y + J(y)\delta[a_s, x]\}\ b_k[x], \end{aligned} \tag{4.12}$$

where (a) follows from the definition of belief; (b) follows from Bayes' rule and $\propto$ denotes proportionality up to a normalization factor independent of x; (c) follows from the facts that Y_k is independent of history $\mathcal{H}_k$ given (X, A_k), and X is independent of action A_k given $\mathcal{H}_k$, and by the definition of belief b_k; (d-e) follow by substitution of the pdf of Y_k given in (4.5) and by definition of $J(y)$. We prove the lemma using induction. The lemma holds for b_0 by definition of m_0. Let $0 \leq k \leq L-1$ and b_k be given by (4.9), then using (4.12)(e) normalized to sum to one, we get

$$\begin{aligned} b_{k+1}[x] &= \frac{\exp\{J(y)\delta[a_s, x]\} \exp(m_k[x]\}}{\sum_{l\in\mathcal{I}} \exp\{J(y)\delta[a_s, l]\} \exp\{m_k[l]\}} \\ &= \frac{\exp\{m_{k+1}[x]\}}{\sum_{l\in\mathcal{I}} \exp\{m_{k+1}[l]\}}, \end{aligned} \tag{4.13}$$

where $m_{k+1}[x]$ is given by (4.10). □

Let $\boldsymbol{m}_k \triangleq [m_k[1], \ldots m_k[|\mathcal{I}|]$. Then, the previous lemma demonstrates that $\boldsymbol{m}_k$ is a sufficient statistic for control decisions, since it is sufficient for computing the belief b_k at time k. Therefore, μ can be expressed as $A_k = \mu_k(\boldsymbol{m}_k)$, $\forall 0 \leq k \leq L$, which maps the current preference vector $\boldsymbol{m}_k$ to beam index $A_k \in \mathcal{I}$. This result makes it possible to achieve an efficient implementation, since the belief can be updated according to simple preference update rules as in (4.10), rather than via complex Bayesian belief updates. In the subsequent analysis, we will use $\boldsymbol{m}_k$ rather than b_k as the state.

4.2.1 MDP Formulation

Thanks to the identification of the sufficient statistic $\boldsymbol{m}_k$, we model the optimization problem **P1** as a Markov decision process (MDP) and optimize the decision variables to maximize the alignment probability in the data-communication phase. The MPD is a 5-tuple $\langle \mathcal{T}, \mathcal{S}, \mathcal{I}, f(\boldsymbol{m}_{k+1}|\boldsymbol{m}_k, a_k), r_k(\boldsymbol{m}_k, a_k), \forall k{\in}\mathcal{T} \rangle$, with elements described as follows.

Time Horizon: given as $\mathcal{T}{=}\{0, 1, \ldots, L\}$ where $\mathcal{T}_{\rm BA}{\equiv}\mathcal{T}\backslash\{L\}$ denote the slot indices associated with the beam-alignment phase, whereas at $k{=}L$, the communication parameters are selected and used until the end of the frame.

State space: given as $\mathcal{S} = \mathbb{R}^{|\mathcal{I}|}$, i.e., all possible values of preference vectors $\boldsymbol{m}_k$.

Action space: the set containing all the beam indices, $\mathcal{I}$.

State transition distribution: Given state $\boldsymbol{m}_k = \boldsymbol{m}$ and action $A_k = a_{\rm s}$ used in the kth stage of the beam-alignment phase, the feedback $Y_k = y$ is generated with pdf

$$\begin{aligned} f(y|\boldsymbol{m}, a_{\rm s}) &\triangleq \sum_{x\in\mathcal{I}} f(Y_k = y|X = x, A_k = a_{\rm s}) b_k[x] \\ &= \frac{\exp\{m[a_{\rm s}]\}}{\sum_{l\in\mathcal{I}}\exp\{m[l]\}} \nu e^{-\nu y} + \left[1 - \frac{\exp\{m[a_{\rm s}]\}}{\sum_{l\in\mathcal{I}}\exp\{m[l]\}}\right] e^{-y}, \end{aligned} \tag{4.14}$$

leading to the new state

$$\boldsymbol{m}_{k+1} = \boldsymbol{m} + J(y)\boldsymbol{\delta}[a_{\rm s}], \tag{4.15}$$

where $\boldsymbol{\delta}[a_{\rm s}]{=}[\delta[a_{\rm s}, x]]_{\forall x\in\mathcal{I}}$ is the vector with entries $\delta[a_{\rm s}, x]$.

Reward function: the reward is the probability of choosing a beam index such that $A_{\rm d} = X$ in the data communication phase, so that correct alignment is achieved, yielding

$$r_k(\boldsymbol{m}, a) = \begin{cases} 0, & k \in \mathcal{T}_{\rm BA}, \\ \frac{\exp\{m[a]\}}{\sum_{l\in\mathcal{I}}\exp\{m[l]\}}, & k = L. \end{cases} \tag{4.16}$$

We now formulate the value function iteration for the MDP.

4.2.2 Value Function

The value function under the optimal policy is given as

$$V_k^*(\boldsymbol{m}) = \max_{a_\mathrm{s}\in\mathcal{I}} q_k(\boldsymbol{m}, a_\mathrm{s}), \tag{4.17}$$

where q_k is the Q-function under the state-action pair $(\boldsymbol{m}, a)$, defined recursively as

$$q_L(\boldsymbol{m}, A_\mathrm{d}) = r_L(\boldsymbol{m}, A_\mathrm{d}) = \frac{\exp\{m[A_\mathrm{d}]\}}{\sum_{l\in\mathcal{I}}\exp\{m[l]\}},$$

and for $k \in \mathcal{T}_\mathrm{BA}$, using (4.14),

$$\begin{aligned} q_k(\boldsymbol{m}, a_\mathrm{s}) &= \int_{\mathbb{R}^{|\mathcal{I}|}} V_{k+1}^*(\boldsymbol{m}') f(\boldsymbol{m}_{k+1}{=}\boldsymbol{m}' | \boldsymbol{m}_k{=}\boldsymbol{m}, A_k{=}a_\mathrm{s})\mathrm{d}\boldsymbol{m}' \\ &= \int_0^\infty V_{k+1}^*(\boldsymbol{m} + J(y)\boldsymbol{\delta}[a_\mathrm{s}]) f(y|\boldsymbol{m}, a_\mathrm{s})\mathrm{d}y. \end{aligned} \tag{4.18}$$

This yields the optimal value function in the data communication phase, by choosing the beam index with maximum preference ${A_\mathrm{d}}^* = \arg\max_{A_\mathrm{d}\in\mathcal{I}} m[A_\mathrm{d}]$,

$$V_L^*(\boldsymbol{m}) = \max_{A_\mathrm{d}\in\mathcal{I}} q_L(\boldsymbol{m}, A_\mathrm{d}) = \frac{\exp\{m[{A_\mathrm{d}}^*]\}}{\sum_{l\in\mathcal{I}}\exp\{m[l]\}}. \tag{4.19}$$

In the beam-alignment phase ($k \in \mathcal{T}_\mathrm{BA}$), combining (4.17) and (4.18), we obtain iteratively the value function as

$$V_k^*(\boldsymbol{m}) = \max_{a_\mathrm{s}\in\mathcal{I}} \int_0^\infty V_{k+1}^*(\boldsymbol{m} + J(y)\boldsymbol{\delta}[a_\mathrm{s}]) f(y|\boldsymbol{m}, a_\mathrm{s})\mathrm{d}y.$$

In the following theorem, whose proof is provided in the Appendix, we unveil structural properties of $V_k^*(\boldsymbol{m})$. We find a lower-bound and an upper-bound to the Q-function and show that these bounds are optimized by a policy which, in each stage of the beam-alignment phase, selects the beam index with the *second-best* preference. This result will be the basis for our proposed policy evaluated numerically in Sec. 4.3.

Theorem 4.1. *For $k \in \mathcal{T}_{\mathrm{BA}}$, the Q-function is bounded as*

$$q_k(\boldsymbol{m}, a_{\mathrm{s}}) \geq q_k^{LB}(\boldsymbol{m}, a_{\mathrm{s}}) \triangleq \frac{1}{\sum_{l\in\mathcal{I}} \exp\{m[l]\}} \Bigg[\xi(a_{\mathrm{s}}; \boldsymbol{m}) \tag{4.20}$$
$$+ \exp\left\{ \frac{\min_{x_i \neq x_j} m[x_i] - \nu m[x_j]}{1-\nu} \right\} h(\nu) \frac{g(\nu) - [g(\nu)]^{L-k}}{1-g(\nu)} \Bigg],$$
$$q_k(\boldsymbol{m}, a_{\mathrm{s}}) \leq q_k^{UB}(\boldsymbol{m}, a_{\mathrm{s}}) \triangleq \frac{[1+h(\nu)]^{L-k-1}}{\sum_{l\in\mathcal{I}} \exp\{m[l]\}} \xi(a_{\mathrm{s}}; \boldsymbol{m}), \tag{4.21}$$

where we have defined

$$\xi(a_{\mathrm{s}}; \boldsymbol{m}) \triangleq \begin{cases} \exp\{m[a_{\mathrm{s}}]\}, & \text{if } \max_{\hat{a}\neq a_{\mathrm{s}}} m[\hat{a}] - m[a_{\mathrm{s}}] < \ln \nu, \\ \exp\{\max_{\hat{a}\neq a_{\mathrm{s}}} m[\hat{a}]\} & \\ \quad + h(\nu) \exp\left\{ \frac{m[a_{\mathrm{s}}] - \nu \max_{\hat{a}\neq a_{\mathrm{s}}} m[\hat{a}]}{1-\nu} \right\}, & \text{otherwise,} \end{cases} \tag{4.22}$$

where

$$h(\nu) \triangleq \exp\left\{ \frac{\nu}{1-\nu} \ln \nu \right\} - \exp\left\{ \frac{\ln \nu}{1-\nu} \right\} > 0, \tag{4.23}$$
$$g(\nu) \triangleq \exp\left\{ \frac{\ln \nu}{1-\nu} \right\} \left[\frac{1}{\nu+1} - \frac{\ln \nu}{1-\nu} \right] > 0. \tag{4.24}$$

Let $x_{[1]}, x_{[2]}, \ldots, x_{[|\mathcal{I}|]}$ be an ordering of beam indices in decreasing order of preference, i.e., $m[x_{[1]}] \geq m[x_{[2]}] \geq \cdots, m[x_{[|\mathcal{I}|]}]$, then the optimal value function is bounded as

$$V_k^*(\boldsymbol{m}) \geq \max_{a_{\mathrm{s}}\in\mathcal{I}} q_k^{LB}(\boldsymbol{m}, a_{\mathrm{s}}) = q_k^{LB}(\boldsymbol{m}, x_{[2]}), \ \forall k \in \mathcal{T}_{\mathrm{BA}}, \tag{4.25}$$
$$V_k^*(\boldsymbol{m}) \leq \max_{a_{\mathrm{s}}\in\mathcal{I}} q_k^{UB}(\boldsymbol{m}, a_{\mathrm{s}}) = q_k^{UB}(\boldsymbol{m}, x_{[2]}), \ \forall k \in \mathcal{T}_{\mathrm{BA}}, \tag{4.26}$$

with the maximizer of q_k^{UB} and q_k^{LB} given by the second-best *beam index $x_{[2]}$.*

Proof. The proof is provided in the Appendix 8.F. □

As a result of this Theorem, both the upper and lower bounds of the Q-function are maximized by the *second-best* beam index policy, which selects the beam index with the

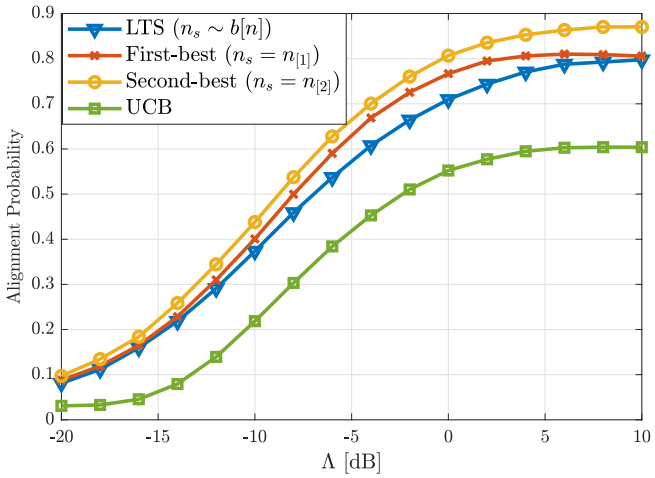

Figure 4.2. Alignment Probability vs Λ; $L = 32$ (beam-alignment takes 16% of frame duration).

second-best preference during the beam-alignment phase. This policy will be evaluated numerically in the next section, against other MAB-based schemes proposed in the literature.

4.3 Numerical Results

In this section, we evaluate the performance of the *second-best* beam index selection scheme (a_{s}=$x_{[2]}$) with analog beamforming, and compare it with three other schemes. The first one is based on LTS, a popular MAB scheme [35]. In LTS, at each slot the action is chosen according to the belief distribution, i.e., $a_{\mathrm{s}}\sim b[x]$. The second scheme is based on scanning the most-likely beam index (a_{s}=$x_{[1]}$) as proposed in [34] (first-best). The third scheme is based on UCB as proposed in [33]. We evaluate the performance of these three schemes in terms of the probability of alignment and spectral efficiency using Monte-Carlo simulation with 10^5 iterations for each simulated point, with parameters as follows: M_{t}=128, M_{r}=1, N_0=−174dBm/Hz, W_{tot}=200MHz, T_{fr}=20ms, T_s=0.1ms, f_c=30GHz, d=10m, [path loss exponent]=2. The BS uses M_{t}=128 antennas and partitions the AoD space into 32 sectors, each with a beamwidth of $\pi/32$rad and with uniform prior $b_0[x] = 1/32$, $\forall x \in \mathcal{I}$; the UE is isotropic, hence it uses M_{r}=1 antenna with a single sector. We use the beamforming design proposed in [27] for ULAs with antenna spacing

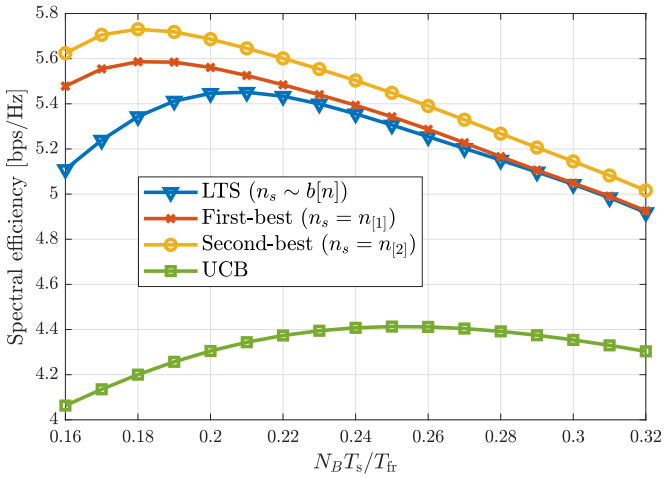

Figure 4.3. Spectral efficiency vs fraction of T_{fr} used for BA $LT_{\text{s}}/T_{\text{fr}}$.

$d_{\text{t}}=\lambda/2$. With this configuration, the main-lobe and side-lobe gains are best approximated by $G \approx 14\text{dB}$, $g \approx -11\text{dB}$.

In Fig. 4.2, we depict the probability of alignment achieved by the aforementioned schemes versus the pre-beamforming SNR Λ. It can be observed that second-best has better performance than the other three schemes, with up to 7%, 10%, and 30% performance gains compared to first-best, LTS-based and UCB-based schemes. The performance gain of second-best is attributed to a better exploration-exploitation trade-off. The first-best scheme suffers from poor exploration since it "greedily" chooses the beam index most likely to succeed, but fails to test other beams that may be under-explored, and is thus prone to make alignment errors. On the other hand, LTS-based scheme suffers from poor exploitation since it may scan least likely beams. The proposed *second-best* scheme, on the other hand, strikes a favorable trade-off between exploration and exploitation: instead of greedily choosing the most likely beam, it chooses the second most likely one, leading to better exploration than *first-best*; simultaneously, by not choosing beam pairs that are unlikely to succeed, it leads to a better exploitation compared to the LTS-based and UCB-based schemes. Finally, compared to UCB, second-best is better tailored to the structure of the model, since it aims to maximize the alignment probability *at the end* of the beam-alignment phase (see (4.16)), rather than the surrogate metric of UCB – the cumulative SNR accrued *during* beam-alignment.

In Fig. 4.3, we depict the spectral efficiency against the fraction of T_{fr} used for BA $LT_{\text{s}}/T_{\text{fr}}$. We fix the SNR for beam-alignment as $\Lambda = 0$dB and the data-communication power as P_{d}^{*}=22dBm. Similar to Fig. 4.2, *second-best* outperforms the three other schemes, owing to improved alignment. The spectral efficiency is maximized at a unique maximizer L^{*}: it increases initially with $L \leq L^{*}$ as the beam-alignment probability improves with L. However, as L increases beyond L^{*}, this gain is offset by the increased overhead and reduced duration of the data communication phase.

5. MOBILITY AND BLOCKAGE-AWARE COMMUNICATIONS IN MILLIMETER-WAVE VEHICULAR NETWORKS

Mobility may degrade the performance of next-generation vehicular networks operating at the millimeter-wave spectrum: frequent mis-alignment and blockages require repeated beam-training and handover, with enormous overhead. Nevertheless, mobility induces temporal correlations in the communication beams and in blockage events. In this chapter, an adaptive design is proposed, that learns and exploits these temporal correlations to reduce the beam-training overhead and make handover decisions. At each time-slot, the serving base station (BS) decides to perform either beam-training, data communication, or handover, under uncertainty in the system state. The decision problem is cast as a partially observable Markov decision process, with the goal to maximize the throughput delivered to the user, under an average power constraint. To address the high-dimensional optimization, an approximate *constrained point-based value iteration* (C-PBVI) method is developed, which simultaneously optimizes the primal and dual functions to meet the power constraint. Numerical results demonstrate a good match between the analysis and a simulation based on 2D mobility and 3D analog beamforming via uniform planar arrays at both BSs and UE, and reveal that C-PBVI performs near-optimally, and outperforms a baseline scheme with periodic beam-training by 38% in spectral efficiency. Motivated by the structure of C-PBVI, two heuristics are proposed, that trade complexity with sub-optimality, and achieve only 4% and 15% loss in spectral efficiency. Finally, the effect of mobility and multiple users on blockage dynamics is evaluated numerically, demonstrating superior performance over the baseline scheme.

5.1 System Model

We consider the scenario of Fig. 5.1, where multiple base stations (BSs) serve user equipments (UEs) moving along a road. At any time, each UE is associated with one BS – the *serving BS*. Each UE and the serving BS use beamforming with large antenna arrays

↑A version of this chapter was previously published by IEEE Transactions on Vehicular Technology [7], [67]

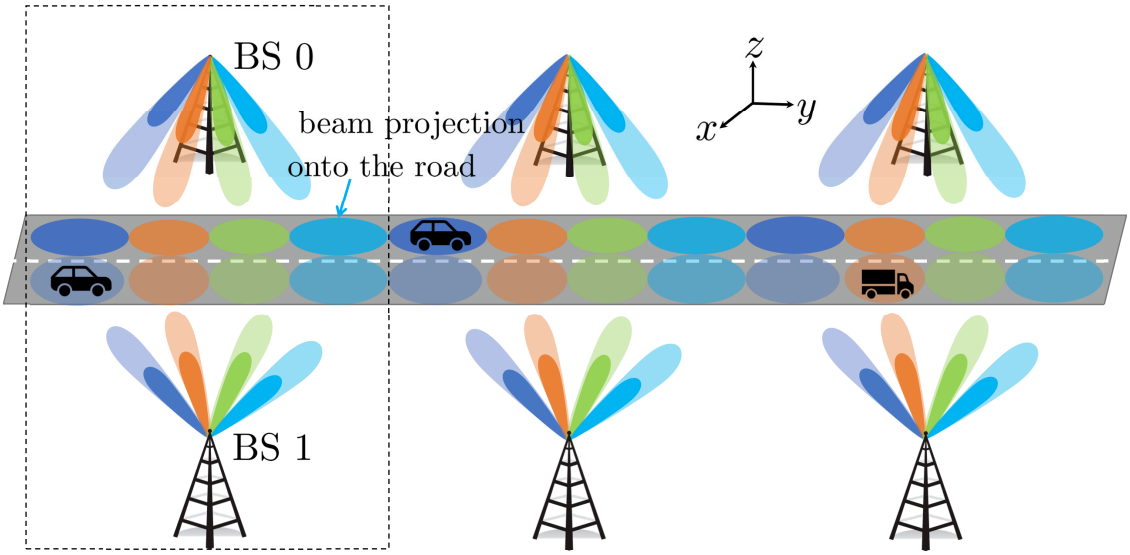

Figure 5.1. A cell deployment with BSs on both side of the road.

to achieve directional data transmission (DT); they use beam-training (BT) to maintain alignment. The communication links are subject to time-varying blockages, which cause the signal quality to drop abruptly and DT to fail. As soon as the serving BS detects blockage, it may decide to perform handover (HO) to the BS on the other side of the road, which then continues the process of BT/DT/HO, until either another blockage event is detected, or the UE exits the coverage area of the two BSs.

In this work, we focus on a specific segment of the road link covered by a pair of BSs and a single UE,[1] as depicted in the framed area of Fig. 5.1. Within this segment, the BT/DT/HO process continues until the UE exits the coverage region of the two BSs, denoted by the area $\mathcal{X} \subset \mathbb{R}^2$. In this context, we investigate the design of the BT/DT/HO strategy during a transmission episode, defined as the time interval between the two instants when the UE enters and exits the coverage area of the two BSs. The goal is to maximize the average throughput delivered to the UE subject to an average power constraint. Note that, when the episode terminates, the UE enters the coverage area of another pair of BSs, and the same analysis may be applied to each segment traversed.

[1]↑The proposed system model and techniques can be applied to a multi-user scenario by partitioning the BS resources using orthogonal frequency division multiple access (OFDMA) and multiple RF chains or time division duplexing (TDD)[68].

Time is discretized into time-slots of duration Δ_{t}, corresponding to the transmission of a beacon signal during BT or of a data fragment during DT. Next, we describe the signal, channel and UE mobility and blockage dynamics models used throughout the paper.

5.1.1 Signal and Channel Models

Let $I \in \{0,1\} \triangleq \mathcal{I}$ denote the index of the serving BS at time k. Let $\mathbf{x}_k \in \mathbb{C}^L$ be the transmitted signal with $\mathbb{E}[\|\mathbf{x}_k\|_2^2]=L$, where L denotes the number of symbols transmitted. The received signal at the UE is expressed as

$$\mathbf{y}_k = \sqrt{P_k}\mathbf{f}_k^H \mathbf{H}_k^{(I)} \mathbf{c}_k \mathbf{x}_k + \mathbf{w}_k, \tag{5.1}$$

where P_k is the average transmit power of the serving BS I; $\mathbf{c}_k \in \mathbb{C}^{M_{\mathrm{tx}}^{(I)} \times 1}$ and $\mathbf{f}_k \in \mathbb{C}^{M_{\mathrm{rx}} \times 1}$ are unit-norm beamforming vectors with $M_{\mathrm{tx}}^{(I)}$ and M_{rx} antenna elements at BS I and the reference UE, respectively; $\mathbf{H}_k^{(I)} \in \mathbb{C}^{M_{\mathrm{rx}} \times M_{\mathrm{tx}}^{(I)}}$ is the channel matrix; $\mathbf{w}_k \sim \mathcal{CN}(0, \sigma_w^2 \mathbf{I})$ with $\sigma_w^2 = (1+F) N_0 W_{\mathrm{tot}}$ is additive white Gaussian noise, N_0 is the noise power spectral density, W_{tot} is the signal bandwidth, F is the receiver noise figure.

In this chapter, we model $\mathbf{H}_k^{(I)}$ as a single line of sight (LOS) path with binary blockage [59] and diffuse multipath [69],

$$\mathbf{H}_k^{(I)} = \underbrace{\sqrt{M_{\mathrm{tx}}^{(I)} M_{\mathrm{rx}}} B_k^{(I)} h_k^{(I)} \mathbf{d}_{\mathrm{rx}}(\theta^{(I)}(X_k)) \mathbf{d}_{\mathrm{tx}}^{(I)}(\phi^{(I)}(X_k))^H}_{\mathbf{H}_{k,\mathrm{LOS}}^{(I)}} + \underbrace{\sum_{l=1}^{N_{\mathrm{DIF}}} \sqrt{M_{\mathrm{tx}}^{(I)} M_{\mathrm{rx}}} \tilde{h}_{k,l}^{(I)} \mathbf{d}_{\mathrm{rx}}(\tilde{\theta}_{k,l}^{(I)}) \mathbf{d}_{\mathrm{tx}}^{(I)}(\tilde{\phi}_{k,l}^{(I)})^H}_{\mathbf{H}_{k,\mathrm{DIF}}^{(I)}},$$

where $B_k^{(I)} \in \{0,1\}$ denotes the binary blockage variable of BS I, equal to 1 if the LOS path is unobstructed, equal to 0 otherwise; $\mathbf{d}_{\mathrm{tx}}^{(I)}(\phi) \in \mathbb{C}^{M_{\mathrm{tx}}^{(I)}}$ and $\mathbf{d}_{\mathrm{rx}}(\theta) \in \mathbb{C}^{M_{\mathrm{rx}}}$ are the unit-norm array response vectors of BS I and UE, as a function of the AoD ϕ and AoA θ (note that these include both azimuth and elevation information for UPAs); $\phi^{(I)}(X_k)$ and $\theta^{(I)}(X_k)$ are the AoD and AoA of the LOS path with respect to BS I and the UE in posi-

tion $X_k \in \mathcal{X}$;[2] $h_k^{(I)} \sim \mathcal{CN}(0, \sigma_{h,I}^2)$ is the complex channel gain of the LOS component, i.i.d. over slots, with $\sigma_{h,I}^2 = 1/\ell(d_I(X_k))$; $\ell(d_I(X_k)) = (4\pi d_I(X_k)/\lambda_c)^2$ denotes the pathloss as function of the BS I-UE distance $d_I(X_k)$; $\lambda_c = c/f_c$ is the wavelength. Finally, $\mathbf{H}_{k,\mathrm{DIF}}^{(I)}$ denotes the channel corresponding to diffuse multipath components with coefficients $\tilde{h}_{k,l}$, AoD $\tilde{\phi}_{k,l}^{(I)}$ and AoA $\tilde{\theta}_{k,l}^{(I)}$; we model $\mathbf{H}_{k,\mathrm{DIF}}^{(I)}$ as zero-mean complex Gaussian, with i.i.d. entries (over time and over antennas), each with variance $\sigma_{\mathrm{DIF},I}^2$. These components have been shown to be much weaker than the LOS path (up to 100× weaker at a BS-UE distance of only 10 meters [59]), so that $\sigma_{\mathrm{DIF},I}^2 \ll \sigma_{h,I}^2$. Then, letting $G_{\mathrm{tx}}^{(I)}(\mathbf{c}_k, x) = M_{\mathrm{tx}}^{(I)} |\mathbf{d}_{\mathrm{tx}}^{(I)}(\phi^{(I)}(x))^H \mathbf{c}_k|^2$ and $G_{\mathrm{rx}}(\mathbf{f}_k, x) = M_{\mathrm{rx}} |\mathbf{d}_{\mathrm{rx}}(\theta^I(x))^H \mathbf{f}_k|^2$ be the beamforming gains of the serving BS I and UE, respectively, with respect to the LOS path, and $\Theta_k = \angle \mathbf{d}_{\mathrm{tx}}^{(I)}(\phi^{(I)}(X_k))^H \mathbf{c}_k + \angle \mathbf{f}_k^H \mathbf{d}_{\mathrm{rx}}(\theta^{(I)}(X_k))$ be the unknown phase of the overall gain, the signal received at the UE can be expressed as

$$\mathbf{y}_k = \sqrt{P_k} \left[B_k^{(I)} h_k^{(I)} \sqrt{G_{\mathrm{tx}}^{(I)}(\mathbf{c}_k, X_k) G_{\mathrm{rx}}(\mathbf{f}_k, X_k)} e^{j\Theta_k} + \Omega_k^{(I)} \right] \mathbf{x}_k + \mathbf{w}_k, \tag{5.2}$$

where $\Omega_k^{(I)} \triangleq \mathbf{f}_k^H \mathbf{H}_{k,\mathrm{DIF}}^{(I)} \mathbf{c}_k \sim \mathcal{CN}(0, \sigma_{\mathrm{DIF},I}^2)$ is the contribution due to the diffuse multipath channel components. The SNR averaged over the fading coefficients is then given as

$$\mathrm{SNR}_k = \frac{P_k}{\sigma_w^2} \left[B_k^{(I)} \frac{G_{\mathrm{tx}}^{(I)}(\mathbf{c}_k, X_k) G_{\mathrm{rx}}(\mathbf{f}_k, X_k)}{\ell(d_I(X_k))} + \sigma_{\mathrm{DIF},I}^2 \right]. \tag{5.3}$$

5.1.2 Codebook Structure

Each BS has a codebook of beamformers to cover the intended coverage region $\mathcal{X}$ on the road. The beamforming codebook of BS I is denoted by $\mathcal{C}_I \triangleq \{\mathbf{c}_{I,1}, \ldots, \mathbf{c}_{I,|\mathcal{C}_I|}\}$. The UE uses the codebook $\mathcal{F} \triangleq \{\mathbf{f}_1, \ldots, \mathbf{f}_{|\mathcal{F}|}\}$. Let $\mathcal{V}_I \triangleq \mathcal{C}_I \times \mathcal{F}$ denote the joint codebook containing all possible beamforming codeword pairs of BS I and UE. We index these codeword pairs by the *beam pair index* (BPI), with values in $\bar{\mathcal{S}}_I \triangleq \{1, 2, \ldots, |\mathcal{C}_I||\mathcal{F}|\}$; let $(\mathbf{c}_I^{(j)}, \mathbf{f}_I^{(j)})$ be the jth such pair, with $j \in \bar{\mathcal{S}}_I$. With this definition, note that, if the UE is in position $X_k = x$ and is

[2] ↑Note that the AoA $\theta^{(I)}(X_k)$ should also depend on the angle of rotation (azimuth and elevation) of the antenna array of the UE; herein, we assume that it only depends on the UE position X_k. This is a good approximation in vehicular networks, where the antenna array may be mounted on the rooftop of the vehicle; the more general case with non-fixed array orientation can be addressed by including the angle of rotation information in the AoA, which may be estimated using a gyroscope sensor [70].

being served by BS I, then the maximum beamforming gain is achieved with the strongest BPI (SBPI), which also yields the maximum SNR in (5.3), defined as

$$s_I^*(x) \triangleq \arg\max_{j \in \bar{\mathcal{S}}_I} G_{\text{tx}}^{(I)}(\mathbf{c}_I^{(j)}, x) G_{\text{rx}}(\mathbf{f}_I^{(j)}, x). \tag{5.4}$$

Let $\mathcal{S}_I \triangleq \{s_I^*(x) : x \in \mathcal{X}\} \subseteq \bar{\mathcal{S}}_I$ be the set of SBPIs across all possible UE positions. Note that this set can be constructed over time utilizing the feedback from the UE and excluding the BPIs that do not yield significant signal power [22]. It follows that the directional communication between BS I and UE can be achieved by restricting the choice of beamforming codewords to the optimal set $\mathcal{S}_I$, since any other beam pair achieves lower SNR. This can be obtained using a coordinated beamforming strategy where, before start of BT or DT, the serving BS I and UE coordinate to select a subset of BPIs from the set $\mathcal{S}_I$ to be scanned synchronously during BT or used for DT, as explained in Section 5.1.5.

5.1.3 Mobility and Blockage Dynamics

Note that, to achieve directional communication, the pair of BS I and UE should detect the SBPI $s_I^*(X_k)$ via beam-training – a source of severe overhead; the mobility of the UE along the road induces temporally correlated dynamics on the SBPI $s_I^*(X_k)$, which may be exploited to reduce the training overhead via POMDPs. Similarly, the blockage state exhibits temporal and spatial correlations, which can be exploited to efficiently detect/predict blockages and perform HO if needed. To define such POMDP model, we now define a Markov model on the SBPI and blockage states, induced by the UE mobility. Let $S_k = (s_0^*(X_k), s_1^*(X_k))$ be the pair of SBPIs at both BSs, taking values from $\mathcal{S} \triangleq \{(s_0^*(x), s_1^*(x)) : x \in \mathcal{X}\}$. Let $B_k \triangleq (B_k^{(0)}, B_k^{(1)}) \in \{0,1\}^2$ be the pair of binary blockage states with $B_k^{(I)}$ denoting the

blockage with respect to BS I. Then, the one-step transition probability of (S_k, B_k) is expressed as

$$\begin{aligned}\mathbf{P}_{s'b'|sb} &\triangleq \mathbb{P}(S_{k+1}=s', B_{k+1}=b' | S_k=s, B_k=b) \\ &= \underbrace{\mathbb{P}(S_{k+1}=s'|S_k=s)}_{\mathbf{S}_{s'|s}} \underbrace{\mathbb{P}(B_{k+1}=b'|B_k=b, S_k=s, S_{k+1}=s')}_{\mathbf{B}_{b'|bss'}}.\end{aligned} \tag{5.5}$$

Here, it is assumed that the next SBPI S_{k+1} is independent of the current blockage state B_k, given the current beam index pair S_k (indeed, the dynamics of SBPI depend solely on UE mobility). Note that $\sum_{s',b'} \mathbf{P}_{s'b'|sb} \leq 1$, since the UE might exit the coverage area of the two BSs. In practice, (5.5) can be estimated based on estimated time-series of SBPI and blockage pairs, $\{(\hat{s}_k, \hat{b}_k, \hat{s}_{k+1}, \hat{b}_{k+1}), k \in T_{\text{sound}}\}$, which in turn may be acquired at times $k \in T_{\text{sound}}$ via exhaustive search beam-training methods. Based on these time-series, the BSs can estimate the transition probabilities in (5.5) as

$$\hat{\mathbf{S}}_{s'|s} = \frac{\sum_{k \in T_{\text{sound}}} \chi(\hat{s}_k = s, \hat{s}_{k+1} = s')}{\sum_{k \in T_{\text{sound}}} \chi(\hat{s}_k = s)}, \tag{5.6}$$

$$\hat{\mathbf{B}}_{b'|bss'} = \frac{\sum_{k \in T_{\text{sound}}} \chi(\hat{s}_k = s, \hat{B}_k = b, \hat{s}_{k+1} = s', \hat{B}_{k+1} = b')}{\sum_{k \in T_{\text{sound}}} \chi(\hat{s}_k = s, \hat{B}_k = b, \hat{s}_{k+1} = s')}, \tag{5.7}$$

where $\chi(\cdot)$ is the indicator function. Note that the estimates $\hat{\mathbf{S}}_{s'|s}$ and $\hat{\mathbf{B}}_{b'|bss'}$ can be improved over time as more samples of $(\hat{s}_k, \hat{b}_k, \hat{s}_{k+1}, \hat{b}_{k+1})$ become available. This approach does not require a dedicated learning phase; instead, estimated time-series can be collected based on beam-training and data communication feedback, so that the estimation overhead is minimal. Following their updates, the proposed policies can be updated accordingly. As more and more samples are collected, the estimation accuracy improves, leading to policies that more optimally leverage the mobility and blockage dynamics within the environment, yielding a more efficient use of resources.

5.1.4 Sectored antenna model

In this chapter, we use the *sectored antenna model* to approximate the beamforming gain, as also used in [4], [41]. As we will show in Section 5.5, when coupled with an appropriate design of the BSs beamforming codebooks $\mathcal{C}_I, I \in \mathcal{I}$ and of the UE beamforming codebook $\mathcal{F}$ [27], the sectored model provides an accurate and analytically tractable approximation of the actual beamforming gain. Consider the BPI $j \in \mathcal{S}_I$ and let $G^{(I)}(j,x) \triangleq G^{(I)}_{\text{tx}}(\mathbf{c}^{(j)}_I, x) G^{(I)}_{\text{rx}}(\mathbf{f}^{(j)}_I, x)$ be the overall gain between BS I and UE position x, under the beamforming codeword pair $(\mathbf{c}^{(j)}_I, \mathbf{f}^{(j)}_I)$. Under the sectored model, if the UE is aligned with BS I under the BPI j, i.e., its position x is such that the SBPI $s^*_I(x) = j$, then the *aligned gain* satisfies $G^{(I)}(j,x) \gg 1$ with gain-to-pathloss ratio $G^{(I)}(j,x)/\ell(d^{(I)}(x)) \approx \Upsilon^{(j)}_I, \forall x : j = s^*_I(x)$. On the other hand, if the UE is mis-aligned with BS I under the BPI j, i.e., $s^*_I(x) \neq j$, then the mis-aligned beamforming gain of BPI $j \in \mathcal{S}_I$ is such that $G^{(I)}(j,x) \approx g^{(I)}_j \ll 1, \forall x : j \neq s^*_I(x)$ (i.e., it is small and equal to the sidelobe gain $g^{(I)}_j$ for all positions x such that j is not the SBPI). Based on this model, we now derive expressions for the transmission power to achieve a target SNR at the receiver. We denote the case with the aligned beam pair and no blockage ($j = s^*_I(x)$ and $b_I = 1$) as "active SBPI" and the complementary case of blockage or UE in the sidelobe ($j \neq s^*_I(x)$ or $b_I = 0$) as "inactive SBPI". In the case of active SBPI, from (5.3) we have

$$\text{SNR}_{\text{act}} = \frac{P^{(I)}_j}{\sigma^2_w} \left[\Upsilon^{(I)}_j + \sigma^2_{\text{DIF},I} \right] \Leftrightarrow P^{(I)}_j = \frac{\sigma^2_w \text{SNR}_{\text{act}}}{\Upsilon^{(I)}_j + \sigma^2_{\text{DIF},I}}, \tag{5.8}$$

which yields the transmission power to achieve a target SNR equal to SNR_{act} in case of active SBPI. In the case of inactive SBPI, we can express the SNR in (5.3) using (5.8) as

$$\begin{aligned} \text{SNR}_{\text{iact}} &= \frac{P^{(I)}_j}{\sigma^2_w} \left[B^{(I)} \frac{G^{(I)}(j,x)}{\ell(d_I(x))} + \sigma^2_{\text{DIF},I} \right] \\ &= \left[B^{(I)} \frac{G^{(I)}(j,x)}{\ell(d_I(x))} + \sigma^2_{\text{DIF},I} \right] \frac{\text{SNR}_{\text{act}}}{\Upsilon^{(I)}_j + \sigma^2_{\text{DIF},I}}. \end{aligned} \tag{5.9}$$

Note that, to help the BS detect the inactive SBPI condition, this value of SNR should be as small as possible; for this reason, we determine the worst case SNR under inactive SBPI

by maximizing (5.9) over all possible blockage states $B^{(I)}\in\{0,1\}$, mis-aligned beam j and UE position $x\in\mathcal{X}$, as

$$\begin{aligned}&\mathrm{SNR}_{\mathrm{iact}}\\&\leq \max_{x\in\mathcal{X}}\max_{j\in\mathcal{S}_I\setminus\{s_I^*(x)\}}\left[B^{(I)}\frac{G^{(I)}(j,x)}{\ell(d_I(x))}+\sigma^2_{\mathrm{DIF},I}\right]\frac{\mathrm{SNR}_{\mathrm{act}}}{\Upsilon_j^{(I)}+\sigma^2_{\mathrm{DIF},I}}\\&\triangleq \rho_I\mathrm{SNR}_{\mathrm{act}}.\end{aligned}\tag{5.10}$$

In other words, to achieve a target $\mathrm{SNR}_{\mathrm{act}}$ within the mainlobe, the BS should transmit with power given by (5.8); however, if the signal is blocked or the UE receives on the sidelobe (or both), the associated worst-case SNR is $\rho_I\mathrm{SNR}_{\mathrm{act}}$.[3] In this case, data transmission is in outage since $\rho_I \ll 1$ (numerically, we found $\rho_I = -15$dB, $\forall I$ based on the setup of Section 5.5).

5.1.5 Beam-Training (BT) and Data Transmission (DT)

We now introduce the BT and DT operations.

BT phase: At the start of a BT phase, the serving BS I selects a set of BPIs $\mathcal{S}_{\mathrm{BT}}\subseteq\mathcal{S}_I$ over which the beacons $\mathbf{x}_k$ are sent, and a target SNR $\mathrm{SNR}_{\mathrm{BT}}$. The beacon transmission is done in sequence over $|\mathcal{S}_{\mathrm{BT}}|$ time-slots, using one slot for each BPI $j\in\mathcal{S}_{\mathrm{BT}}$, with the serving BS transmitting using the beamforming vector $\mathbf{c}_I^{(j)}$, and the UE synchronously receiving using the combining vector $\mathbf{f}_I^{(j)}$. Therefore, the duration of the BT phase is $T_{\mathrm{BT}}\triangleq|\mathcal{S}_{\mathrm{BT}}|+1$, including the last slot for feedback signaling from the UE to the BS. Let $i\in\{0,\ldots,T_{\mathrm{BT}}-2\}$ be the ith time-slot of the BT phase, and $j_i\in\mathcal{S}_{\mathrm{BT}}$ be the BPI scanned by the BS I and UE in this slot. The UE processes the received signal $\mathbf{y}_{k+i}$ with a matched filter,

$$\Gamma_{j_i}\triangleq\frac{|\mathbf{x}_{k+i}^H\mathbf{y}_{k+i}|^2}{(1+F)N_0W_{\mathrm{tot}}\|\mathbf{x}_{k+i}\|_2^2}.\tag{5.11}$$

[3]↑For the sake of analytical tractability, ρ_I (found by maximizing over $j\neq s_I^*(x)$) is the worst case over the BPI $j\in\mathcal{S}_I$. The model can be generalized to express the dependence of ρ_I on j, leading to a more complicated BT feedback analysis, possibly not in closed form.

Upon collecting the sequence $\{\Gamma_j, \forall j \in \mathcal{S}_{\rm BT}\}$, the UE generates the feedback signal

$$Y = \begin{cases} j^* \triangleq \arg\max_{j\in\mathcal{S}_{\rm BT}} \Gamma_j, & \max_{j\in\mathcal{S}_{\rm BT}} \Gamma_j > \eta_{\rm BT}^{(I)}, \\ \emptyset, & \max_{j\in\mathcal{S}_{\rm BT}} \Gamma_j \leq \eta_{\rm BT}^{(I)}. \end{cases} \tag{5.12}$$

In other words, if all the matched filter outputs are smaller than $\eta_{\rm BT}^{(I)}$, $Y{=}\emptyset$ indicates that no beam pair is deemed sufficient for data transmission, either due to blockage ($B_k^{(I)}{=}0$), or the UE receiving on the sidelobes of the BPIs $j \in \mathcal{S}_{\rm BT}$. Otherwise, $Y{=}j^*$ indicates the index of the strongest BPI detected.

We now perform a probabilistic analysis of feedback. To this end, let $S_I{=}s_I^*(X_k)$ and $B_I{=}B_k^{(I)}$ be the SBPI and blockage state under BS I at the beginning of the BT phase. We assume that these state variables do not change during the transmission of the beacon sequences, i.e., $s_I^*(X_{k+i}){=}S_I$ and $B_{k+i}^{(I)}{=}B_I, \forall i{\in}\{0,\ldots,T_{\rm BT}{-}2\}$. This is a reasonable assumption, since the duration of the BT phase ($\times$0.1ms) is typically much shorter than the time required by the UE to change beam ($\times$100ms) or the time-scales of blockage ($\times$100ms). With this assumption, given the state (S_I, B_I) of BS I during BT, the signal sequence $\{\Gamma_j, \forall j{\in}\mathcal{S}_{\rm BT}\}$ is independent across j, due to the i.i.d. nature of $h_{k+i}^{(I)}$, $\Omega_{k+i}^{(I)}$ and $\mathbf{w}_{k+i}$. In addition, in case of active SBPI ($S_I{=}j$ and $B_I{=}1$), by using (5.2) and (5.8), Γ_j has exponential distribution with mean $1{+}{\rm SNR}_{\rm BT}L$, $\Gamma_j{\sim}\mathcal{E}(1{+}{\rm SNR}_{\rm BT}L)$; otherwise (inactive SBPI, $S_I{\neq}j$ or $B_I{=}0$) $\Gamma_j{\sim}\mathcal{E}(1{+}\rho_I{\rm SNR}_{\rm BT}L)$. It follows that

$$\begin{cases} \Sigma_{I,1} \triangleq \mathbb{P}(\Gamma_j{\leq}\eta_{\rm BT}^{(I)} | S_I = j, B_I = 1) = 1 - e^{\frac{-\eta_{\rm BT}^{(I)}}{1+{\rm SNR}_{\rm BT}L}}, \\ \Sigma_{I,0} \triangleq \mathbb{P}(\Gamma_j{\leq}\eta_{\rm BT}^{(I)} | S_I \neq j \text{ or } B_I = 0){=}1{-}e^{\frac{-\eta_{\rm BT}^{(I)}}{1+\rho_I{\rm SNR}_{\rm BT}L}}. \end{cases}$$

Now, let us consider separately the two events $\{S_I \notin \mathcal{S}_{\rm BT}\} \cup \{B_I = 0\}$ ("inactive SBPI in $\mathcal{S}_{\rm BT}$") and $\{S_I \in \mathcal{S}_{\rm BT}\} \cap \{B_I = 1\}$ ("active SBPI $S_I \in \mathcal{S}_{\rm BT}$"). In case of inactive SBPI in

$\mathcal{S}_{\mathrm{BT}}$, the probability of generating the feedback signal $Y = \emptyset$ (i.e., of correctly detecting inactive SBPI within the $\mathcal{S}_{\mathrm{BT}}$ scanned in the BT phase) is

$$\mathbb{P}(Y = \emptyset|\text{inactive SBPI in } \mathcal{S}_{\mathrm{BT}}) = \prod_{j\in\mathcal{S}_{\mathrm{BT}}} \mathbb{P}(\Gamma_j \leq \eta_{\mathrm{BT}}^{(I)}|S_I \neq j \text{ or } B_I = 0) = \Sigma_{I,0}^{|\mathcal{S}_{\mathrm{BT}}|}, \quad (5.13)$$

since $Y{=}\emptyset$ is equivalent to $\Gamma_j \leq \eta_{\mathrm{BT}}^{(I)}, \forall j \in \mathcal{S}_{\mathrm{BT}}$, and Γ_j are independent across j, conditional on (S_I, B_I). Similarly, in case of active SBPI $S_I \in \mathcal{S}_{\mathrm{BT}}$, the probability of incorrectly detecting inactive SBPI is

$$\begin{aligned}
&\mathbb{P}(Y = \emptyset|\text{active SBPI } S_I \in \mathcal{S}_{\mathrm{BT}}) \\
&= \mathbb{P}(\Gamma_j{\leq}\eta_{\mathrm{BT}}^{(I)}|S_I{=}j, B_I{=}1) \prod_{j\in\mathcal{S}_{\mathrm{BT}}\setminus\{S_I\}} \mathbb{P}(\Gamma_j{\leq}\eta_{\mathrm{BT}}^{(I)}|S_I{\neq}j, B_I{=}1) \\
&= \Sigma_{I,1}\Sigma_{I,0}^{|\mathcal{S}_{\mathrm{BT}}|-1},
\end{aligned} \quad (5.14)$$

since S_I is the SBPI, implying $\Gamma_{s_I} \sim \mathcal{E}(1 + \mathrm{SNR}_{\mathrm{BT}}L)$.

In case of inactive SBPI in $\mathcal{S}_{\mathrm{BT}}$, the probability of generating the feedback signal $j^* \in \mathcal{S}_{\mathrm{BT}}$ (i.e., of incorrectly detecting an active SBPI) is

$$\mathbb{P}(Y = j^*|\text{inactive SBPI in } \mathcal{S}_{\mathrm{BT}}) = \frac{1}{|\mathcal{S}_{\mathrm{BT}}|}\Big[1{-}\mathbb{P}(Y{=}\emptyset|\text{inactive SBPI in } \mathcal{S}_{\mathrm{BT}})\Big]{=}\frac{1{-}\Sigma_{0,I}^{|\mathcal{S}_{\mathrm{BT}}|}}{|\mathcal{S}_{\mathrm{BT}}|}; \quad (5.15)$$

in fact, Γ_j are i.i.d. across beams, conditional on inactive SBPI, so that incorrect detections are uniform across the feedback outcomes $j^* \in \mathcal{S}_{\mathrm{BT}}$.

Instead, in case of active SBPI $S_I \in \mathcal{S}_{\mathrm{BT}}$, we need to further distinguish between the two

cases $j^*=S_I$ (the SBPI is detected correctly) and $j^*\in\mathcal{S}_{\rm BT}\backslash\{S_I\}$ (incorrect detection). The probability of correctly detecting the SBPI is found as

$$\begin{aligned}
&\mathbb{P}(Y=S_I|\text{active SBPI } S_I\in\mathcal{S}_{\rm BT})\\
&=\mathbb{P}(\Gamma_{S_I}>\eta_{\rm BT}^{(I)},\Gamma_{S_I}>\Gamma_j,\forall j\in\mathcal{S}_{\rm BT}\backslash\{S_I\}|\text{active SBPI } S_I\in\mathcal{S}_{\rm BT})\\
&=\int_{\eta_{\rm BT}^{(I)}}^{\infty}\Big[f(\Gamma_{S_I}=\tau|\text{active SBPI } S_I\in\mathcal{S}_{\rm BT})\prod_{j\in\mathcal{S}_{\rm BT}\backslash\{S_I\}}\mathbb{P}(\Gamma_j<\tau|S_I\neq j,B_I=1)\Big]{\rm d}\tau\\
&=\int_{\eta_{\rm BT}^{(I)}}^{\infty}\left[\frac{1}{1+{\rm SNR}_{\rm BT}L}\exp\left\{-\frac{\tau}{1+{\rm SNR}_{\rm BT}L}\right\}\left(1-\exp\left\{-\frac{\tau}{1+\rho_I{\rm SNR}_{\rm BT}L}\right\}\right)^{|\mathcal{S}_{\rm BT}|-1}\right]{\rm d}\tau\\
&=\sum_{n=0}^{|\mathcal{S}_{\rm BT}|-1}\binom{|\mathcal{S}_{\rm BT}|-1}{n}\frac{(-1)^n(1-\Sigma_{I,1})(1-\Sigma_{I,0})^n}{1+\frac{1+{\rm SNR}_{\rm BT}L}{1+\rho_I{\rm SNR}_{\rm BT}L}n},
\end{aligned}\tag{5.16}$$

where in the first step we used the definition of $Y{=}S_I$, i.e., Γ_{S_I} must be greater than the threshold $\eta_{\rm BT}^{(I)}$, and all other Γ_j must be smaller than Γ_{S_I}; in the last step, we used Newton's binomial theorem to solve the integral. Finally, the probability of incorrectly detecting the SBPI, $j^*\in\mathcal{S}_{\rm BT}\backslash\{S_I\}$ is

$$\mathbb{P}(Y=j^*|\text{active SBPI } S_I\in\mathcal{S}_{\rm BT}){=}\frac{1}{|\mathcal{S}_{\rm BT}|{-}1}\Big[1{-}\sum_{y\in\{S_I,\emptyset\}}\mathbb{P}(Y{=}y|\text{active SBPI } S_I{\in}\mathcal{S}_{\rm BT})\Big]\tag{5.17}$$

since, similarly to (5.15), erroneous detections are uniform across the remaining $|\mathcal{S}_{\rm BT}|{-}1$ beams.

Since $Y{=}\emptyset$ represents the fact that the inactive SBPI condition has been detected, we choose $\eta_{\rm BT}^{(I)}$ so that the misdetection and false alarm probabilities are both equal to $\delta_{\rm BT}$, yielding from (5.14)-(5.15) (over all $j{\in}\mathcal{S}_{\rm BT}$),

$$\delta_{\rm BT}^{(I)}=1{-}\Sigma_{I,0}^{|\mathcal{S}_{\rm BT}|}=\Sigma_{I,1}\Sigma_{I,0}^{|\mathcal{S}_{\rm BT}|-1}.\tag{5.18}$$

For a given ${\rm SNR}_{\rm BT}$ and $|\mathcal{S}_{\rm BT}|$, the value of $\eta_{\rm BT}^{(I)}$ and the corresponding $\delta_{\rm BT}^{(I)}$ can be found numerically using the bisection method, since the left- and right- hand sides of (5.18) are decreasing and increasing functions of $\eta_{\rm BT}^{(I)}$, respectively.

DT phase: At the start of the DT phase, the BS I chooses a BPI $j\in\mathcal{S}_I$ used for data transmission, along with the duration $T_{\rm DT}$ of the DT frame, the target average SNR at the receiver $\mathrm{SNR}_{\rm DT}$, and a target transmission rate $\bar{R}_{\rm DT}$; the last slot is used for the feedback signal from the UE to the BS, as described below. We assume that a fixed fraction $\kappa\in(0,1)$ out of L symbols in each slot is used for channel estimation. Consider slot $t\in\{k,\ldots,k+T_{\rm DT}-2\}$ of data communication; then, if $s_I^*(X_t)\neq j$ or $B_t^{(I)}=0$, i.e., the selected BPI j is inactive, then the communication is in outage; otherwise ($s_I^*(X_t)=j$ and $B_t^{(I)}=1$, i.e., the selected BPI j is an active SBPI) assuming that channel estimation errors are negligible compared to the noise level (achieved with a sufficiently long pilot sequence κL), from the signal model (5.2), we find that outage occurs if (note that $\mathbb{E}[|h_t^{(I)}|^2\ell(d_I(X_t))]=1$)

$$W_{\rm tot}\log_2(1+|h_t^{(I)}|^2\ell(d_I(X_t))\mathrm{SNR}_{\rm DT})<\bar{R}_{\rm DT}, \tag{5.19}$$

yielding the outage probability

$$\begin{aligned}\mathbb{P}_{\rm OUT}(\bar{R}_{\rm DT},\mathrm{SNR}_{\rm DT})&=\mathbb{P}\Big(|h_t^{(I)}|^2\ell(d_I(X_t))<\frac{2^{\bar{R}_{\rm DT}/W_{\rm tot}}-1}{\mathrm{SNR}_{\rm DT}}\Big)\\&=1-\exp\Big\{-\frac{2^{\bar{R}_{\rm DT}/W_{\rm tot}}-1}{\mathrm{SNR}_{\rm DT}}\Big\}.\end{aligned} \tag{5.20}$$

In this chapter, we design $\bar{R}_{\rm DT}$ based on the notion of $\epsilon-$outage capacity, i.e., $\bar{R}_{\rm DT}$ is the largest rate such that $\mathbb{P}_{\rm OUT}(\bar{R}_{\rm DT},\mathrm{SNR}_{\rm DT})\leq\epsilon$, for a target outage probability $\epsilon<1$. Imposing (5.20) equal to ϵ, this can be expressed as

$$\bar{R}_{\rm DT}=C_\epsilon(\mathrm{SNR}_{\rm DT})=W_{\rm tot}\log_2\left(1-\mathrm{SNR}_{\rm DT}\ln(1-\epsilon)\right). \tag{5.21}$$

With this choice, the transmission is successful with probability $1-\epsilon$, and the average rate (throughput) is

$$\mathcal{T}(\epsilon,\mathrm{SNR}_{\rm DT})\triangleq(1-\kappa)(1-\epsilon)C_\epsilon(\mathrm{SNR}_{\rm DT}), \tag{5.22}$$

where $(1-\kappa)$ accounts for the channel estimation overhead. In what follows, we select ϵ to maximize the throughput, yielding the optimal $\epsilon^*(\mathrm{SNR}_{\mathrm{DT}})$ at a given SNR $\mathrm{SNR}_{\mathrm{DT}}$ as the unique fixed point of $\mathrm{d}\mathcal{T}(\epsilon, \mathrm{SNR}_{\mathrm{DT}})/\mathrm{d}\epsilon = 0$, or equivalently,

$$\ln\Big(1-\mathrm{SNR}_{\mathrm{DT}}\ln(1-\epsilon)\Big)\Big(1-\mathrm{SNR}_{\mathrm{DT}}\ln(1-\epsilon)\Big) = \mathrm{SNR}_{\mathrm{DT}}.$$

We denote the resulting throughput maximized over ϵ as $\mathcal{T}^*(\mathrm{SNR}_{\mathrm{DT}}) \triangleq \mathcal{T}(\epsilon^*(\mathrm{SNR}_{\mathrm{DT}}), \mathrm{SNR}_{\mathrm{DT}})$.

We envision a mechanism in which the pilot signal transmitted in the last data transmission slot (at time $t = k + T_{\mathrm{DT}} - 2$) is used to generate the binary feedback signal

$$Y = \begin{cases} j, & \Gamma_j > \eta_{DT}^{(I)}, \\ \emptyset, & \Gamma_j \le \eta_{DT}^{(I)}, \end{cases} \tag{5.23}$$

transmitted by the UE to the BS in the last slot of the DT phase (at time $t = k + T_{\mathrm{DT}} - 1$). As in (5.11) for the BT feedback, $Y{=}j$ denotes active SBPI detected, whereas $Y = \emptyset$ denotes inactive SBPI detection, due to either loss of alignment or blockage. Similarly to (5.11),

$$\Gamma_j \triangleq \frac{|\mathbf{x}_{k+T_{\mathrm{DT}}-2}^{(p)H}\mathbf{y}_{k+T_{\mathrm{DT}}-2}^{(p)}|^2}{(1+F)N_0W_{\mathrm{tot}}\|\mathbf{x}_{k+T_{\mathrm{DT}}-2}^{(p)}\|_2^2}$$

is based on the pilot signal $\mathbf{x}_{k+T_{\mathrm{DT}}-2}^{(p)}$ (of duration κL) and on the corresponding signal $\mathbf{y}_{k+T_{\mathrm{DT}}-2}^{(p)}$ received on the second last slot of the DT phase. The distribution of the feedback conditional on $s_I^*(X_t){=}S_I$ and $B_t^{(I)}{=}B_I$ at the 2nd last slot ($t{=}k{+}T_{\mathrm{DT}}{-}2$) can be computed as a special case of (5.14) and (5.15) with $|\mathcal{S}_{\mathrm{BT}}|{=}\,1$ (since in the DT phase only one beam j is used for data transmission) and κL in place of L (since only κL symbols are used as pilot signal), yielding the probability of incorrectly detecting an active SBPI as

$$\mathbb{P}(Y{=}j|S_I \neq j \text{ or } B_I = 0){=}\exp\left\{-\frac{\eta_{DT}}{1{+}\rho_I\kappa\mathrm{SNR}_{\mathrm{DT}}L}\right\}, \tag{5.24}$$

and that of incorrectly detecting j to be an inactive SBPI as

$$\mathbb{P}(Y{=}\emptyset|S_I=j,B_I=1){=}1{-}\exp\left\{-\frac{\eta_{DT}}{1{+}\kappa\mathrm{SNR}_{\mathrm{DT}}L}\right\}. \tag{5.25}$$

As in the BT phase, we choose $\eta_{DT}^{(I)}$ so that the probabilities of misdetection and false alarm are both equal to $\delta_{\mathrm{DT}}^{(I)}$, yielding

$$\delta_{\mathrm{DT}}^{(I)}{=}\exp\left\{\frac{-\eta_{DT}}{1{+}\rho_I\kappa\mathrm{SNR}_{\mathrm{DT}}L}\right\}{=}1{-}\exp\left\{\frac{-\eta_{DT}}{1{+}\kappa\mathrm{SNR}_{\mathrm{DT}}L}\right\}. \tag{5.26}$$

5.2 POMDP Formulation

We now formulate the problem of optimizing the BT, DT and HO strategy as a constrained POMDP. In the following, we define the elements of this POMDP.

States: the state at time k is denoted by Z_k. We introduce the state $\bar{z}$ to characterize the episode termination, so that $Z_k{=}\bar{z}$ if the UE exited the coverage area of the two BSs, i.e., $X_k{\notin}\mathcal{X}$. Otherwise ($Z_k{\neq}\bar{z}$), we define the state as $Z_k{\triangleq}(U_k,I_k)$, where $I_k{\in}\mathcal{I}$ is the index of the serving BS, $U_k{\triangleq}(S_k,B_k)$ is the joint SBPI-blockage state, taking values from the set $\mathcal{U}{=}\mathcal{S}{\times}\{0,1\}^2$, $S_k{=}(S_k^{(0)},S_k^{(1)}){\in}\mathcal{S}$ with $S_k^{(i)}{\triangleq}s_i^*(X_k)$ is the SBPI at the current UE position X_k, $B_k{=}(B_k^{(0)},B_k^{(1)})$ is the blockage state of the two BSs. The overall state space, including the absorbing $\bar{z}$, is then $\mathcal{Z}{=}(\mathcal{U}{\times}\mathcal{I}){\cup}\{\bar{z}\}$. Note that the position of the UE and the blockage state cannot be directly observed, thereby making the state U_k unobservable. We model such state uncertainty via a belief β_k, representing the probability distribution of U_k, given the information collected (actions selected and feedback) up to time k.

Actions: the serving BS can perform three actions: beam-training (BT), data transmission (DT), or handover (HO). However, differently from standard POMDPs in which each action takes one slot, in this chapter we generalize the model to actions taking multiple slots, as explained next.

If action HO is chosen, the data plane is transferred to the other BS, which becomes the serving one for the successive time-slots, until HO is chosen again or the episode terminates.

HO requires $T_{\rm HO}$ time-slots to complete, due to the delay to coordinate the transfer of the data traffic between the two BSs.

If actions BT is chosen, the serving BS I chooses the BPI set $\mathcal{S}_{\rm BT}\subseteq\mathcal{S}_I$ to scan and the target SNR $\mathrm{SNR}_{\rm BT}$. The transmission power is then found via (5.8), and the feedback error probability $\delta_{\rm BT}^{(I)}$ is found by solving (5.18). The action duration is $T_{\rm BT}=|\mathcal{S}_{\rm BT}|+1$: $|\mathcal{S}_{\rm BT}|$ slots for scanning the BPI set $\mathcal{S}_{\rm BT}$, and one slot for the feedback back to the serving BS.

If action DT is chosen, then the serving BS I selects the BPI $j\in\mathcal{S}_I$ to perform data communication with the UE, along with the duration $T_{\rm DT}\geq 2$ of the data communication session, and the target SNR $\mathrm{SNR}_{\rm DT}$. The transmission power is then determined via (5.8), and the transmission rate is given by (5.21) to achieve ϵ-outage capacity, so that the resulting throughput (in case of LOS and correct alignment) is $\mathcal{T}^*(\mathrm{SNR}_{\rm DT})$. The duration of the data communication session $T_{\rm DT}$ includes the second last slot for the feedback signal, which is transmitted from the UE to the BS in the last slot. The feedback error probability $\delta_{\rm DT}^{(I)}$ is the unique fixed point of (5.26).

We represent compactly these actions as $(c,\Pi_c)\in\mathcal{A}_I$, with action space $\mathcal{A}_I$, where $c\in\{\mathrm{BT},\mathrm{DT},\mathrm{HO}\}$ refers to the action class and $\Pi_c=(\mathcal{S}_c,\mathrm{SNR}_c,T_c)$ specifies the corresponding parameters: $\mathcal{S}_c\subseteq\mathcal{S}_I$ is a subset of BPIs of serving BS I, used during the action, SNR_c is the target SNR, so that the corresponding transmission power is given by (5.8), and T_c is the action duration. For HO, we set $\mathrm{SNR}_{\rm HO}=0$ and $\mathcal{S}_{\rm HO}=\emptyset$.

Observations: after selecting action $A_k\in\mathcal{A}_I$ of duration T in slot k and executing it in state $u_k\in\mathcal{U}$, the BS observes Y_{k+T} taking value from the observation space $\bar{\mathcal{Y}}\triangleq\mathcal{Y}\cup\{\bar{z}\}$, where $\mathcal{Y}\triangleq\mathcal{S}_1\cup\mathcal{S}_2\cup\{\emptyset\}\cup\{\bar{z}\}$. $Y_{k+T}=\bar{z}$ denotes that $Z_k=\bar{z}$, so that the UE exited the coverage area of the two BSs and the episode terminates; otherwise, Y_{k+T} denotes the feedback signal after the action is completed, as described in (5.12) and (5.23) for the BT and DT actions ($Y_k=\emptyset$ under the HO action).

Transition and Observation probabilities: Let $\mathbb{P}(Z_{k+T}=z',Y_{k+T}=y|Z_k=z,A_k=a)$ be the probability of moving from a non-absorbing state $z=(u,I)\in\mathcal{Z}\setminus\{\bar{z}\}$ to state $z'\in\mathcal{Z}$ and observing $y\in\bar{\mathcal{Y}}$ under action $a\in\mathcal{A}_I$ of duration T. If the episode does not terminate ($Z_{k+T}\neq\bar{z}$ and $y\neq\bar{z}$), let $Z_{k+T}=(u',I')$ be the next state. Note that the new serving BS I' is a function $\mathbb{I}(a,I)$ of the chosen action: if a is the HO action then $I'=\mathbb{I}(a,I)=1-I$,

otherwise $I'=\mathbb{I}(a,I)=I$. Using the law of conditional probability, the transition probability is then expressed as

$$\begin{aligned}&\mathbb{P}(Z_{k+T}=(u',I'),Y_{k+T}=y|Z_k=(u,I),A_k=a)\\&=\mathbb{P}(U_{k+T}=u',Y_{k+T}=y|U_k=u,I_k=I,A_k=a)\chi(I'=\mathbb{I}(a,I)),\end{aligned}\tag{5.27}$$

since (U_{k+T},Y_{k+T}) is conditionally independent of I_{k+T} given (U_k,I_k,A_k). To characterize the first term in (5.27), under the HO action $a=(\text{HO},\emptyset,0,T_{\text{HO}})$, of duration $T=T_{\text{HO}}$, the observation signal is deterministically $Y_{k+T}=\emptyset$, yielding

$$\mathbb{P}(U_{k+T}=(s',b'),Y_{k+T}=\emptyset|U_k=(s,b),I_k=I,A_k=a)=\mathbf{P}_{s'b'|sb}(T),\tag{5.28}$$

where $\mathbf{P}_{s'b'|sb}(T)$ is the T steps transition probability from $U_k=(s,b)$ to $U_{k+T}=(s',b')$, found recursively as $\mathbf{P}_{s'b'|sb}(T)=\sum_{s'',b''}\mathbf{P}_{s'b'|s''b''}(T-1)\mathbf{P}_{s''b''|sb}$ with $\mathbf{P}_{s'b'|sb}(1)=\mathbf{P}_{s'b'|sb}$. In other words, the UE moves from s to s' and the BSs's blockage states move from b to b', in T slots.

Under the BT action $a=(\text{BT},\mathcal{S}_{\text{BT}},\text{SNR},T)$, of duration $T=|\mathcal{S}_{\text{BT}}|+1$, the observation signal is $Y_{k+T}=y\in\mathcal{S}_{\text{BT}}\cup\{\emptyset\}$ (see the BT signaling mechanism in Section 5.1). Therefore,

$$\begin{aligned}&\mathbb{P}\Big(U_{k+T}=(s',b'),Y_{k+T}=y|U_k=(s,b),I_k=I,A_k=a\Big)\\&=\mathbb{P}(Y_{k+T}=y|\mathcal{S}_{\text{BT}},S_k^{(I)}=s_I,B_k^{(I)}=b_I,I_k=I)\mathbf{P}_{s'b'|sb}(T),\end{aligned}$$

where $\mathbb{P}(Y=y|\mathcal{S},S_k^{(I)}=s_I,B_k^{(I)}=b_I,I_k=I)$ has been defined in (5.13)-(5.17) for the cases of active SBPI $\{s_I\in\mathcal{S}\}\cap\{b_I=1\}$ and inactive SBPI $\{s_I\notin\mathcal{S}\}\cup\{b_I=0\}$.

Finally, under the DT action $a=(\text{DT},\{j\},\text{SNR},T)$, the observation signal is $Y_{k+T}=y\in\{j,\emptyset\}$ (see the DT signaling in Section 5.1). However, in this case the feedback signal is generated based on the second last slot, i.e., it depends on the state U_{k+T-2} at time $k+T-2$. By marginalizing with respect to $S_{k+T-2}=s''$ and $B_{k+T-2}=b''$, we then obtain (5.29) given at the top of page 103. To explain it, note that: the system moves from $(S_k,B_k)=(s,b)$ to $(S_{k+T-2},B_{k+T-2})=(s'',b'')$ in $T-2$ steps; then, the feedback signal Y_{k+T} is generated with distribution $\mathbb{P}(Y_{k+T}=y|\{j\},S_{k+T-2}^{(I)}=s_I'',B_{k+T-2}^{(I)}=b_I'',I_{k+T-2}=I)$, given in (5.24), (5.25) for the

$$\mathbb{P}\Big(U_{k+T}=(s',b'),Y_{k+T}=y|U_k=(s,b),I_k=I,A_k=a\Big) \tag{5.29}$$
$$=\sum_{s''\in\mathcal{S},b''\in\{0,1\}^2}\mathbb{P}\Big(U_{k+T}=(s',b'),Y_{k+T}=y,S_{k+T-2}=s'',B_{k+T-2}=b''|U_k=(s,b),I_k=I,A_k=a\Big)$$
$$=\sum_{s''\in\mathcal{S},b''\in\{0,1\}^2}\Big[\mathbf{P}_{s''b''|sb}(T-2)\mathbb{P}\Big(Y_{k+T}=y|\{j\},S^{(I)}_{k+T-2}=s''_I,B^{(I)}_{k+T-2}=b''_I,I_{k+T-2}=I\Big)\mathbf{P}_{s'b'|s''b''}(2)\Big]$$

cases of active or inactive SBPI in $\{j\}$; finally, in the remaining 2 steps ,the system moves from $(S_{k+T-2},B_{k+T-2})=(s'',b'')$ to $(S_{k+T},B_{k+T})=(s',b')$.

The probability of terminating the episode ($z'=\bar{z}$ and $y=\bar{z}$) is equivalent to the probability of exiting the coverage area of the two BSs within T steps,

$$\mathbb{P}(Z_{k+T}=\bar{z},Y_{k+T}=\bar{z}|Z_k=(u,I),A_k=a)=1-\sum_{u'\in\mathcal{U},y\in\mathcal{Y}}\mathbb{P}(U_{k+T}=u',Y_{k+T}=y|U_k=u,I_k=I,A_k=a)$$

since it is the complement event of $\cup_{z\in\mathcal{Z}\setminus\{\bar{z}\}}\cup_{y\in\mathcal{Y}}\{Z_k=z,Y_{k+T}=y\}$.

Costs and Rewards: for every state $z=(u,I)\in\mathcal{Z}\setminus\{\bar{z}\}$ and action a, we let $r(u,I,a)$ and $e(u,I,a)$ be the expected number of bits transmitted from the serving BS to the UE and the expected energy cost, respectively. Under the HO and BT actions, we have that $r(u,I,a)=0$ (since no bits are transmitted during these actions). On the other hand, under the DT action $a=(\mathrm{DT},\{j\},\mathrm{SNR},T_{\mathrm{DT}})$ taken in slot k, the expected throughput in the tth communication slot, $t\in\{0,\ldots,T_{\mathrm{DT}}-2\}$, is $\mathcal{T}^*(\mathrm{SNR})$ as in (5.22), maximized over ϵ, if the current state is such that $S^{(I)}_{k+t}=j$ and $B^{(I)}_{k+t}=1$ (i.e., j is an active SBPI); otherwise, outage occurs and the expected throughput is zero. Therefore, we find that

$$\begin{aligned}&r((s,b),I,(\mathrm{DT},\{j\},\mathrm{SNR},T_{\mathrm{DT}}))\\&=\mathcal{T}^*(\mathrm{SNR})\sum_{t=0}^{T_{\mathrm{DT}}-2}\mathbb{P}(S^{(I)}_{k+t}=j,B^{(I)}_{k+t}=1|S_k=s,B_k=b)\\&=\mathcal{T}^*(\mathrm{SNR})\sum_{t=0}^{T_{\mathrm{DT}}-2}\sum_{(s',b')\in\mathcal{U}}\mathbf{P}_{s'b'|sb}(t)\chi(s'_I=j,b'_I=1).\end{aligned} \tag{5.30}$$

The energy cost of a HO action is $e(u,I,a)$=0; that of DT or BT action a=$(c,\mathcal{S},\text{SNR},T)$ is found from (5.8) as (note that T=$|\mathcal{S}|$+1 for a BT action and $|\mathcal{S}|$=1 for a DT action)

$$e(u,I,a)=\frac{(T-1)\Delta_t}{|\mathcal{S}|}\sum_{j\in\mathcal{S}}\frac{\sigma_w^2}{\Upsilon_{j,I}+\sigma_{\text{DIF},I}^2}\text{SNR}. \tag{5.31}$$

Note that the last slot of the DT or BT phases is reserved to the feedback transmission, with no energy cost for the BS.

Policy and Belief updates: Since the agent cannot directly observe the pairs of BPI S and blockage B, we define the POMDP state as (β,I), where β denotes the belief, i.e., the probability distribution over U=(S,B), given the information collected so far and I is the index of the serving BS. The belief β takes values from belief space $\mathcal{B}\triangleq\{\beta\in\mathbb{R}^{|\mathcal{U}|}:\beta(u)\geq 0\ \forall u\in\mathcal{U},\sum_{u\in\mathcal{U}}\beta(u)=1\}$. Given (β,I), the serving BS selects an action a according to a policy $a=\pi(\beta,I)$, that is part of our design in Section 5.3; then, after executing the action a and receiving the feedback signal $y\in\mathcal{Y}$, the BS I updates the belief according to Bayes' rule as

$$\beta'(u')=\mathbb{P}(u'\mid y,a,\beta,I)=\frac{\sum_{u\in\mathcal{U}}\beta(u)\mathbb{P}(u',y|u,I,a)}{\sum_{u\in\mathcal{U}}\beta(u)\sum_{u''\in\mathcal{U}}\mathbb{P}(u'',y|u,I,a)}, \tag{5.32}$$

with $\mathbb{P}(u',y|u,I,a)$ given by (5.28)-(5.29), and the serving BS becomes $I'=\mathbb{I}(a,I)$. We denote the function that maps the belief β, action a and observation y under the serving BS I as $\beta'=\mathbb{B}_I(y,a,\beta)$. Note that Y=$\bar{z}$ indicates episode termination.

5.3 Optimization Problem

Our goal is to determine a policy π (a map from beliefs to actions) maximizing the expected throughput, under an average power constraint $\bar{P}_{\text{avg}}$, starting from an initial belief β_0=β_0^* and serving BS I_0=I_0^*. From Little's Theorem [71], the average rate and power consumption can be expressed as

$$\bar{T}^\pi\triangleq\frac{\bar{R}_{\text{tot}}^\pi}{\bar{D}_{\text{tot}}},\ \bar{P}^\pi\triangleq\frac{\bar{E}_{\text{tot}}^\pi}{\bar{D}_{\text{tot}}}, \tag{5.33}$$

where $\bar{R}^{\pi}_{\text{tot}}$, $\bar{E}^{\pi}_{\text{tot}}$ are the total expected number of bits transmitted and energy cost during an episode; $\bar{D}_{\text{tot}}$ is the expected episode duration, which only depends on the mobility process but is independent of the policy π. Therefore, we aim to solve

P1:

$$\max_{\pi} \bar{R}^{\pi}_{\text{tot}} \triangleq \mathbb{E}_{\pi}\Big[\sum_{n=0}^{\infty} r(u_{t_n}, i_{t_n}, a_{t_n})\chi(Z_{t_n}\neq\bar{z})\Big|\beta_0=\beta_0^*, I_0=I_0^*\Big],$$

s.t.

$$\bar{E}^{\pi}_{\text{tot}} \triangleq \mathbb{E}_{\pi}\Big[\sum_{n=0}^{\infty} e(u_{t_n},i_{t_n},a_{t_n})\chi(Z_{t_n}\neq\bar{z})\Big|\beta_0=\beta_0^*,I_0=I_0^*\Big]\leq E_{\max},$$

where $E_{\max}\triangleq\bar{D}_{\text{tot}}\bar{P}_{\text{avg}}$; t_n is the time index of the n-th decision round, recursively computed as $t_{n+1}=t_n+T_n$, where T_n is the duration (number of slots) of the action taken in the n-th decision round and $t_0=0$. We opt for a Lagrangian relaxation to handle the cost constraint, and define $\mathcal{L}_{\lambda}(u,i,a) = r(u,i,a)-\lambda e(u,i,a)$ for $\lambda\geq 0$. For a generic policy π, we define its value function as[4]

$$V^{\pi}_{\lambda}(\beta,I)=\mathbb{E}_{\pi}\Big[\sum_{n=0}^{\infty}\mathcal{L}_{\lambda}(u_{t_n},i_{t_n},a_{t_n})\chi(Z_{t_n}\neq\bar{z})\mid\beta_0=\beta,I_0=I\Big].$$

The goal is to determine the optimal policy π^* which maximizes the value function, i.e.,

$$V^*_{\lambda}(\beta,I)\triangleq\max_{\pi} V^{\pi}_{\lambda}(\beta,I). \tag{5.34}$$

The optimal dual variable is then found via the dual problem

$$\lambda^* = \arg\min_{\lambda\geq 0} V^*_{\lambda}(\beta_0^*,I_0^*)+\lambda E_{\max}. \tag{5.35}$$

[4]↑Note that the convergence of this series is guaranteed by the presence of the absorbing state $\bar{z}$.

It is well known that the optimal value function for a given λ uniquely satisfies Bellman's optimality equation [37] $V_\lambda^* = H_\lambda[V_\lambda^*]$, where we have defined the operator $\hat{V} = H_\lambda[V]$ as

$$\hat{V}(\beta, I) = \max_{a \in \mathcal{A}} \sum_{u \in \mathcal{U}} \beta(u) \Big[\mathcal{L}_\lambda(u, I, a) + \sum_{(u', y) \in \mathcal{U} \times \mathcal{Y}} \mathbb{P}(u', y | u, I, a) V\Big(\mathbb{B}_I(y, a, \beta), \mathbb{I}(a, I) \Big) \Big], \forall (\beta, I) \in \mathcal{B} \times \mathcal{I}.$$

The optimal value function V_λ^* can be arbitrarily well approximated via the value iteration algorithm $V_{n+1} = H_\lambda[V_n]$, where $V_0(\beta, I) = 0, \forall (\beta, I) \in \mathcal{B} \times \mathcal{I}$. Moreover, V_n is a piece-wise linear and convex function [37], so that, at any stage of value iteration, it can be expressed by a finite set of hyperplanes $\mathcal{Q}_n^{(I)} \equiv \{(\alpha_{n,I,\ell}^{(r)}, \alpha_{n,I,\ell}^{(e)})\}_{\ell=1}^{N_n^{(I)}}$ of cardinality $N_n^{(I)}$,

$$V_n(\beta, I) = \max_{\alpha_I \in \mathcal{Q}_n^{(I)}} \langle \beta, \alpha_I^{(r)} - \lambda \alpha_I^{(e)} \rangle, \tag{5.36}$$

where $\langle \beta, \alpha \rangle = \sum_u \beta(u) \alpha(u)$ denotes inner product. Each hyperplane $(\alpha_I^{(r)}, \alpha_I^{(e)}) \in \mathcal{Q}_n^{(I)}$ is associated with an action $a_{\alpha_I} \in \mathcal{A}_I$, so that the maximizing hyperplane α_I^* in (5.36) defines the policy $\pi_n(\beta, I) = a_{\alpha_I^*}$. Note that a distinguishing feature of our approach compared to [37] is that we define distinct hyperplanes $\alpha_I^{(r)}$ for the reward and $\alpha_I^{(c)}$ for the cost; as we will see later, this approach will be key to solving the dual optimization problem to optimize the power constraint, since it allows to more efficiently track changes in the dual variable λ, as part of the dual problem (5.35), and to approximate the expected total reward and cost as

$$\begin{aligned} \bar{R}_n(\beta, I) = \langle \beta, \alpha_I^{(r)*} \rangle, \;\; \bar{E}_n(\beta, I) = \langle \beta, \alpha_I^{(e)*} \rangle, \\ \text{where } (\alpha_I^{(r)*}, \alpha_I^{(e)*}) = \arg \max_{\alpha_I \in \mathcal{Q}_n^{(I)}} \langle \beta, \alpha_I^{(r)} - \lambda \alpha_I^{(e)} \rangle. \end{aligned} \tag{5.37}$$

It can be shown (see for instance [36]) that the set of hyperplanes is updated recursively as

$$\begin{aligned} \mathcal{Q}_{n+1}^{(I)} \equiv \Big\{ (r(\cdot, I, a), e(\cdot, I, a)) + \sum_{u' \in \mathcal{U}, y \in \mathcal{Y}} \mathbb{P}(u', y | \cdot, I, a) \left(\alpha_{I',y}^{(r)}(u'), \alpha_{I',y}^{(e)}(u') \right) : \\ a \in \mathcal{A}_I, I' = \mathbb{I}(a, I), [(\alpha_{I',y}^{(r)}, \alpha_{I',y}^{(e)})]_{\forall y \in \mathcal{Y}} \in (Q_n^{(I')})^{|\mathcal{Y}|} \Big\}, \end{aligned} \tag{5.38}$$

so that the cardinality grows as $N_{n+1}^{(I)} = |\mathcal{Q}_{n+1}^{(I)}| = \mathcal{O}(|\mathcal{A}|^{|\mathcal{Y}|^n})$ – doubly exponentially with the number of iterations.

For this reason, computing optimal planning solutions for POMDPs is an intractable problem for any reasonably sized task. This calls for approximate solution techniques, e.g., PERSEUS [37], which we introduce next.

PERSEUS [37] is an approximate PBVI algorithm for unconstrained POMDPs. Its key idea is to define an approximate backup operator $\tilde{H}_\lambda[\cdot]$ (in place of $H_\lambda[\cdot]$), restricted to a discrete subset of POMDP states in $\tilde{\mathcal{B}}_0 \cup \tilde{\mathcal{B}}_1$, where $\tilde{\mathcal{B}}_I$ is discrete set of POMDP states with the serving BS I, chosen as representative of the entire belief space $\mathcal{B}$; in other words, for a given value function $\tilde{V}_n$ at stage n, PERSEUS builds a value function $\tilde{V}_{n+1} = \tilde{H}[\tilde{V}_n]$ that improves the value of all POMDP states (β, I) with $\beta \in \tilde{\mathcal{B}}_I$, without regard for the POMDP states outside of this discrete set, $\beta \notin \tilde{\mathcal{B}}_I$. For each $I \in \mathcal{I}$, the goal of the algorithm is to provide a $|\tilde{\mathcal{B}}_I|$-dimensional set of hyperplanes $\alpha_I = (\alpha_I^{(r)}, \alpha_I^{(e)}) \in \mathcal{Q}_I$ and associated actions a_{α_I}. Given such set, the value function at any other POMDP state, (β, I) is then approximated via (5.36) as $\tilde{V}(\beta, I) = \langle \beta, \alpha_I^{(r)*} - \lambda \alpha_I^{(e)*} \rangle$, where $\alpha_I^* = (\alpha_I^{(r)*}, \alpha^{(e)*}) = \arg\max_{(\alpha_I^{(r)}, \alpha^{(e)}) \in \mathcal{Q}^{(I)}} \langle \beta, \alpha_I^{(r)} - \lambda \alpha_I^{(e)} \rangle$, which defines an approximately optimal policy $\pi(\beta, I) = a_{\alpha_I^*}$.

Key to the performance of PBVI is the design of $\tilde{\mathcal{B}}_I$, which should be representative of the belief points encountered in the system dynamics. In the PBVI literature [36], most of the strategies to design $\tilde{\mathcal{B}}_I$ focus on selecting reachable belief points, rather than covering uniformly the entire belief simplex. We choose the beliefs in the following two steps. For each $I \in \mathcal{I}$, an initial belief set $\mathcal{B}_I^{(0)}$ is selected deterministically to cover uniformly the belief space. followed by expansion of $\{\mathcal{B}_I^{(0)}, I \in \mathcal{I}\}$ using the *Stochastic simulation and exploratory action* (SSEA) algorithm [36] to yield the expanded belief points set $\{\tilde{\mathcal{B}}_I, I \in \mathcal{I}\}$. After initializing $\mathcal{B}_I^{(0)}$, given $\mathcal{B}_I^{(n)}$ at iteration n, for each $\beta \in \mathcal{B}_I^{(n)}$, SSEA performs a one step forward simulation with each action in the action set, thus producing new POMDP states $\{(\beta_a, I_a), \forall a \in \mathcal{A}_I\}$. At this point, it computes the L1 distance between each new β_a and its closest neighbor in $\mathcal{B}_{I_a}^{(n)}$, and adds the point β_{a^*} to $\mathcal{B}_{I_{a^*}}^{(n)}$ if $\min_{\beta \in \mathcal{B}_{I_{a^*}}^{(n)}} \|\beta_{a^*} - \beta\|_1 \geq \min_{\beta \in \mathcal{B}_{I_a}^{(n)}} \|\beta_a - \beta\|_1, \forall a \in \mathcal{A}_I$, so as to more widely cover the belief space. This expansion is performed multiple times to obtain $\{\tilde{\mathcal{B}}_I, I \in \mathcal{I}\}$.

The approximate backup operation of PERSEUS is given by Algorithm 1, which takes as input the index of the serving BS I, the set of belief points $\tilde{\mathcal{B}}_I$ associated with BS I, the sets of hyperplanes $\{\mathcal{Q}_n^{(i)}, i \in \mathcal{I}\}$ and the corresponding actions, and outputs a new set $\mathcal{Q}_{n+1}^{(I)}$ along with their corresponding actions. To do so: in line 4, a belief is chosen randomly from $\hat{\mathcal{B}}_I$; in lines 5–7, the hyperplane associated with each action $a \in \mathcal{A}$ is computed; in particular, line 6 computes the hyperplane associated with the future value function $V_n(\mathbb{B}_I(y,a,\beta),\mathbb{I}(a,I))$, for each possible observation y resulting in the belief update $\mathbb{B}_I(y,a,\beta)$; line 7 instead performs the backup operation to determine the new hyperplane of $V_{n+1}(\beta,I)$ associated to action a; line 8 determines the optimal action that maximizes the value function, so that lines 5-8 overall approximate the value iteration update $V_{n+1}(\beta,I) = \max_a \mathbb{E}_{U,Y|a,\beta,I}[\mathcal{L}_\lambda(U,I,a)+V_n(\mathbb{B}_I(Y,a,\beta),\mathbb{I}(a,I))]$; in lines 9-12, the new hyperplane and the associated action is added to the set $\mathcal{Q}_{n+1}^{(I)}$, but only if it yields an improvement in the value function $V_{n+1}(\beta,I) > \tilde{V}_n(\beta,I)$; otherwise, the previous hyperplane is used; finally, lines 13-14 update the set of un-improved POMDP states based on the newly added hyperplane; only the belief points that have not been improved are part of the next iterations of the algorithm, and the process continues until the set $\hat{\mathcal{B}}_I$ is empty. Overall, the algorithm guarantees monotonic improvements of the value function in $\tilde{\mathcal{B}}_I$. Note that PERSEUS can be executed in parallel by each serving BS, thereby reducing the computation time.

The basic routine for C-PBVI is given in Algorithm 2. However, differently from [37], we also embed the dual optimization (5.35) by updating the dual variable λ in line 6. In line 4, we perform one backup operation via PERSEUS (Algorithm 1); in line 5, we compute the new value function $V_{n+1}(\beta,I)$ (based on the new hyperplane sets $\mathcal{Q}_{n+1}^{(I)}$); in line 6, we compute the approximate cost $\bar{E}_{n+1}$ starting from state (β_0^*, I_0^*), based on the optimal hyperplane α^*; this is used in line 7 to update the dual variable λ via projected subgradient descent, with the goal to solve the dual problem (5.35) (note that $E_{\max} - \bar{E}_{n+1}$ is a subgradient of the dual function, see [72]): as a result, λ_n is decreased if the estimated cost $\bar{E}_{n+1} < E_{\max}$, to promote throughput maximization over energy cost minimization, otherwise it is increased; the algorithm continues until the KKT conditions are approximately satisfied [72], i.e., $\max_{I\in\mathcal{I}} \max_{\beta\in\tilde{\mathcal{B}}_I} |V_{n+1}(\beta,I) - V_n(\beta,I)| < \epsilon_V$ (i.e., an approximately fixed point of $V_{n+1} =$

Algorithm 1: function PERSEUS

input : I, $\tilde{\mathcal{B}}_I$, $\{\mathcal{Q}_n^{(i)}\}_{i\in\mathcal{I}}$, $\{a_{\alpha_i}^n, \alpha_i \in \mathcal{Q}_n^{(i)}\}, \forall i \in \mathcal{I}$, λ

Init: $\tilde{V}_{n+1}(\beta, I) = -\infty, \forall \beta \in \tilde{\mathcal{B}}_I$; $\hat{\mathcal{B}}_I \equiv \tilde{\mathcal{B}}_I$; $\mathcal{Q}_{n+1}^{(I)} = \emptyset$

$\tilde{V}_n(\beta, I) \leftarrow \max_{\alpha_I \in \mathcal{Q}_n^{(I)}} \langle \beta, \alpha_I^{(r)} - \lambda \alpha_I^{(e)} \rangle$, and maximizer $(\alpha_{\beta,I}^{(r)}, \alpha_{\beta,I}^{(e)}), \forall \beta \in \tilde{\mathcal{B}}_I$

while $\hat{\mathcal{B}}_I \neq \emptyset$ **do** `// Unimproved beliefs`

  Sample β from $\hat{\mathcal{B}}_I$ (e.g., uniformly)

  for *each action* a **do**

    $I' = \mathbb{I}(a, I)$; $\alpha_{y,a}^* = \arg\max_{\alpha \in \mathcal{Q}_n} \langle \mathbb{B}_I(y,a,\beta), \alpha_{I'}^{(r)} - \lambda \alpha_{I'}^{(e)} \rangle, \forall y \in \mathcal{Y}$

    $\hat{\alpha}_a^* = (r(\cdot, I, a), e(\cdot, I, a)) + \sum_{u', y} \mathbb{P}(u', y | \cdot, I, a)(\alpha_{y,a}^{*(r)}(u'), \alpha_{y,a}^{*(e)}(u'))$

  Solve $V_{n+1}(\beta, I) = \max_{a \in \mathcal{A}} \langle \beta, \hat{\alpha}_a^{*(r)} - \lambda \hat{\alpha}_a^{*(e)} \rangle$ and maximizing action a^* and $\hat{\alpha} = \hat{\alpha}_{a^*}^*$

  if $V_{n+1}(\beta, I) > \tilde{V}_n(\beta, I)$ **then** `// $\hat{\alpha}$ improves value`

    $\mathcal{Q}_{n+1}^{(I)} \leftarrow \mathcal{Q}_{n+1}^{(I)} \cup \{\hat{\alpha}\}$; $a_{\hat{\alpha}}^{n+1} = a^*$ `// add` $\hat{\alpha}$ `to` $\mathcal{Q}_{n+1}^{(I)}$ `and define action associated with` $\hat{\alpha}$;

  else `// keep previous hyperplane` $\alpha_{\beta,I}$

    $\hat{\alpha} = \alpha_{\beta,I}$; $\mathcal{Q}_{n+1}^{(I)} \leftarrow \mathcal{Q}_{n+1}^{(I)} \cup \{\hat{\alpha}\}$; $a_{\hat{\alpha}}^{n+1} = a_{\hat{\alpha}}^n$

  $\tilde{V}_{n+1}(\tilde{\beta}, I) \leftarrow \max\{\langle \tilde{\beta}, \hat{\alpha}^{(r)} - \lambda \hat{\alpha}^{(e)} \rangle, \tilde{V}_{n+1}(\tilde{\beta}, I)\}, \forall \tilde{\beta} \in \tilde{\mathcal{B}}_I$

  $\hat{\mathcal{B}}_I \leftarrow \{\tilde{\beta} \in \hat{\mathcal{B}}_I : \tilde{V}_{n+1}(\tilde{\beta}, I) < \tilde{V}_n(\tilde{\beta}, I)\}$ `// New set of unimproved beliefs`

return $\mathcal{Q}_{n+1}^{(I)}$, $\{a_\alpha^{n+1}, \forall \alpha \in \mathcal{Q}_{n+1}^{(I)}\}$ `// new hyperplanes and associated actions`

$\tilde{H}[V_n]$ has been determined and PERSEUS converged), $\bar{E}_{n+1} \leq E_{\max}$ (primal feasibility constraint satisfied) and $\lambda_n |\bar{E}_{n+1} - E_{\max}| < \epsilon_E$ (complementary slackness; note that dual feasibility $\lambda_n \geq 0$ is enforced automatically in line 7).

After returning the sets of hyperplanes $\{\mathcal{Q}_{n+1}^{(I)}\}_{I\in\mathcal{I}}$, the associated actions $\{a_\alpha^{n+1}, \forall \alpha \in \mathcal{Q}_{n+1}^{(I)}\}$, and the dual variable λ_n, the (approximately) optimal action to be selected when operating under the state (β, I) can be computed as

$$\pi^*(\beta, I) = a_{\alpha^*}^{n+1}, \text{ where } \alpha^* = \arg\max_{\alpha \in \mathcal{Q}_{n+1}^{(I)}} \langle \beta, \alpha^{(r)} - \lambda_n \alpha^{(e)} \rangle,$$

along with the approximate expected reward and cost via (5.37).

In Fig. 5.2, we plot a time-series of the following variables for a portion of an episode executed under the C-PBVI policy (Algorithms 1 and 2) under the numerical setup of Section 5.5, with simulation parameters listed in Table 5.1: serving BS index I_k, BPI $S_k^{(I_k)}$

Algorithm 2: Constrained point based value iteration (C-PBVI)

Init: beliefs $\{\tilde{\mathcal{B}}_i\}_{i\in\mathcal{I}}$; hyperplanes $\mathcal{Q}_0^{(I)}=\{(\mathbf{0},\mathbf{0})\}, \forall I\in\mathcal{I}$; optimal actions $a^0_{(\mathbf{0},\mathbf{0})}=$ HO; value function $V_{n+1}(\beta,i)=0, \forall\beta\in\tilde{\mathcal{B}}_i, \forall i\in\mathcal{I}$; $\lambda_0\geq 0$; stepsize $\{\gamma_n=\gamma_0/(n+1), n\geq 0\}$

for $n=0,\ldots$ **do**

  for *each* $I\in\mathcal{I}$ **do**

    $(\mathcal{Q}_{n+1}^{(I)}, \{a_\alpha^{n+1}, \forall\alpha\in\mathcal{Q}_{n+1}^{(I)}\}) = \text{PERSEUS}(I, \tilde{\mathcal{B}}_I, \{\mathcal{Q}_n^{(I)}\}_{I\in\mathcal{I}}, \{a_\alpha^n, \alpha\in\mathcal{Q}_n^{(I)}\}, \lambda_n)$

    $V_{n+1}(\beta, I) = \max\limits_{\alpha\in\mathcal{Q}_{n+1}^{(I)}} \langle\beta, \alpha^{(r)}-\lambda_n\alpha^{(e)}\rangle, \forall\beta\in\tilde{\mathcal{B}}_I$

  Let $\bar{E}_{n+1}=\langle\beta_0^*, \alpha_{\beta_0,I_0}^{(e)*}\rangle$, where $\alpha^*_{\beta_0,I_0}=\arg\max\limits_{\alpha\in\mathcal{Q}_{n+1}^{(I_0^*)}} \langle\beta_0^*, \alpha^{(r)}-\lambda_n\alpha^{(e)}\rangle$

  $\lambda_{n+1}=\max\{\lambda_n+\gamma_n(\bar{E}_{n+1}-E_{\max}), 0\}$

  if $\max_{I\in\mathcal{I}}\max_{\beta\in\tilde{\mathcal{B}}_I}|V_{n+1}(\beta,I)-V_n(\beta,I)|<\epsilon_V$, $\bar{E}_{n+1}\leq E_{\max}$ *and* $\lambda_n|\bar{E}_{n+1}-E_{\max}|<\epsilon_E$ **then**

    return $\{\mathcal{Q}_{n+1}^{(I)}\}_{I\in\mathcal{I}}$, $\{a_\alpha^{n+1}, \forall\alpha\in\mathcal{Q}_{n+1}^{(I)}\}$, λ_n

and blockage state $B_k^{(I_k)}$ of the serving BS I_k, the action class $c\in\{\text{DT}, \text{BT}, \text{HO}\}$, the BT and DT feedbacks Y_{BT} and Y_{DT} as defined in (5.12) and (5.23). It can be observed in the figure that, at 0.915s, 0.985s and 1.025s, NACKs ($Y_{\text{DT}}=\emptyset$) are received after executing the DT action. After each one of these NACKs, the policy executes the BT action. If the BT feedback $Y_{\text{BT}}\neq\emptyset$, then DT is performed; otherwise, blockage is detected and the HO action is executed.

It should be noted that, although Algorithm 2 returns an approximately optimal design, it incurs substantial computational cost in POMDPs with large state and action spaces (hence large number of representative belief points). To remedy this, in the subsequent section we propose simple heuristic policies, inspired by the behavior of the C-PBVI policy described earlier and depicted in Fig. 5.2. These policies will be shown numerically to trade complexity with sub-optimality and achieve satisfactory performance.

5.4 Heuristic Policies

In this section, we present two heuristic policies, namely a belief-based heuristic (B-HEU) and a finite-state-machine (FSM)-based heuristic (FSM-HEU) and present closed-form

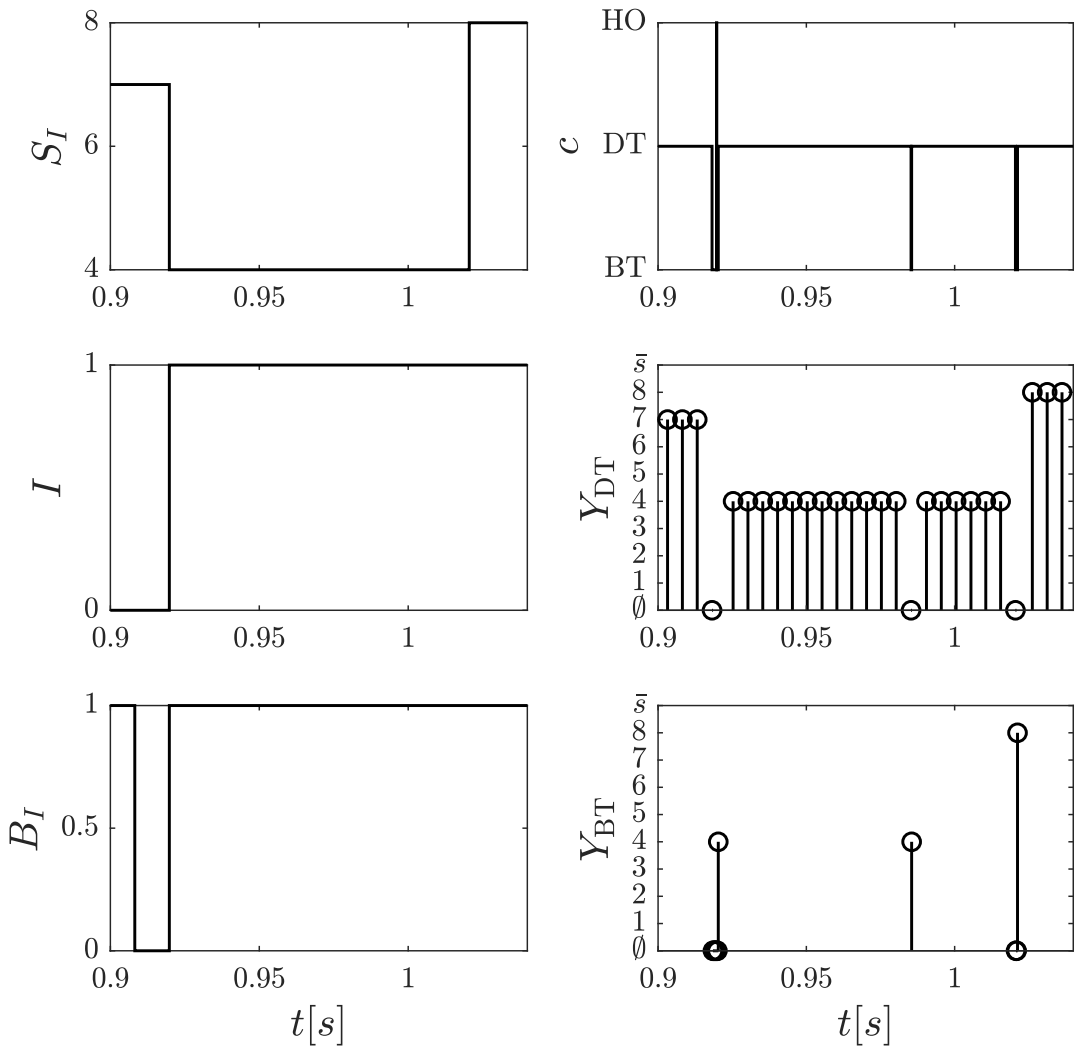

Figure 5.2. Execution of policy π^*.

expressions for the performance of FSM-HEU. Similarly to C-PBVI, B-HEU needs to track the belief β, whereas FSM-HEU is solely based on the current observation signal that defines transitions in a FSM. For this reason, FSM-HEU has lower complexity than B-HEU, while achieving only a small degradation in performance (see Section 5.5).

5.4.1 FSM-based Heuristic policy (FSM-HEU)

The key idea of FSM-HEU is that it selects actions based solely on a FSM, whose states define the action to be selected, and whose transitions are defined by the observation signal, as depicted in Fig. 5.3 and described next. In FSM-HEU, we consider the following actions:

- the HO action $A_k = (\text{HO}, \emptyset, 0, T_{\text{HO}})$ of duration T_{HO};
- the BT action $A_k = (\text{BT}, \mathcal{S}_I, \text{SNR}_{\text{BT}}, T_{\text{BT}})$ of duration $T_{\text{BT}} = |\mathcal{S}_I|+1$; in other words, the serving BS performs an exhaustive search over the entire set of SBPIs, with a fixed SNR SNR_{BT} (determined offline), followed by feedback;
- the $|\mathcal{S}_I|$ DT actions $(\text{DT}, j, \text{SNR}_{\text{DT}}, T_{\text{DT}})$, where $j \in \mathcal{S}_I$; in other words, the serving BS performs DT with fixed SNR SNR_{DT} and duration T_{DT} (both determined offline).

For notational convenience, we compactly refer to these actions as HO, BT and $(\text{DT}, j), j \in \mathcal{S}_I$, respectively. Let $A_k \in \{\text{BT}, \text{HO}\} \cup \{(\text{DT}, j) : j \in \mathcal{S}_I\}$ be the selected action of the serving BS I (the state of the FSM at time k), of duration T, and Y_{k+T} be the observation signal generated by such action, as described in Section 5.2; then, the FSM moves to state $A_{k+T} = \mathbb{A}_I(A_k, Y_{k+T})$, which defines the next action A_{k+T} to be selected in the next decision round. Note that $\mathbb{A}_I$ defines transitions in the FSM, and the process continues until the episode terminates.

Let us consider the transitions in the FSM, defined by the function $\mathbb{A}_I$, depicted in Fig. 5.3. If A_k=BT and the observation signal is Y_{k+T}=j∈$\mathcal{S}_I$, then the BS detects the strongest beam j; hence FSM-HEU switches to DT and uses the DT action $A_{k+T} = (\text{DT}, j) = \mathbb{A}_I(\text{BT}, j)$ of serving BS I in the next decision round, of duration T_{DT}. On the other hand, if the observation signal is Y_{k+T}=$\emptyset$, the BS detects blockage and performs HO to the non-serving BS, so that the new action is A_{k+T}=HO=$\mathbb{A}(\text{BT}, \emptyset)$ of serving BS I.

If A_k=(DT, j) of serving BS I, i.e., the DT action is executed on beam j, of duration T_{DT}, and the signal Y_{k+T}=j is observed, then the BS infers that the signal is still sufficiently strong to continue DT on the same beam, and the same action A_{k+T}=(DT, j)=$\mathbb{A}_I((\text{DT}, j), j)$ of the serving BS I is selected again. Otherwise (Y_{k+T}=$\emptyset$), the BS detects a loss of alignment, hence the BT action A_{k+T}=BT=$\mathbb{A}_I((\text{DT}, j), \emptyset)$ of the serving BS I is executed next.

Finally, if A_k=HO of serving BS I (the HO action is chosen, with observation signal Y_{k+T}=$\emptyset$), then the new serving BS $I' = 1 - I$ executes the BT action A_{k+T}=BT=$\mathbb{A}_I(\text{HO}, \emptyset)$ next. This procedure continues until the episode terminates.

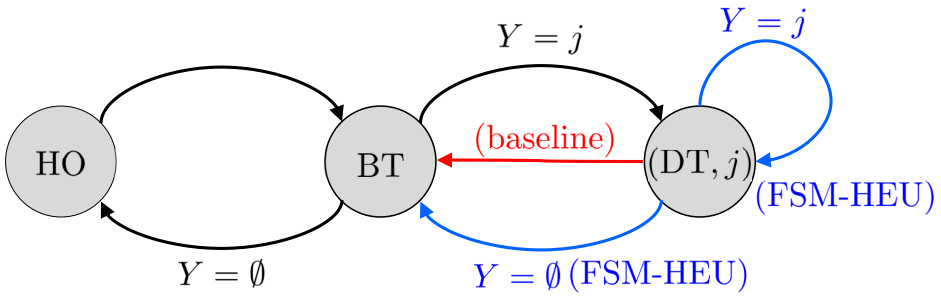

Figure 5.3. Evolution of the selected action A_k of the serving BS based on the observation signal Y_{k+T}. Black lines represent the transitions under both FSM-HEU and baseline policies; blue lines represent transitions under the FSM-HEU policy only; the red line represents the transition under the baseline policy only.

The performance of FSM-HEU can be computed in closed form. In fact, $G_k{=}(U_k, I_k, A_k)$, i.e., the system state (U_k, I_k) and action A_k, form a Markov chain, taking values from the state space

$$\mathcal{G} \equiv \bigcup_{I\in\mathcal{I}} \mathcal{U} \times \{I\} \times [\{\mathrm{BT}, \mathrm{HO}\} \cup \{(\mathrm{DT}, j) : j \in \mathcal{S}_I\}]. \tag{5.39}$$

To see this, note that the observation Y_{k+T} and next state (U_{k+T}, I_{k+T}) (where T is the duration of the selected action A_k) have joint distribution given by (5.27), which solely depends on G_k; then, in view of the FSM of Fig. 5.3, $A_{k+T} = \mathbb{A}(A_k, Y_{k+T})$ is a deterministic function of A_k and Y_{k+T}. The state transition probability is then obtained by computing the marginal with respect to the observation signal Y_{k+T}, yielding

$$\begin{aligned}
&\mathbb{P}\left(G'_{k+T} = (u', I', a') | G_k = (u, I, a)\right) \\
&= \sum_{y\in\mathcal{Y}:\mathbb{A}_I(a,y)=a'} \Big[\mathbb{P}\left(U_{k+T}{=}u', Y_{k+T}{=}y | U_k{=}u, I_k = I, A_k{=}a\right) \mathbb{P}(I_{k+T} = I' | I_k = I, A_k{=}a)\Big]. \\
&= \sum_{y\in\mathcal{Y}} \mathbb{P}(u', y | u, I, a) \chi(I' = \mathbb{I}(a, I)) \chi(a' = \mathbb{A}_I(a, y)).
\end{aligned} \tag{5.40}$$

We remind that $\mathbb{P}(u', y|u, I, a)$ is given by (5.28)-(5.29). Let $\bar{R}^{\text{FSM}}_{\text{tot}}(g)$ and $\bar{E}^{\text{FSM}}_{\text{tot}}(g)$ be the total expected number of bits delivered and energy cost under FSM-HEU, starting from state g. Then, with $\mathbb{P}(g'|g)$ defined in (5.40) and $g = (u, I, a)$,

$$\bar{R}^{\text{FSM}}_{\text{tot}}(u,I,a) = r(u,I,a) + \sum_{(u',I',a')\in\mathcal{G}} \mathbb{P}(u',I',a'|u,I,a)\bar{R}^{\text{FSM}}_{\text{tot}}(u',I',a'),$$
$$\bar{E}^{\text{FSM}}_{\text{tot}}(u,I,a) = e(u,I,a) + \sum_{(u',I',a')\in\mathcal{G}} \mathbb{P}(u',I',a'|u,I,a)\bar{E}^{\text{FSM}}_{\text{tot}}(u',I',a'),$$

where $r(\cdot)$ and $e(\cdot)$ are given by (5.30)-(5.31). We can solve these equations in closed form, yielding

$$\bar{\mathbf{R}}^{\text{FSM}}_{\text{tot}} = (\mathbf{I} - \mathbf{P}^{\text{FSM}})^{-1}\mathbf{r}, \;\; \bar{\mathbf{E}}^{\text{FSM}}_{\text{tot}} = (\mathbf{I} - \mathbf{P}^{\text{FSM}})^{-1}\mathbf{e}, \tag{5.41}$$

where $\bar{\mathbf{R}}^{\text{FSM}}_{\text{tot}}=[\bar{R}^{\text{FSM}}_{\text{tot}}(g)]_{g\in\mathcal{G}}$, $\bar{\mathbf{E}}^{\text{FSM}}_{\text{tot}}=[\bar{E}^{\text{FSM}}_{\text{tot}}(g)]_{g\in\mathcal{G}}$, $\mathbf{r}=[r(g)]_{g\in\mathcal{G}}$, $\mathbf{e}=[e(g)]_{g\in\mathcal{G}}$, $[\mathbf{P}^{\text{FSM}}]_{g,g'}=\mathbb{P}(g'|g)$.

5.4.2 Belief-based Heuristic policy (B-HEU)

Unlike FSM-HEU, this policy exploits the POMDP state (β_k, I_k) in the decision-making process. However, B-HEU selects actions in a heuristic fashion as described next, as opposed to C-PBVI (Algorithm 1), which selects actions (approximately) optimally. The decision making under B-HEU are depicted in the flow chart of Fig. 5.4. To describe this policy, let (β, I) be the current POMDP state. Let $\Xi_I(j)$ be the marginal probability of the UE occupying the jth BPI with no blockage under the serving BS I, defined as

$$\Xi_I(j) \triangleq \frac{\sum\limits_{(s,b):(s_I,b_I)=(j,1)} \beta(s,b)}{\sum\limits_{j'\in\mathcal{S}_I}\sum\limits_{(s,b):(s_I,b_I)=(j',1)} \beta(s,b)}. \tag{5.42}$$

Then, $\Lambda_I \triangleq \sum\limits_{j\in\mathcal{S}_I}\sum\limits_{(s,b):(s_I,b_I)=(j,1)} \beta(s,b)$ can be interpreted as the probability of no blockage under the serving BS I. Given these quantities, B-HEU operates as follows, with thresholds η_1, η_2 and η_3 determined offline: if $\Lambda_I < \eta_1$, then blockage is detected, hence the HO action is selected; otherwise ($\Lambda_I \geq \eta_1$), let $\hat{j}_I = \arg\max_{j\in\mathcal{S}_I} \Xi_I(j)$ be the most likely BPI occupied by the UE: if $\Xi_I(\hat{j}_I) \geq \eta_2$, i.e., the serving BS I is confident that the UE belongs to BPI $\hat{j}_I \in \mathcal{S}_I$

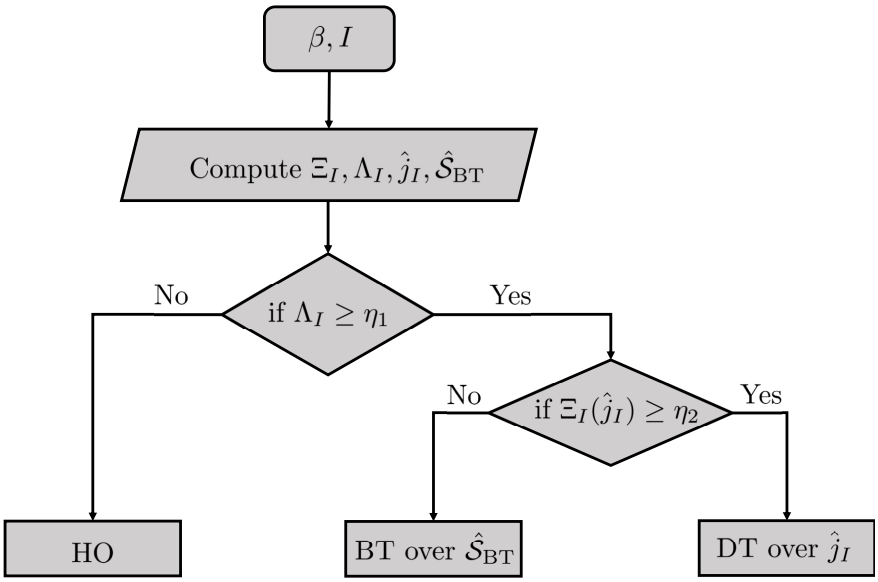

Figure 5.4. Flow chart for B-HEU Policy.

and there is no blockage, then the BS performs DT over BPI $\hat{j}_I$, with SNR $\mathrm{SNR}_{\mathrm{DT}}$ and duration T_{DT} determined offline. Otherwise ($\Lambda_I \geq \eta_1$ and $\Xi_I(\hat{j}_I) < \eta_2$), the BS is uncertain on the BPI of the UE, hence it performs BT over the smallest BPI set $\hat{\mathcal{S}}_{\mathrm{BT}}$ with aggregate probability greater or equal to η_3, defined as

$$\hat{\mathcal{S}}_{\mathrm{BT}} \triangleq \arg\min_{\mathcal{S}\subseteq\mathcal{S}_I} |\mathcal{S}| \text{ s.t.: } \sum_{j\in\mathcal{S}} \Xi_I(j) \geq \eta_3. \tag{5.43}$$

By doing so, it neglects the least likely set of beams whose aggregate probability is less than η_3.

After selecting the appropriate action based on the belief, the next serving BS with index $I' = \mathbb{I}(a, I)$ collects the observation Y_{k+T} and updates its belief using (5.32). Note that, unlike FSM-HEU which performs an exhaustive search during the BT phase, B-HEU exploits the current belief β to perform BT only on the most likely beams, and therefore reduces the BT overhead. However, it incurs higher complexity than FSM-HEU, since the belief needs to be tracked.

5.5 Numerical results

In this section, we perform numerical evaluations of the proposed policies. We compare their performance with a baseline policy, which is the same as FSM-HEU except for one key difference: after executing the DT action, it executes the BT action irrespective of the binary feedback. In other words, $\mathbb{A}_I((\mathrm{DT}, j), Y)=\mathrm{BT}, \forall Y$. Note that, if no blockage is detected, this baseline mimics the periodic exhaustive search. Its performance can be analyzed in closed form in a similar fashion as for FSM-HEU (see its FSM representation in Fig. 5.3).

The simulation parameters are listed in Table 5.1. The BSs and UE are both equipped with uniform planar arrays (deployed in the yz-plane) with $M_{\mathrm{tx}}^{(I)} = M_{\mathrm{tx},z}^{(I)} \times M_{\mathrm{tx},y}^{(I)}$ and $M_{\mathrm{rx}} = M_{\mathrm{rx},z} \times M_{\mathrm{rx},y}$ antennas, respectively. The BS and UE codebooks are based on array steering vectors, designed to provide coverage to a road segment of length 30m. For numerical simulation, we adopt a blockage dynamic model independent of the UE location, and with blockage states of the two BSs independent of each other. This models a worst-case scenario, where the blockage states of two BSs are independent and they show no correlation with the current and future UE position. In this case, the blockage transition probability can be expressed as $\mathbf{B}_{b'|bss'} = \mathbf{B}^{(0)}_{b'_0|b_0}\mathbf{B}^{(1)}_{b'_1|b_1}$. The transition probabilities can be expressed in terms of average blockage duration $D_0^{(I)}$[s] and steady state blockage probability $\pi_0^{(I)}$ as

$$\mathbf{B}_{01}^{(I)} = \frac{\Delta_t}{D_0^{(I)}}, \ \mathbf{B}_{10}^{(I)} = \frac{\pi_0^{(I)}}{1-\pi_0^{(I)}}\frac{\Delta_t}{D_0^{(I)}}. \tag{5.44}$$

Using the throughput and power metrics defined in (5.33), the average spectral efficiency (bps/Hz) under policy π is expressed as $\bar{T}^{\pi}/W_{\mathrm{tot}}$. We choose the initial BS $I = 1$ and the initial belief $\beta_0^*(u)=\chi(u=u_0)$, where $u_0=(s_0, b_0)$ with s_0 denoting the first pair of BS-UE BPI and $b_0 = (1, 1)$ denoting absence of blockage with respect to both BSs.

We define a 2D mobility model for a two lane straight highway with lane separation of $\Delta_{\mathrm{lane}} = 3.7$m as depicted in Fig 5.1.[5] The UE position along the road (y-axis) follows a

[5]↑The proposed system model and schemes can be used for multi-lane highway with any arbitrary road shape.

Table 5.1. Simulation parameters.

Parameter	Symbol	Value
Number of BS antennas	$M_{\text{tx}}^{(I)}$	$256 = (32 \times 8)$
Number of UE antennas	$M_{\text{rx}}^{(I)}$	$32 = (8 \times 4)$
Number of BS beam	$\lvert\mathcal{C}_I\rvert$	8
Number of UE beams	$\lvert\mathcal{F}\rvert$	8
Slot duration	Δ_t	100μs
Distance of BS to Rd center	D	22m
Lane separation	Δ_{lane}	3.5m
BS height	h_{BS}	10m
Bandwidth	W_{tot}	100MHz
Carrier frequency	f_c	30GHz
Noise psd	N_0	-174dBm/Hz
Noise figure	F	10dB
Sidelobe/mainlobe SNR ratio	ρ	-15dB
Fraction of DT slot for channel estimation	κ	0.01
HO delay	T_{HO}	1 slot
DT duration	T_{DT}	$\{20, 30, 40, 50\}$ slots
Steady state blockage prob.	$\pi_0^{(1)}, \pi_0^{(2)}$	0.2
Avg blockage duration	$D_0^{(1)}, D_0^{(2)}$	200ms
UE average speed	μ_v	30m/s
UE speed st. dev.	σ_v	10
UE mobility memory param.	γ	0.2
UE lane change prob.	$q_{1\to 2} = q_{2\to 1}$	0.01
Accuracy for Algorithm 2	ϵ_E, ϵ_V	0.01
B-HEU thresholds	(η_1, η_2, η_3)	(0.1,0.8,0.60)

Gauss-Markov mobility model and it changes lanes on the road with probability $q_{l\to l'}$. The speed V_k and position $X_{y,k}$ of the UE along the road (y-axis) follow the dynamics

$$\begin{cases} V_k = \gamma V_{k-1} + (1-\gamma)\mu_v + \sigma_v\sqrt{1-\gamma^2}\tilde{V}_{k-1}, \\ X_{y,k} = X_{y,k-1} + \Delta_{\text{t}} V_{k-1}, \end{cases} \tag{5.45}$$

where, unless otherwise stated, $\mu_v = 30$m/s is the average speed; $\sigma_v = 10$m/s is the standard deviation of speed; $\gamma = 0.2$ is the memory parameter; $\tilde{V}_{k-1} \sim \mathcal{N}(0,1)$, i.i.d. over slots. Note that, under this model, the SBPI $S_k = (s_0^*(X_k), s_1^*(X_k))$ does *not* follow Markovian dynamics, causing a mismatch between the analysis (based on the assumption of Markov state dynamics) and actual state trajectories (which do not follow Markovian dynamics). In addition, there is a mismatch between the sectored antenna model used in the analysis and the actual beamforming gain, which depends on the beam design and the actual AoA and

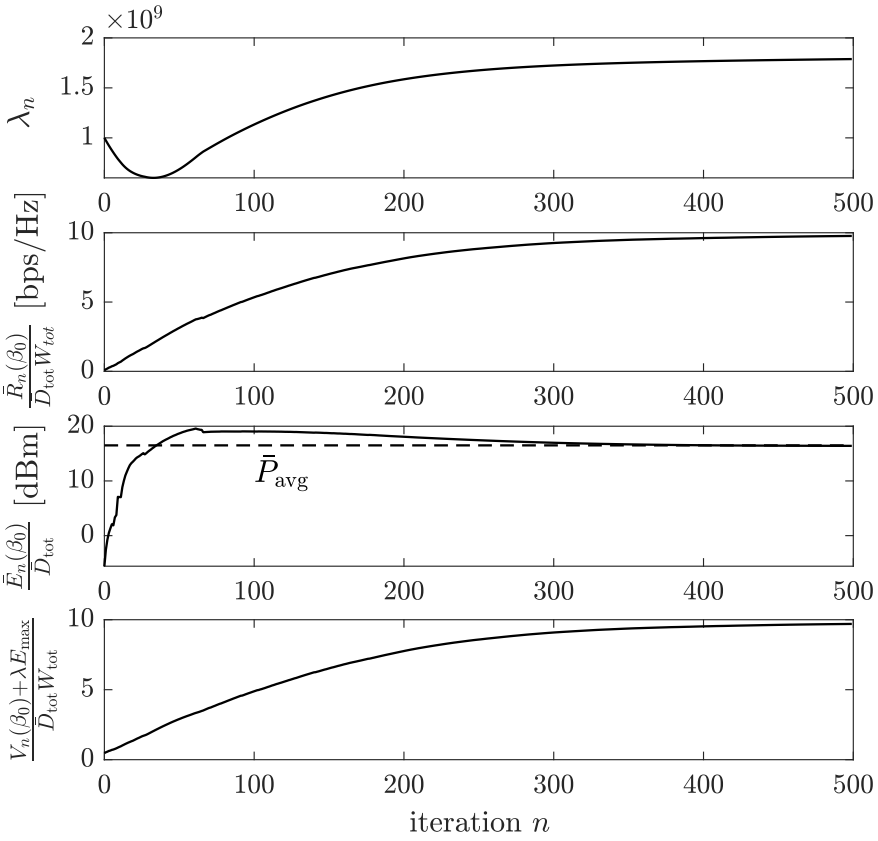

Figure 5.5. Convergence of C-PBVI Algorithm 2.

AoD associated with the current UE position X_k (see (5.2)). This mismatch might cause the POMDP based policy to underperform. To evaluate the accuracy of our analysis under this more realistic setting, in the simulations, we show the results corresponding to the analytical model presented in the paper – where the transition model $\mathbf{S}_{s'|s}$ is estimated from simulations of $10,000$ trajectories under the Gauss-Markov model (5.45), as described in Section 5.1.3 – as well as the results obtained through Monte-Carlo simulation using the array steering based analog beamforming and the Gauss-Markov mobility model: in this case, the position X_k is generated as in (5.45); the beamforming gain is based on the AoA and AoD associated with UE position X_k (see (5.2)) rather than the sectored antenna approximation used in the analytical model (see Section 5.1.4); the UE's feedback signal Y_k is generated as in (5.12); the belief is then updated using (5.32); actions are selected according to the policy under consideration – either based on the belief (C-PBVI and B-HEU policies) or feedback signaling (FSM-HEU and baseline policies). Table 5.1 summarizes the numerical parameters.

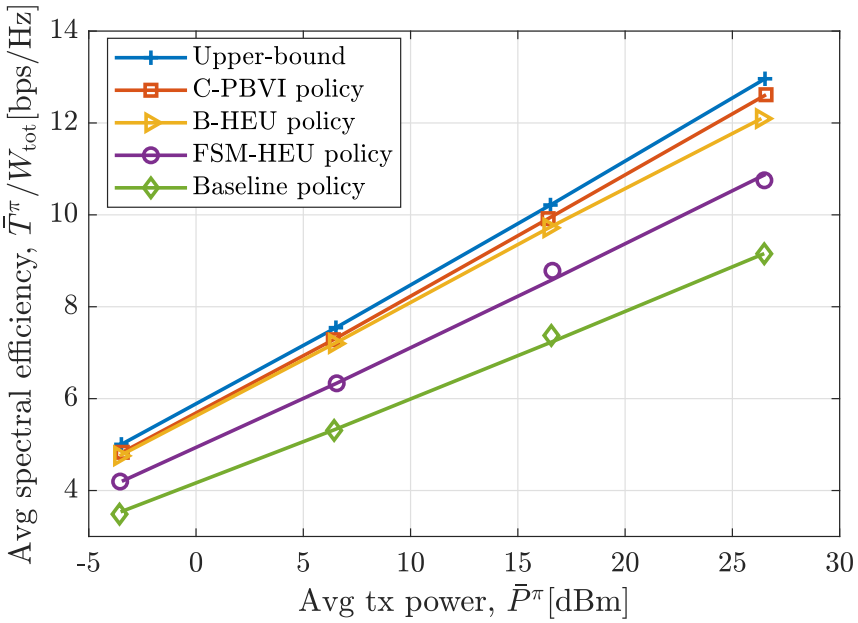

Figure 5.6. Average spectral efficiency versus average power consumption. The continuous lines represent the analytical curves based on the sectored model and synthetic mobility (generated based on the beam transition probability $\mathbf{S}_{ss'}$, see Eq. (5.5)), whereas the markers represent the simulation using analog beamforming and actual mobility.

In Fig. 5.5, we show the convergence of the C-PBVI Algorithm 2, which optimizes both the policy π and the dual variable λ to meet the power constraint $\bar{P}^{\pi} \leq \bar{P}_{\text{avg}}$. It can be observed that the dual variable λ, expected spectral efficiency $\bar{R}_n/\bar{D}_{\text{tot}}/W_{\text{tot}}$, average power $\bar{E}_n/\bar{D}_{\text{tot}}$ and Lagrangian function $[V_n(\beta_0)+\lambda_n E_{\max}]/\bar{D}_{\text{tot}}/W_{\text{tot}}$ converge, and $\bar{E}_n/\bar{D}_{\text{tot}}$ converges to the desired average power constraint $\bar{P}_{\text{avg}} = 16$dBm. In Fig. 5.6, we depict the average spectral efficiency versus the average power consumption. For the heuristic policies, we set T_{DT}=10 and SNR_{BT}=$\text{SNR}_{\text{DT}} = \text{SNR}_{\text{pre}} M_{\text{tx}}^{(I)} M_{\text{rx}}, \forall I \in \mathcal{I}$, where SNR_{pre}, representing the minimum pre-beamforming SNR, is varied from -12dB to 18dB.[6] The upper-bound shown in the figure is obtained by a genie-aided policy that always executes DT with perfect knowledge of the state (u, I). It should be noted that this upper-bound is loose since it is found by assuming perfect state knowledge. The C-PBVI policy π^* yields the best performance with negligible performance gap with respect to the upper-bound. It shows a performance gain of up to 4%, 17% and 38% compared to B-HEU, FSM-HEU and baseline,

[6] $\uparrow M_{\text{tx}}^{(I)} M_{\text{rx}}$ is the peak beamforming gain for array steering based analog beamforming [73].

respectively. It is also observed that B-HEU shows 12% performance gain over FSM-HEU. On the other hand, the baseline scheme yields up to 24% and 15% degraded performance compared to B-HEU and FSM-HEU, respectively: in fact, it neglects the DT feedback and instead performs periodic BT, thus incurring significant overhead. We also observe that the curves, obtained through the proposed analytical model, and the markers, representing simulation points obtained considering analog beam design and Gauss Markov mobility, closely match, thereby demonstrating the accuracy of our analysis in realistic settings.

In Fig. 5.7, we plot the spectral efficiency versus the DT time duration T_{DT} used in B-HEU, FSM-HEU and baseline schemes. As observed previously, the C-PBVI policy outperforms B-HEU and FSM-HEU, and all of them outperform the baseline scheme. B-HEU achieves near-optimal performance with an optimized value of $T_{\mathrm{DT}} \simeq 70$[slots] followed by FSM-HEU which performs best with $T_{\mathrm{DT}} \simeq 40$. Most remarkably, near-optimal performance is achieved by B-HEU at a fraction of the complexity of C-PBVI. It is observed that the spectral efficiency initially improves by increasing T_{DT} due to reduced overhead of BT and feedback time. However, after achieving a maximum value at an optimal T_{DT}, the spectral efficiency decreases as T_{DT} is further increased. This is attributed to the fact that during very large data transmission periods, loss of alignment and blockages are more likely to occur before the serving BS is able to react to these events. It is also observed that the baseline scheme achieves peak performance at a much higher value of $T_{\mathrm{DT}} \simeq 125$[slots]. In fact, since baseline performs periodic BT, it incurs severe overhead, hence there is a stronger incentive to reduce the overhead by extending the duration of DT, as opposed to B-HEU and FSM-HEU which adapt the duration of DT based on the DT feedback signal. In Fig. 5.8, we evaluate the impact of mobility and multiple users on blockage dynamics, based on the probabilistic model developed in [38]: this model defines a relationship between the dynamics of the blockage process, the number of UEs in the coverage area and their average speed. In fact, mobile UEs may cause time-varying obstructions of the signal (blockages) which may severely degrade the performance of vehicular mm-wave systems, especially in dense and highly-mobile scenarios. In the figure, we plot the total average spectral efficiency versus the number of users and the mean UE speed. The system performance is evaluated via Monte-Carlo simulation. Moreover, we assume that the proposed policies are executed in parallel

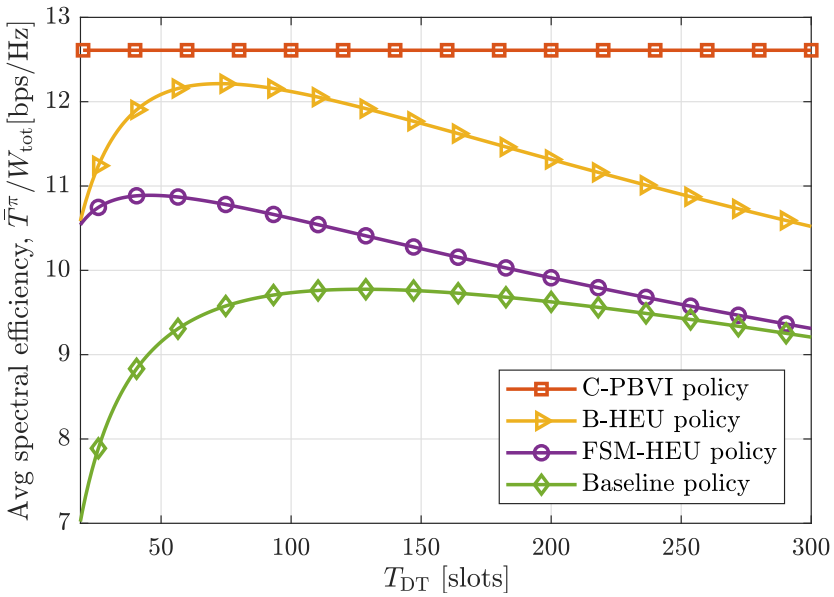

Figure 5.7. Average spectral efficiency versus T_{DT}; $\text{SNR}_{\text{pre}} = 18\text{dB}$.

across multiple UEs, using OFDMA [68] to orthogonalize their transmission resources. It can be seen that, for all policies, the spectral efficiency decreases as the mean speed increases: in fact, at higher speed, the UEs not only experience more frequent beam mis-alignments, but also the frequency of occurrence of blockages is exacerbated. The spectral efficiency also degrades as the number of UEs increases: in fact, nearby UEs contribute to creating obstructions and more frequent blockages, as well as a reduced time duration for the unblocked intervals. As previously noted, B-HEU achieves the best performance, followed by FSM-HEU and baseline. Most importantly, the two heuristics B-HEU and FSM-HEU achieve 50% and 25% higher spectral efficiency than the baseline scheme, respectively, demonstrating their robustness in mobile and dense user scenarios.

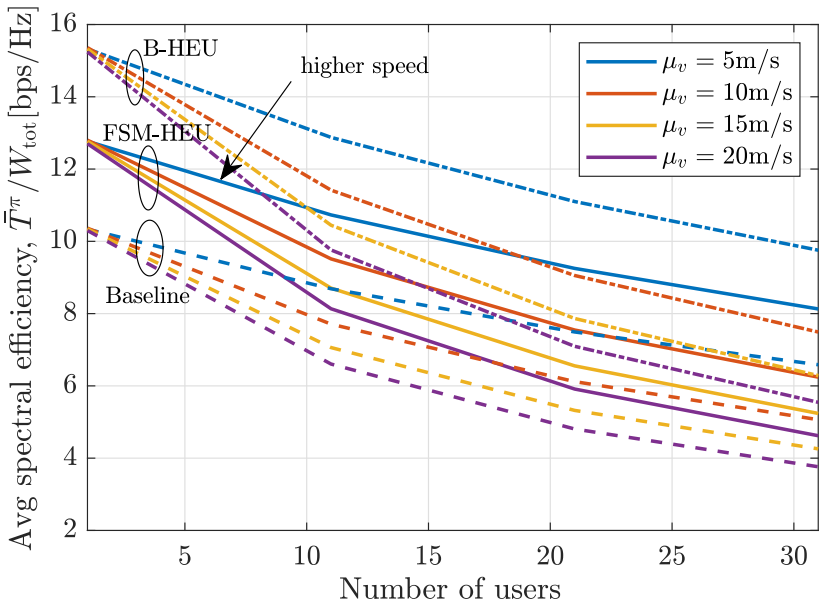

Figure 5.8. Total average spectral efficiency versus number of UEs for different UE mean speed μ_v; $\sigma_v = 10$m/s, $\text{SNR}_{\text{pre}} = 18$dB, $T_{\text{DT}} = 50$.

6. LEARNING AND ADAPTATION IN MILLIMETER-WAVE COMMUNICATIONS VIA DEEP VARIATIONAL AUTOENCODERS AND POMDPS

Millimeter-wave vehicular networks using narrow-beam communications incur enormous beam-training overhead. To mitigate it, this chapter proposes a learning and adaptation framework, in which the dynamics of the communication beams are learned and then exploited to design adaptive beam-training procedures. Specifically, a dual timescale approach is proposed: on a long timescale, a recurrent deep variational autoencoder (R-VAE) uses noisy beam-training observations to learn a probabilistic model of beam dynamics; on a short timescale, an adaptive beam-training procedure is formulated as a partially observable (PO-) Markov decision process (MDP), and optimized using *point-based value iteration* (PBVI) by leveraging beam-training feedback and a probabilistic knowledge of the strongest beam pair provided by the R-VAE. In turn, beam-training observations are used to refine the R-VAE via stochastic gradient descent in a continuous process of learning and adaptation. It is shown that the proposed R-VAE mobility learning framework learns accurate beam dynamics: it reduces the Kullback-Leibler divergence between the ground-truth and the learned beam dynamics model by 86%, with respect to the Baum-Welch algorithm and by 92% with respect to a naive mobility learning approach that neglects feedback errors. The proposed dual timescale approach yields negligible loss of spectral efficiency with respect to a genie-aided scheme that operates under error-free feedback and knowledge of the ground-truth mobility model. Finally, a low-complexity policy is proposed by reducing the POMDP to an MDP. It is shown that the PBVI-based and MDP-based policies yield a spectral efficiency gain of up to 46% and 37%, respectively, over a policy that scans exhaustively the likely beam pairs.

↑A version of this chapter is pending publication in IEEE Transactions on Vehicular Technology.

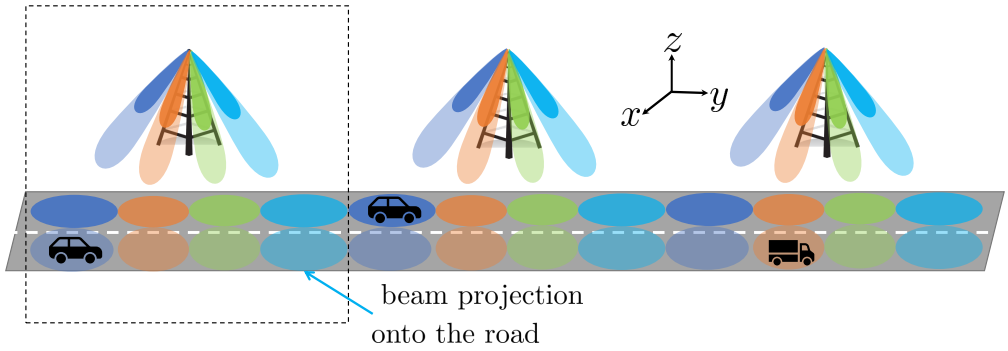

Figure 6.1. A mobile millimeter wave network.

6.1 System Model

We consider a mobile millimeter-wave network scenario as depicted in Fig. 6.1. The UE is moving on the road, covered by multiple roadside base-stations (BSs). At each time, the UE is connected to one BS, referred to as the serving BS. The BS and UE both use 3D beamforming with large antenna arrays to achieve highly directional communications. If the UE exits its serving BS coverage area, a handover is performed to the next BS, at which point the data/control planes are transferred to the next serving BS. In this chapter, we restrict the beam alignment and communication design to the UE and its serving BS, as depicted in the box in Fig. 6.1.

We consider a time-slotted system: frames of duration T_{frame} are divided into K slots, each of duration $T_{\text{slot}} \triangleq T_{\text{frame}}/K$. We assume $T_{\text{frame}} \leq T_b$, where T_b is the *beam coherence time*, i.e., the time duration over which the BS-UE beams remain aligned. For example, using the analysis of [58], $T_b \simeq 1$[s] for a UE velocity of 100[km/h]. Each frame is split into a beam-training (BT) phase of variable duration, followed by a data transmission (DT) phase until the end of the frame, as shown in Fig. 6.2.

The mobility of the UE and of the surrounding propagation environment induce mobility in the optimal beams that should be used at the transmitter and receiver for data communication. Since $T_{\text{frame}} \leq T_b$, the optimal beam pair is assumed to remain constant during the entire frame duration, but may change across frames. The goal pursued in this chapter is to learn these dynamics to enable predictions of the optimal beam pair and reduce the beam-training overhead.

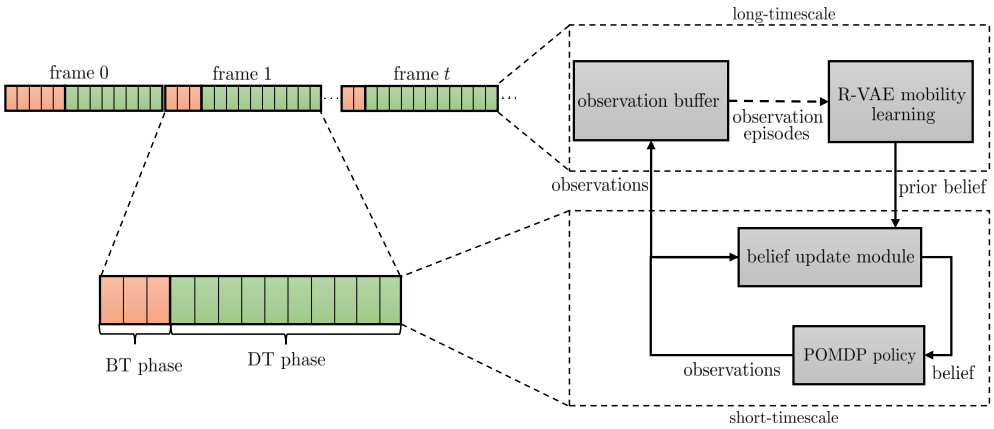

Figure 6.2. Beam-training and data transmission phases. Short-timescale interactions shown with solid arrows; long-timescale interactions are shown by dashed arrows;

To achieve this goal, in this chapter we propose a learning and adaptation framework based on a dual timescale approach, depicted as a block diagram in Fig. 6.2: the short timescale is the duration of one frame T_{frame}; the long-timescale refers to the duration of time during which the UE stays within the BS coverage area, of the order of several hundred frames. In the long-timescale, the mobility learning module aims to learn the dynamics of the optimal beam pair induced by the mobility of the UE and of the propagation environment, based on previous interactions with the UE, and with previous UEs. In the short timescale, the goal is to maximize the throughput within the frame duration, by optimizing the BT and DT strategy and by exploiting beam-training feedback as well as prior statistical information on the optimal beam pair (prior belief) provided by the mobility learning module. To achieve the desired goals of the dual timescale strategy, we make the following design choices, depicted in Fig. 6.2:

1. We model the dynamics of the strongest beam pair index (SBPI) as a Markov process. The BS leverages the mobility model to provide side information at the start of each frame, represented as a prior belief over SBPIs (a probability distribution over SBPIs).

2. We formulate a POMDP (Section 6.2) that leverages the prior belief to optimize the decision-making process on the short-timescale, and propose a PBVI algorithm (Section 6.2.1) to find the approximately optimal policy. The policy uses the prior belief over

the SBPI and BT feedback received within the frame to adaptively select the BT and DT actions on the short-timescale.

3. We develop an R-VAE-based framework (Section. 6.3) to learn the Markov SBPI model, based on SNR observations acquired during previous interactions (see Fig. 6.2) and collected in the observation buffer.

Next, we describe the signal and channel models.

6.1.1 Channel and Signal Model

Let $\mathbf{x}\in\mathbb{C}^{L_{\text{sy}}}$ be the transmitted signal with $\mathbb{E}[\|\mathbf{x}\|_2^2]=L_{\text{sy}}$, where L_{sy} denotes the number of symbols transmitted. The signal received at the UE in frame $t \in \mathbb{N}$ and slot $k \in \mathcal{K} \triangleq \{0, 1, \ldots, K-1\}$ within the frame is expressed as

$$\mathbf{y}_{t,k} = \sqrt{P_{t,k}}\mathbf{f}_{t,k}^H \mathbf{H}_{t,k}\mathbf{c}_{t,k}\mathbf{x} + \mathbf{w}_{t,k}, \tag{6.1}$$

where $P_{t,k}$ is the average transmit power of the BS; $\mathbf{c}_{t,k}\in\mathbb{C}^{M_{\text{tx}}\times 1}$ and $\mathbf{f}_{t,k}\in\mathbb{C}^{M_{\text{rx}}\times 1}$ are unit-norm beamforming vectors at the BS and UE, with M_{tx} and M_{rx} antennas, respectively; $\mathbf{w}_{t,k} \sim \mathcal{CN}(\mathbf{0}, \sigma_w^2\mathbf{I})$ is the additive white Gaussian noise (AWGN), with variance σ_w^2. In this chapter, we adopt the diffused multipath channel model with one dominant line of sight (LOS) path, used also in our previous work [7], expressed as

$$\mathbf{H}_{t,k}=\underbrace{\sqrt{M_{\text{tx}}M_{\text{rx}}}h_{t,k}\mathbf{d}_{\text{rx}}(\theta_t)\mathbf{d}_{\text{tx}}(\phi_t)^H}_{\mathbf{H}_{t,k,\text{LOS}}} + \underbrace{\sum_{l=1}^{N_{\text{DIF}}}\sqrt{M_{\text{tx}}M_{\text{rx}}}\tilde{h}_{t,k,l}\mathbf{d}_{\text{rx}}(\tilde{\theta}_{t,k,l})\mathbf{d}_{\text{tx}}(\tilde{\phi}_{t,k,l})^H}_{\mathbf{H}_{t,k,\text{DIF}}}, \tag{6.2}$$

where $h_{t,k} \sim \mathcal{CN}(0, \sigma_{h,t}^2)$ is the complex gain of the LOS component, with $\sigma_{h,t}^2 = 1/\ell_t$; $\ell_t = [4\pi/\lambda_c]^2 d_t^2$ is the path loss as a function of the UE-BS's distance d_t; θ_t and ϕ_t are the angle of arrival (AoA) and of departure (AoD) of the LOS path (azimuth and elevation angles), respectively. $\mathbf{d}_{\text{tx}}(\phi)$ and $\mathbf{d}_{\text{rx}}(\theta)$ are the array response vectors of the BS and UE antenna arrays, respectively; $\tilde{h}_{t,k,l}$, $\tilde{\theta}_{t,k,l}$ and $\tilde{\phi}_{t,k,l}$ denote the complex channel gain, AoA and AoD of the l^{th} diffused multipath component, respectively. Note that d_t, θ_t and ϕ_t may

evolve across frames as a result of mobility of the UE and of the surrounding propagation environment, however, they remain fixed during the frame duration, as subsumed by the assumption $T_{\text{frame}} \leq T_b$ discussed earlier. On the other hand, $h_{t,k}$ is i.i.d over the frame slots. We model $\mathbf{H}_{t,k,\text{DIF}}$ as zero mean Gaussian with i.i.d entries (over slot indices and antenna elements), each with variance σ^2_{DIF}. Experimental studies in [59] demonstrated that the diffused multipath components are much weaker than the LOS path (up to 100× weaker at a BS-UE distance of only 10 meters), so that $\sigma^2_{\text{DIF}} \ll \sigma^2_{h,t}$. Let $\ell(x)$, $\phi(x)$ and $\theta(x)$ be the pathloss, AoD and AoA of the LOS path when the UE is in position $x \in \mathcal{X}$ within the coverage are $\mathcal{X}$ of the BS. Let X_t be the UE's position in frame t, $G_{\text{tx}}(\mathbf{c}, x) = M_{\text{tx}} |\mathbf{d}_{\text{tx}}(\phi(x))^H \mathbf{c}|^2$ and $G_{\text{rx}}(\mathbf{f}, x) = M_{\text{rx}} |\mathbf{d}_{\text{rx}}(\theta(x))^H \mathbf{f}|^2$ be the beamforming gains of the BS and UE, respectively, with respect to the LOS path, and $\Theta_{t,k} = \angle \mathbf{d}_{\text{tx}}(\phi(X_t))^H \mathbf{c}_{t,k} + \angle \mathbf{f}_{t,k}^H \mathbf{d}_{\text{rx}}(\theta(X_t))$ be the unknown phase of the overall gain. Then, the received signal is expressed as

$$\mathbf{y}_{t,k} = \sqrt{P_{t,k}} \left[h_{t,k} \sqrt{G_{\text{tx}}(\mathbf{c}_{t,k}, X_t) G_{\text{rx}}(\mathbf{f}_{t,k}, X_t)} e^{j\Theta_{t,k}} + \Omega_{t,k} \right] \mathbf{x}_{t,k} + \mathbf{w}_{t,k}, \tag{6.3}$$

where $\Omega_{t,k} \triangleq \mathbf{f}_{t,k}^H \mathbf{H}_{t,k,\text{DIF}} \mathbf{c}_{t,k} \sim \mathcal{CN}(0, \sigma^2_{\text{DIF}})$ is the contribution due to the diffuse multipath channel components. The SNR averaged over the fading coefficients is then given as

$$\text{SNR}_{t,k} = \frac{P_{t,k}}{\sigma_w^2} \left[\frac{G_{\text{tx}}(\mathbf{c}_{t,k}, X_t) G_{\text{rx}}(\mathbf{f}_{t,k}, X_t)}{\ell(X_t)} + \sigma^2_{\text{DIF}} \right]. \tag{6.4}$$

6.1.2 Codebook Structure

The BS and UE both use pre-designed analog beamforming codebooks, denoted by $\mathcal{C} \triangleq \{\mathbf{c}_1, \ldots, \mathbf{c}_{|\mathcal{C}|}\}$ and $\mathcal{F} \triangleq \{\mathbf{f}_1, \ldots, \mathbf{f}_{|\mathcal{F}|}\}$, respectively. Let $\mathcal{V} \triangleq \mathcal{C} \times \mathcal{F}$ denote the joint codebook containing all possible beamforming codeword pairs of the BS and UE. These are indexed by the *beam pair index* (BPI), taking values from $\mathcal{S} \triangleq \{1, 2, \ldots, |\mathcal{C}||\mathcal{F}|\}$. We denote the jth beam pair by $(\mathbf{c}^{(j)}, \mathbf{f}^{(j)}) \in \mathcal{V}$. Let $G(j, x) \triangleq G_{\text{tx}}(\mathbf{c}^{(j)}, x) G_{\text{rx}}(\mathbf{f}^{(j)}, x)$ be the beamforming gain achieved under the jth BPI with the UE in position x. Similarly to [7], we

define the strongest BPI (SBPI) at a given UE's position x as the one achieving maximum beamforming gain,

$$s^*(x) \triangleq \arg\max_{j\in\mathcal{S}} G(j,x). \tag{6.5}$$

It follows that the optimal beam pair that should be used in the tth frame to carry out the data communication is the one indexed by $s^*(X_t)$, based on the current UE's position X_t.

6.1.3 Sectored Antenna Model

In this chapter, we use the *sectored antenna model* to approximate the beamforming gain [4], [7], [41], which provides an analytically tractable yet valuable approximation of the actual beam pattern, as demonstrated in our numerical evaluations in Section 6.4. We partition the coverage region $\mathcal{X}$ into $|\mathcal{S}|$ regions $\{\mathcal{X}_j, j\in\mathcal{S}\}$, where $\mathcal{X}_j = \{x\in\mathcal{X} : s^*(x) = j\}$ is the set of positions (possibly, empty) in which the SBPI is j. We denote the condition $x\in\mathcal{X}_j$ as the *alignment condition* along BPI j; conversely, we denote the condition $x\notin\mathcal{X}_j$ as the *misalignment condition* along BPI j. Consider the BPI j. Under the alignment condition (i.e., $x\in\mathcal{X}_j$), we approximate the SNR as

$$\mathrm{SNR}_{\mathrm{align}} = \frac{P_j}{\sigma_w^2}\left[\Upsilon_j + \sigma_{\mathrm{DIF}}^2\right] \tag{6.6}$$

where $\Upsilon_j = \min_{x\in\mathcal{X}_j} G(j,x)/\ell(x)$. Note that (6.6) also gives the transmit power P_j required to achieve a certain target SNR $\mathrm{SNR}_{\mathrm{align}}$ along the BPI j, under the alignment condition. On the other hand, under the misalignment condition $x\notin\mathcal{X}_j$, we model the mis-aligned SNR as

$$\mathrm{SNR}_{\mathrm{misalign}} \le \rho\mathrm{SNR}_{\mathrm{align}}, \tag{6.7}$$

where $\rho \ll 1$ is the ratio of the worst-case misaligned SNR and aligned SNR. In this case, data transmission is in outage since $\rho \ll 1$. Let $S_t \triangleq s^*(X_t)$ denote the current SBPI. With these definitions, the achieved SNR is only a function of whether the alignment is achieved

or not with the current choice of the BPI j: if $j = S_t$ is the selected BPI, then alignment is achieved, and the received is SNR $\text{SNR}_{\text{align}}$; in contrast, if $j \neq S_t$, then misalignment is achieved, and the received SNR is $\text{SNR}_{\text{misalign}}$.

6.1.4 Strongest Beam Pair Index Dynamics

The UE mobility along the road and/or temporal environment changes induce temporally correlated dynamics on the SBPI S_t, which can be exploited to reduce the training overhead. We assume that the process $\{S_t, t \geq 0\}$ is stationary Markovian. We will demonstrate numerically in Section 6.4 that this assumption yields a good approximation of non-Markovian dynamics (for instance, a vehicle moving at constant speed along a road). Let $p(s'|s) \triangleq \mathbb{P}(S_{t+1}{=}s'|S_t{=}s), \forall s \in \mathcal{S}, \forall s' \in \bar{\mathcal{S}}$ be the one-frame transition probability from the current SBPI s to the next SBPI s'. Here, $\bar{\mathcal{S}} \triangleq \mathcal{S} \cup \{\bar{s}\}$ includes the additional state $\bar{s}$, which indicates that the UE has moved out of the coverage area of the BS and can no longer be served. In practice, the transition model $p(\cdot|\cdot)$ is not known a-priori and has to be estimated using the history of observation (BT/DT actions and their feedback) sent to the BS from the UE. This represents a departure from our previous work [7], which assumed prior knowledge of $p(\cdot|\cdot)$.

A straightforward approach to estimate $p(\cdot|\cdot)$ is to use the detected SBPI obtained during the BT procedure (e.g., exhaustive search) to generate a sequence $\{(\hat{s}_t, \hat{s}_{t+1}), t \in \mathcal{T}_{\text{sound}}\}$ of SBPIs' transitions recored at time frames $t \in \mathcal{T}_{\text{sound}}$. Then, $p(\cdot|\cdot)$ may be estimated as

$$\hat{p}(s'|s) = \frac{\sum_{t \in \mathcal{T}_{\text{sound}}} 1[\hat{s}_t = s, \hat{s}_{t+1} = s']}{\sum_{t \in \mathcal{T}_{\text{sound}}} 1[\hat{s}_t = s]}, \ \forall s \in \mathcal{S}, \ s' \in \bar{\mathcal{S}}, \tag{6.8}$$

where $1[\cdot]$ is the indicator function. However, transition models estimated with this approach, which we term *naive mobility learning*, suffer from errors in the detected SBPI $\hat{s}_t$ caused by noise and beam imperfections. The design of estimation procedures that are robust against measurement noise and beam imperfections is developed in Section 6.3 using a novel technique based on recurrent variational auto-encoder (R-VAE).

6.1.5 Beam-Training (BT) and Data Transmission (DT)

We now introduce the BT and DT operations. As shown in Fig. 6.2, each frame comprises a BT phase of variable duration and a DT phase for the remainder of the frame. In the following, we outline the BT and DT phases and describe the feedback model.

BT phase: Within the BT phase, the BS selects and executes a sequence of BT actions. For each BT action, a set of BPIs $\hat{\mathcal{S}} \subseteq \mathcal{S}$ is first chosen; then, a sequence of beacons $\mathbf{x}$ are sent in sequence over $|\hat{\mathcal{S}}|$ slots, using one slot for each BPI $j \in \hat{\mathcal{S}}$, during which the BS transmits using the beamforming vector $\mathbf{c}^{(j)}$, while the UE receives synchronously using the combining vector $\mathbf{f}^{(j)}$. After the sequence of beacons have been transmitted, an additional slot is used for feedback from the UE back to the BS, so that the overall duration of the BT action $\hat{\mathcal{S}}$ is $L \triangleq |\hat{\mathcal{S}}|+1$. The feedback signal is generated as follows, similarly to [7]. Letting i_j be the slot index over which BPI $j \in \hat{\mathcal{S}}$ is transmitted, the UE process the received signal $\mathbf{y}_{t,i_j}$ using a matched filter as

$$\Gamma_t^{(j)} \triangleq \frac{|\mathbf{x}^H \mathbf{y}_{t,i_j}|^2}{\sigma_w^2 \|\mathbf{x}\|_2^2}, \tag{6.9}$$

which represents an estimate of the SNR when BPI j is used.

Upon collecting the sequence $\{\Gamma_t^{(j)}, \forall j \in \hat{\mathcal{S}}\}$, the UE detects the BPI with strongest signal as

$$Y = \begin{cases} j^* \triangleq \arg\max_{j \in \hat{\mathcal{S}}} \Gamma_t^{(j)}, & \max_{j \in \hat{\mathcal{S}}} \Gamma_t^{(j)} > \eta, \\ \emptyset, & \max_{j \in \hat{\mathcal{S}}} \Gamma_t^{(j)} \leq \eta. \end{cases} \tag{6.10}$$

In other words, if all the SNR estimates are smaller than a threshold η, $Y=\emptyset$ indicates that no beam pair in $\hat{\mathcal{S}}$ is aligned. Otherwise, $Y=j^*$ indicates the index of the SBPI detected. The feedback signal is then sent back to the BS.

The probabilistic analysis of the BT feedback can be carried out as in [7], where closed-form expressions of $\mathbb{P}(y|s,\hat{\mathcal{S}})$ are derived. Herein, we briefly describe the BT feedback distribution, without providing explicit expressions. Let $\hat{\mathcal{S}}$ be the BT action selected. Then, when the BPI $j \in \hat{\mathcal{S}}$ is scanned, the corresponding matched filter output $\Gamma_t^{(j)}$ has exponential

distribution with mean $1+\text{SNR}_{\text{align}}$ if $S_t = j$ or $1+\text{SNR}_{\text{misalign}}$ if $S_t \neq j$. The feedback distribution $\mathbb{P}(Y = y|S_t = s, \hat{\mathcal{S}})$ can then be derived in closed form [7] for the two cases $S_t \in \hat{\mathcal{S}}$ or $S_t \notin \hat{\mathcal{S}}$. In particular, for the case when $S_t \in \hat{\mathcal{S}}$, the closed-form expressions of the probability of mis-detection $\mathbb{P}(Y = \emptyset|S_t = s, \hat{\mathcal{S}}, s \in \mathcal{S}) \triangleq p_{\text{md}}(\eta, |\hat{\mathcal{S}}|, \text{SNR})$ and the probability of correct detection $\mathbb{P}(Y = s|S_t = s, \hat{\mathcal{S}}, s \in \mathcal{S}) \triangleq p_{\text{correct}}(\eta, |\hat{\mathcal{S}}|, \text{SNR})$ can be derived in closed-form, each as function of threshold η, number of BPIs $|\hat{\mathcal{S}}|$ and target SNR SNR. Then, the probability of incorrect detection is computed by using the fact that incorrect detection is i.i.d among $|\hat{\mathcal{S}}|-1$ BPIs in $\hat{\mathcal{S}} \setminus \{s\}$, as follows

$$\mathbb{P}(Y \in \hat{\mathcal{S}} \setminus \{s\}|S_t = s, \hat{\mathcal{S}}, s \in \mathcal{S}) \triangleq \frac{1 - p_{\text{correct}}(\eta, |\hat{\mathcal{S}}|, \text{SNR}) - p_{\text{md}}(\eta, |\hat{\mathcal{S}}|, \text{SNR})}{|\hat{\mathcal{S}}|-1} \tag{6.11}$$

On the other hand, for the mis-aligned case ($S_t \notin \hat{\mathcal{S}}$), the probability of making no false-alarm $\mathbb{P}(Y = \emptyset|S_t = s, \hat{\mathcal{S}}, s \notin \mathcal{S}) \triangleq 1 - p_{\text{fa}}(\eta, |\hat{\mathcal{S}}|, \text{SNR})$ can be computed in closed-form as function of the threshold η, number of BPIs $|\hat{\mathcal{S}}|$, and target SNR SNR. Then, the probability of making false alarm to any BPI in $\hat{\mathcal{S}}$ is computed as

$$\mathbb{P}(Y \in \hat{\mathcal{S}}|S_t = s, \hat{\mathcal{S}}, s \notin \mathcal{S}) \triangleq \frac{p_{\text{fa}}(\eta, |\hat{\mathcal{S}}|, \text{SNR})}{|\hat{\mathcal{S}}|} \tag{6.12}$$

For any given target SNR SNR and number of SBPI scanned $|\hat{\mathcal{S}}|$, the threshold η is determined by enforcing the total false alarm probability equal to the total misdetection probability and solving for η, yielding

$$p_{\text{mf}}(|\hat{\mathcal{S}}|, \text{SNR}) = p_{\text{md}}(\eta^*, |\hat{\mathcal{S}}|, \text{SNR}) = p_{\text{fa}}(\eta^*, |\hat{\mathcal{S}}|, \text{SNR}), \tag{6.13}$$

where η^* is solution to $p_{\text{md}}(\eta, |\hat{\mathcal{S}}|, \text{SNR}) = p_{\text{fa}}(\eta, |\hat{\mathcal{S}}|, \text{SNR})$. We omit the exact analysis of the feedback due to space constraint and can be found in [7]. In the next section, we present the POMDP, which exploits the feedback model.

DT phase: At the start of the DT phase, the BS chooses a BPI $\hat{s}_{\text{DT}}$. Data transmission then occurs with a fixed rate R until the end of the frame, by having the BS transmit with

beamforming vector $\mathbf{c}^{(\hat{s}_{\mathrm{DT}})}$ and the UE receive with combining vector $\mathbf{f}^{(\hat{s}_{\mathrm{DT}})}$. If $S_t \neq \hat{s}_{\mathrm{DT}}$, the selected BPI $\hat{s}_{\mathrm{DT}}$ is mis-aligned, resulting in communication outage; otherwise ($S_t{=}\hat{s}_{\mathrm{DT}}$, i.e., the selected BPI $\hat{s}_{\mathrm{DT}}$ is aligned) from the signal model (6.3), we find that a successful transmission occurs if

$$R < W_{\mathrm{tot}} \log_2(1 + |h_{t,k}|^2 \ell(X_t)\mathrm{SNR}_{\mathrm{align}}), \tag{6.14}$$

where W_{tot} is the bandwidth. Then, using the fact that $h_{t,k} \sim \mathcal{CN}(0, \sigma_{h,t}^2)$, the success probability

$$\mathbb{P}_{\mathrm{succ}}{=}\mathbb{P}\Big(|h_{t,k}|^2\ell(X_t){>}\frac{2^{R/W_{\mathrm{tot}}}{-}1}{\mathrm{SNR}_{\mathrm{align}}}\Big){=}1{-}\exp\Big\{-\frac{2^{R/W_{\mathrm{tot}}}-1}{\mathrm{SNR}_{\mathrm{align}}}\Big\}, \tag{6.15}$$

and the expected throughput per slot $\bar{R} = \mathbb{P}_{\mathrm{succ}} R$.

6.2 Short Timescale: Adaptive BT via Point-based Value iteration

This section introduces the POMDP and proposes an efficient PBVI algorithm to determine an approximately optimal BT and DT policy within the frame duration. The individual components of the POMDP are defined as follows.

Time horizon: $\mathcal{K} = \{0, 1, \ldots, K-1\}$ denotes the time horizon of the decision period, corresponding to the frame duration (K slots per frame).

State space: the set of BPIs $\mathcal{S}$; the state at frame t is the unobserved SBPI $S_t \in \mathcal{S}$, which remains constant during the frame duration, but may change from one frame to the next according to the transition model $p(\cdot|\cdot)$ (see Sec. 6.1.4).

Prior belief $\beta_{t,0}$: the probability mass function of the SBPI S_t available at the beginning of the frame, before BT is executed, so that $\beta_{t,0}(s) = \mathbb{P}(S_t = s)$. The prior belief $\beta_{t,0}$ is provided by the mobility learning module discussed in Sec. 6.3.

Action space: In slot k within the frame, the BS selects whether to perform BT or DT. If DT is selected, then the BS also selects the BPI $\hat{s}_{\mathrm{DT}} \in \mathcal{S}$ used for data communication for the remainder of the frame; the DT action space is then $\mathcal{A}_{\mathrm{DT}} \equiv \mathcal{S}$. Otherwise (BT is selected), the BS selects a set of BPIs $\hat{\mathcal{S}}$ to scan; since the duration of this action is $L{\triangleq}|\hat{\mathcal{S}}|{+}1$,

the selected action must be such that $|\hat{\mathcal{S}}| \leq K-1-k$, so that the BT action space in slot k is given by

$$\mathcal{A}_{\mathrm{BT,k}} \equiv \{\hat{\mathcal{S}} \subseteq \mathcal{S} : |\hat{\mathcal{S}}| \leq K-1-k\}.$$

Note that if a DT action is selected, it is used until the end of the frame, so that the decision period terminates; otherwise, when a BT action of duration L is chosen at time k, the next decision is taken in the slot $k+L$.

Observation model: After taking a BT action $\hat{\mathcal{S}}$, the BS observes the feedback signal Y, taking value from the observation space $\mathcal{Y} \triangleq \hat{\mathcal{S}} \cup \{\emptyset\}$ as described in Section 6.1.5.

Reward: we measure the reward as the expected number of data communication bits successfully delivered to the UE. Hence, under a BT action, no reward is accrued. On the other hand, if the DT action $\hat{s}_{\mathrm{DT}} \in \mathcal{S}$ is selected in slot $k \in \mathcal{K}$, with $S_t = s$ being the ground truth SBPI, then the reward accrued is

$$r_k(s, \hat{s}_{\mathrm{DT}}) = (K-k) T_{\mathrm{slot}} \bar{R} \cdot 1[s = \hat{s}_{\mathrm{DT}}], \tag{6.16}$$

i.e., it is zero if the SBPI is detected incorrectly ($s \neq \hat{s}_{\mathrm{DT}}$), otherwise it is equal to the total expected throughput delivered during the remaining frame duration $K-k$.

Belief Update: Since the SBPI $S \in \mathcal{S}$ is not observable, we use the belief β over $S \in \mathcal{S}$ as a POMDP state. The belief is the probability distribution over $s \in \mathcal{S}$, given the history of actions and observations from the beginning of the current frame until the current slot. Initially, $\beta_{t,0}$ is the prior belief at the beginning of the frame, provided by the mobility learning module. Let $\beta_{t,k}$ be the belief in slot k. This is updated whenever BT observations are received. If a DT action is taken in slot k, then it is executed until the end of the frame, so that a belief update is not required in this case and $\beta_{t,\ell} = \beta_{t,k},\ \forall \ell = k, \ldots, K$. Otherwise, consider the BT action $\hat{\mathcal{S}}$ of duration L taken in slot k. Since the corresponding feedback signal is received at the end of slot $k+L-1$, no belief updates occur in slots $k, \ldots, k+L-1$,

hence $\beta_{t,\ell} = \beta_{t,k},\ \forall \ell = k, \ldots, k+L-1$. Upon receiving the feedback signal $Y_{k+L-1} = y$ at the end of slot $k+L-1$, the BS updates the belief in slot $k+L$ based on Bayes' rule as

$$\beta_{t,k+L}(s) = \frac{\beta_{t,k+L-1}(s)\mathbb{P}(y|s,\hat{\mathcal{S}})}{\sum_{j\in\mathcal{S}} \beta_{t,k+L-1}(j)\mathbb{P}(y|j,\hat{\mathcal{S}})},\ \forall s \in \mathcal{S}, \tag{6.17}$$

where $\mathbb{P}(y|s,\hat{\mathcal{S}})$ is the feedback distribution given in Section 6.1.5. We define the belief update via the mapping $\mathbb{B}(\beta, y, \hat{\mathcal{S}})$, so that $\beta_{t,k+L} = \mathbb{B}(\beta_{t,k+L-1}, Y_{k+L-1}, \hat{\mathcal{S}})$, expressed as the function of the previous belief $\beta_{t,k+L-1}$, feedback Y_{k+L-1} and BT action $\hat{\mathcal{S}}$. At the end of the frame, with $\beta_{t,K}$ computed with the procedure described above, the prior belief for the next frame can be computed by the mobility learning module using the learned transition model $\hat{p}(\cdot|\cdot)$ as

$$\beta_{t+1,0}(s') = \sum_{s\in\mathcal{S}} \hat{p}(s'|s)\beta_{t,K}(s), \forall s' \in \mathcal{S}, \tag{6.18}$$

so that the POMDP optimization can be carried out in the next frame, and so on. Estimating the transition model $p(\cdot|\cdot)$ accurately is critical to achieve good performance, as shown numerically in Section 6.4. In fact, a poor estimate of $\hat{p}(\cdot|\cdot)$ may lead to inaccurate predictions of the SBPI, resulting in increased BT overhead and decreased spectral efficiency. To address this challenge, in Section 6.3, we will propose a state-of-art estimation of the mobility model based on R-VAE. **Policy:** a mapping from the current belief β_k to a BT or DT action. We break the policy into two steps. First, given β_k, the BS decides whether to perform BT or DT based on the policy $\mu_k(\beta_k)$, representing the probability of choosing DT in slot k. If DT is selected, then the BS further selects the DT action $\hat{s}_{\mathrm{DT}} \in \mathcal{S}$ according to a policy $\Delta_k(\beta_k)$, which is executed until the end of the frame; otherwise (BT selected, with probability $1-\mu_k(\beta_k)$), the BS selects a BT training set according to the policy $\Sigma_k(\beta_k) \in \mathcal{A}_{\mathrm{BT,k}}$. In this case, with $L = |\Sigma_k(\beta_k)|+1$ being the BT action duration, the next action is chosen in slot $k+L$, and so on until the end of the frame.

6.2.1 Optimization Problem

In this section, we present the optimization problem and a computationally efficient value iteration algorithm to solve the problem. We want to design a policy $\pi = (\mu, \Delta, \Sigma)$ which dictates whether to perform BT or DT and, if BT is chosen, it selects the BT set, with the goal to maximize the expected frame spectral efficiency. Let $k_{\mathrm{DT}} \in \{0, 1, \ldots, K\}$ be the slot when the DT action $\hat{s}_{\mathrm{DT}}$ is selected (if no DT action is selected within the frame, we let $k_{\mathrm{DT}} = K$). Then, the frame spectral efficiency is defined as

$$T_{\mathrm{fr}}(S_t) = \frac{1}{T_{\mathrm{frame}} W_{\mathrm{tot}}} r_{k_{\mathrm{DT}}}(S_t, \hat{s}_{\mathrm{DT}}) = \frac{\bar{R}}{W_{\mathrm{tot}}} \left(1 - \frac{k_{\mathrm{DT}}}{K}\right) \cdot 1[S_t = \hat{s}_{\mathrm{DT}}],$$

where $1 - \frac{k_{\mathrm{DT}}}{K}$ represents the loss of efficiency due to the BT overhead. The optimization problem is then stated as

$$\textbf{P1:} \max_{\pi} \ \mathbb{E}_{\pi} \left[T_{\mathrm{fr}}(S_t) | \beta_{t,0} \right], \tag{6.19}$$

where the expectation is conditional on the prior belief $\beta_{t,0}$ under the sequence of actions dictated by policy $\pi = (\mu, \Delta, \Sigma)$. This optimization can be carried out using the value iteration algorithm [36]. We define recursively the value function in slot k under belief $\beta_{t,k}$ and policy $\pi = (\mu, \Delta, \Sigma)$ as $V_K^{\pi}(\beta) = 0$ and, $\forall k = 0, \ldots, K-1$,

$$V_k^{\pi}(\beta) = \mu_k(\beta) \frac{\bar{R}}{W_{\mathrm{tot}}} \left(1 - \frac{k}{K}\right) \beta(\Delta_k(\beta)) \tag{6.20}$$

$$+ [1 - \mu_k(\beta)] \sum_{s \in \mathcal{S}} \beta(s) \sum_{y \in \mathcal{Y}} \mathbb{P}(y | s, \Sigma_k(\beta)) V_{k+|\Sigma_k(\beta)|+1}^{\pi}(\mathbb{B}(\beta, y, \Sigma_k(\beta))). \tag{6.21}$$

In fact, if a DT action is selected, with probability $\mu_k(\beta)$, then the reward is $\frac{\bar{R}}{W_{\mathrm{tot}}} \left(1 - \frac{k}{K}\right)$ until the end of the frame, as long as beam aligned is achieved, $S_t = \Delta_k(\beta)$ with probability $\beta(\Delta_k(\beta))$. On the other hand, if a BT action is selected, with probability $1 - \mu_k(\beta)$, then the BT set $\Sigma_k(\beta)$ is chosen and no reward is collected; since this action has duration $|\Sigma_k(\beta)|+1$, the future value function is taken at time $k + |\Sigma_k(\beta)|+1$, and the belief is updated based on the observation collected using the mapping $\mathbb{B}$. The optimal value function (V_k^*) is

obtained by maximizing with respect to the BT and DT actions, yielding $V_K^*(\beta) = 0$ and, $\forall k = 0, \ldots, K-1$,

$$V_k^*(\beta) = \max \left\{ \frac{\bar{R}}{W_{\text{tot}}} \left(1 - \frac{k}{K}\right) \max_{\hat{s}_{\text{DT}}} \beta(\hat{s}_{\text{DT}}), \max_{\hat{\mathcal{S}} \in \mathcal{A}_{\text{BT,k}}} \sum_{s \in \mathcal{S}} \beta(s) \sum_{y \in \mathcal{Y}} \mathbb{P}(y|s, \hat{\mathcal{S}}) V_{k+|\hat{\mathcal{S}}|+1}^*(\mathbb{B}(\beta, y, \hat{\mathcal{S}})) \right\}, \tag{6.22}$$

yielding the optimal decision between DT or BT action (maximizer of the outer max), the optimal DT action (maximizer of the first inner max) and the optimal BT set (maximizer of the second inner max). Since the value function is piecewise linear [36], it can be expressed using a finite set of $|\mathcal{S}|$-dimensional hyperplanes $\mathcal{Q}_k = \{\alpha_k^{(\ell)}\}_{\ell=1}^{M_k} \subset \mathbb{R}^{|\mathcal{S}|}$ of cardinality M_k:

$$V_k^*(\beta) = \max_{\alpha \in \mathcal{Q}_k} \langle \beta, \alpha \rangle \tag{6.23}$$

where $\langle \beta, \alpha \rangle = \sum_{s \in \mathcal{S}} \beta(s)\alpha(s)$. Since $V_K^*(\beta) = 0$, it follows that $\mathcal{Q}_K = \{\mathbf{0}\}$ with cardinality $M_K = 1$ and, for $k = 0, \ldots, K-1$, the set of hyperplanes is recursively computed as [36]

$$\mathcal{Q}_k \equiv \cup_{i=1}^{|\mathcal{S}|} \left\{ \frac{\bar{R}}{W_{\text{tot}}} \left(1 - \frac{k}{K}\right) \mathbf{e}_i \right\} \bigcup \cup_{\hat{\mathcal{S}} \in \mathcal{A}_{\text{BT},k}} \left\{ \sum_{y \in \mathcal{Y}} \mathbb{P}(y|\cdot, \hat{\mathcal{S}}) \odot \alpha^{(y)} : \ [\alpha^{(y)}]_{y \in \mathcal{Y}} \in \mathcal{Q}_{k+|\hat{\mathcal{S}}|+1}^{|\mathcal{Y}|} \right\}, \tag{6.24}$$

where $\mathbf{e}_i$ is the vector with entries equal to zero except in position i where $\mathbf{e}_i(s) = 1$, and $\mathbf{c} = \mathbf{a} \odot \mathbf{b}$ is the vector with entries $\mathbf{c}(i) = \mathbf{a}(i)\mathbf{b}(i), \ \forall i$. Note that in (6.24), the hyperplane of the form $\frac{\bar{R}}{W_{\text{tot}}} \left(1 - \frac{k}{K}\right) \mathbf{e}_i$ correspond to DT action $i \in \mathcal{A}_{\text{DT}}$. On the other hand, the hyperplanes of form $\sum_{y \in \mathcal{Y}} \mathbb{P}(y|\cdot, \hat{\mathcal{S}}) \odot \alpha^{(y)}$ correspond to BT action $\hat{\mathcal{S}} \in \mathcal{A}_{\text{BT},k}$, where $\alpha^{(y)}$ is the hyperplane corresponding to the future action at the slot $k + |\hat{\mathcal{S}}|+1$.

Note that in exact value iteration, the cardinality of $\mathcal{Q}_k$, M_k is shown to grow doubly exponentially with iteration k, i.e., $M_k = \mathcal{O}(|\mathcal{A}_{\text{DT}}|+(|\mathcal{A}_{\text{BT},k}|)^{|\mathcal{Y}|^{K-k-1}})$, thereby making it computationally intractable to use for a reasonable size task. To address this computational challenge, approximate value iteration techniques are proposed in the POMDP literature

[36]. In the next subsection, we present one such approach, called point-based value iteration (PBVI).

6.2.2 Point-Based Value Iteration

The key idea behind PBVI is to restrict the backup operation defined in (6.24) to a finite set of belief points $\tilde{\mathcal{B}}$, rather than the entire belief space, where $\tilde{\mathcal{B}}$ is chosen as representative of the entire belief-space $\mathcal{B}$.[1] In other words, given the $\tilde{V}_{k+1}, \ldots, \tilde{V}_K$, the PBVI algorithm builds an approximation $\tilde{V}_k = \tilde{H}[\tilde{V}_{k+1}, \ldots, \tilde{V}_K]$. To achieve this, the PBVI algorithm recursively constructs a set of hyperplanes $\mathcal{Q}_k$ and the associated actions $a_\alpha, \forall \alpha \in \mathcal{Q}_k$, for each $k \in \mathcal{K}$ through value iteration. Starting from $\mathcal{Q}_K = \{\mathbf{0}\}$, the set of hyperplanes is computed similar to (6.24), with one key difference that hyperplanes are pruned since the PBVI is restricted to the set $\tilde{\mathcal{B}}$ only. To this end, the hyperplanes which maximize the value function for $\beta \in \tilde{\mathcal{B}}$ are only included in hyperplane set $\mathcal{Q}_k, \forall k$. Given, this set of hyperplanes $\mathcal{Q}_k$, the value function $V_k^*(\beta), \forall \beta \in \mathcal{B}$ is approximated as

$$\tilde{V}_k(\beta) = \max_{\alpha \in \mathcal{Q}_k} \langle \beta, \alpha \rangle, \tag{6.25}$$

where $\langle \beta, \alpha \rangle \triangleq \sum_{s \in \mathcal{S}} = \beta(s)\alpha(s)$ is the inner prod between β and α. The approximately optimal policy is given as

$$\tilde{\pi}_k(\beta) = a_{\alpha_k^*}, \alpha_k^* = \arg\max_{\alpha \in \mathcal{Q}_k} \langle \beta, \alpha \rangle. \tag{6.26}$$

In other words the policy $\tilde{\pi}_k$ chooses that action $a_{\alpha_k^*}$ associated with the hyperplane α_k^*, which maximizes the value function in (6.25).

The back operator of the PBVI is by Algorithm 3, which takes as input $k \in \mathcal{K}$ and the set of belief points $\tilde{\mathcal{B}}$, and all previously computed sets of hyperplanes $\{\mathcal{Q}_{k+1}, \ldots, \mathcal{Q}_K\}$ and returns the set of hyperplane $\mathcal{Q}_k$ and the associated actions $\{a_\alpha^{(k)} : \alpha \in \mathcal{Q}_k\}$. To this end, for each belief $\beta \in \tilde{\mathcal{B}}$, a hyperplane is computed in lines 3–12. In particular, lines 4 compute hyperplane corresponding to optimal DT action Δ^* (computed in line 3). In

[1] $\uparrow \tilde{\mathcal{B}}$ is selected deterministically to uniformly cover the entire belief space.

lines 5–7, the BT hyperplane corresponding to the optimal BT action is computed. In particular, line 5 computes hyperplane for future value function for each action $\hat{\mathcal{S}} \in \mathcal{A}_{\mathrm{BT},k}$ and each possible observation $y \in \mathcal{Y}$; line 6 performs the backup operation to determine the hyperplane $\alpha_{\hat{\mathcal{S}}}$ for each BT action $\hat{\mathcal{S}} \in \mathcal{A}_{\mathrm{BT},k}$; line 7 finds the optimal BT hyperplane $\hat{\alpha}_{\mathrm{BT}}$ for the belief β and the associated optimal BT action Σ^*. Line 8 determines the probability of DT $\mu^* = 1[\langle \beta, \hat{\alpha}_{\mathrm{DT}} \rangle \geq \langle \beta, \hat{\alpha}_{\mathrm{BT}} \rangle]$. In lines 9–12, depending on μ^*, either the optimal DT hyperplane or the optimal BT hyperplane is added to the set of hyperplane $\mathcal{Q}_k$. If $\mu^* = 1$ (DT is optimal) $\hat{\alpha}_{\mathrm{DT}}$ is added to $\mathcal{Q}_k$; otherwise $\hat{\alpha}_{\mathrm{BT}}$ is added to $\mathcal{Q}_k$. Note that at most one hyperplane, which maximizes the value function for each $\beta \in \tilde{\mathcal{B}}$ is added to the set of hyperplane $\mathcal{Q}_k$. Therefore, unlike the exact value iteration, in the PBVI, the cardinality of a set of hyperplanes $|\mathcal{Q}_k| \leq |\tilde{\mathcal{B}}|$ does not grow beyond $|\tilde{\mathcal{B}}|$, yielding a linear-time value iteration algorithm.

The overall PBVI algorithm is given in Algorithm 4, which takes as input the discrete set of belief points $\tilde{\mathcal{B}}$, frame duration K [slots]. For each $k \in \mathcal{K}$, it applies the backup operator of the PBVI (Algorithm 3) to find the set of hyperplane $\mathcal{Q}_k$ and action associated with each hyperplane $\{a_\alpha^{(k)} : \alpha \in \mathcal{Q}_k\}$. The algorithm terminates in K steps and returns the sets of hyperplanes for each $k \in \mathcal{K}$ and associated actions.

6.2.3 Low-Complexity Policy Design

Although the PBVI algorithm finds an approximately optimal policy, it may incur a high computational complexity, especially for high dimensional belief spaces. In this section, we propose an MDP policy based on the assumption of error-free feedback to overcome this computational challenge. This case can be cast as a special case of the POMDP, where the BT observations have distribution

$$\mathbb{P}(y|s, \hat{\mathcal{S}}) = \begin{cases} 1[y = s], & \forall s \in \hat{\mathcal{S}} \\ 1[y = \emptyset], & \forall s \notin \hat{\mathcal{S}}. \end{cases}$$

Algorithm 3: backup method

input : $k, \tilde{\mathcal{B}}, \{\mathcal{Q}_{k+1}, \ldots, \mathcal{Q}_K\}$

init: $\mathcal{Q}_k = \{\}$

for *each* $\beta \in \tilde{\mathcal{B}}$ **do**

- $\Delta^* = \arg\max_{i \in \mathcal{S}} \langle \beta, \alpha_{\mathrm{DT},i} \rangle$, where $\alpha_{\mathrm{DT},i} \triangleq \frac{\bar{R}}{W_{\mathrm{tot}}} \left(1 - \frac{k}{K}\right) \mathbf{e}_i, \forall i \in \mathcal{S}$
- $\hat{\alpha}_{\mathrm{DT}} = \alpha_{\mathrm{DT},\Delta^*}$
- $\alpha'_{\hat{\mathcal{S}},y} = \arg\max_{\alpha \in \mathcal{Q}_{k+L_a}} \langle \mathbb{B}(\beta, y, \hat{\mathcal{S}}), \alpha \rangle, \forall \hat{\mathcal{S}} \in \mathcal{A}_{\mathrm{BT},k}, \forall y \in \mathcal{Y}$
- $\alpha_{\hat{\mathcal{S}}} = \sum_{y \in \mathcal{Y}} \mathbb{P}(y|\cdot, \hat{\mathcal{S}}) \odot \alpha'_{\hat{\mathcal{S}},y}, \forall \hat{\mathcal{S}} \in \mathcal{A}_{\mathrm{BT},k}$
- $\Sigma^* = \arg\max_{\hat{\mathcal{S}} \in \mathcal{A}_{\mathrm{BT},k}} \langle \beta, \alpha_{\hat{\mathcal{S}}} \rangle$, $\hat{\alpha}_{\mathrm{BT}} = \alpha_{\Sigma^*}$
- $\mu^* = 1[\langle \beta, \hat{\alpha}_{\mathrm{DT}} \rangle \geq \langle \beta, \hat{\alpha}_{\mathrm{BT}} \rangle]$
- **if** $\mu^* = 1$ **then**
  - $\mathcal{Q}_k \leftarrow \mathcal{Q}_k \cup \{\hat{\alpha}_{\mathrm{DT}}\}$
- **else**
  - $\mathcal{Q}_k \leftarrow \mathcal{Q}_k \cup \{\hat{\alpha}_{\mathrm{BT}}\}$
- $a^{(k)}_{\hat{\alpha}} = (\mu^*, \Delta^*, \Sigma^*)$

return $\mathcal{Q}_k, \{a^{(k)}_\alpha : \alpha \in \mathcal{Q}_k\}$

Algorithm 4: PBVI Algorithm

input : Belief set $\tilde{\mathcal{B}}$, frame duration K

init: $\mathcal{Q}_K = \{\mathbf{0}\}$

for $k = K-1, \ldots, 0$ **do**

- $\mathcal{Q}_k, \{a^{(k)}_\alpha : \alpha \in Q_k\} = \mathrm{backup}(k, \tilde{\mathcal{B}}, \{\mathcal{Q}_{k+1}, \ldots, Q_K\})$
- $\tilde{V}_k(\beta) = \max_{\alpha \in \mathcal{Q}_k} \beta \cdot \alpha, \forall \beta \in \tilde{\mathcal{B}}$

return $\{\mathcal{Q}_k : k \in \mathcal{K}\}, \{\{a^{(k)}_\alpha : \alpha \in \mathcal{Q}_k\} : k \in \mathcal{K}\}$

The belief update in (6.17) under the BT action $\hat{\mathcal{S}}$ of duration $L = |\hat{\mathcal{S}}|+1$ is then specialized as follows: under the observation $y = \emptyset$,

$$\beta_{k+L}(s) = \frac{\beta_{k+L-1}(s)}{\sum_{j \notin \hat{\mathcal{S}}} \beta_{k+L-1}(j)} 1[s \notin \hat{\mathcal{S}}] \tag{6.27}$$

and for $y \in \hat{\mathcal{S}}$ as

$$\beta_{k+L}(s) = 1[y = s]. \tag{6.28}$$

Since the short-timescale optimization is independent of frame index t, we drop the subscript t for notational simplification. Under such belief updates, it can be shown that the belief β_k is expressed as a function $\hat{\mathbb{B}}(\beta_0, \mathcal{U}_k)$ of the prior belief β_0 and the support $\mathcal{U}_k$ of β_k. This function is defined as follows

$$\beta_k(s) = \frac{\beta_0(s)}{\sum_{j\in\mathcal{U}_k}\beta_0(j)} 1[s \in \mathcal{U}_k], \forall s \in \mathcal{S}. \tag{6.29}$$

For any BT action of length $L \geq 2$ executed in slots $k, \ldots, k+L-1$, since the corresponding feedback signal is received at the end of slot $k+L-1$, no support updates occur in slots $k, \ldots, k+L-1$, hence $\mathcal{U}_\ell = \mathcal{U}_k,\ \forall \ell \in \{k, \ldots, k+L-1\}$. The support is updated at the start of slot $k+L$ following the rule $\mathbb{U}(\mathcal{U}_{k+L-1}, \hat{\mathcal{S}}, y)$, defined as

$$\mathcal{U}_{k+L} \triangleq \mathbb{U}(\mathcal{U}_{k+L-1}, \hat{\mathcal{S}}, y) = \begin{cases} \mathcal{U}_{k+L-1} \setminus \hat{\mathcal{S}}, & y = \emptyset \\ \{s\}, & y = s. \end{cases} \tag{6.30}$$

Given the prior belief β_0, the support $\mathcal{U}_k$ is sufficient statistics for the belief β_k. Therefore, we express the value function as a function of the support and the prior belief as follows. For a given a given prior belief β_0, let $\hat{V}_{\beta_0,k}(\mathcal{U}_k) \triangleq \frac{KW_{\text{tot}}}{R} V_k^*(\hat{\mathbb{B}}(\beta_0, \mathcal{U}_k))$ as the normalized value function expressed as a function of the support $\mathcal{U}_k$. Based on the support, the action-space can also be refined as follows. It can be shown any BT action $\hat{\mathcal{S}} \not\subset \mathcal{U}_k$ and any DT action $\hat{s}_{\text{DT}} \notin \mathcal{U}_k$ are suboptimal. Therefore, by excluding these suboptimal actions from the action-space, we redefine the BT and DT action-spaces as follows

$$\mathcal{A}_{\text{BT},k}(\mathcal{U}) \equiv \{\hat{\mathcal{S}} \subset \mathcal{U} : 1 \leq |\hat{\mathcal{S}}| \leq K-k-1\}, \qquad \mathcal{A}_{\text{DT}}(\mathcal{U}) \equiv \mathcal{U} \tag{6.31}$$

The value iteration algorithm in (6.22) can thus be specialized to

$$\hat{V}_{\beta_0,k}(\mathcal{U}) = \max\left\{(K-k)\frac{\max_{\hat{s}_{\text{DT}}\in\mathcal{U}}\beta_0(\hat{s}_{\text{DT}})}{\sum_{j\in\mathcal{U}}\beta_0(j)}, \max_{\hat{\mathcal{S}}\in\mathcal{A}_{\text{BT},k}(\mathcal{U})} \hat{V}^{(\text{BT})}_{\beta_0,k}(\mathcal{U}, \hat{\mathcal{S}})\right\}, \tag{6.32}$$

where the first term in (6.32) corresponds to reward if DT is performed and the second term in (6.32) corresponds to expected value function if BT is performed (maximized over all BT actions $\hat{\mathcal{S}} \in \mathcal{A}_{\mathrm{BT},k}(\mathcal{U})$). $\hat{V}^{(\mathrm{BT})}_{\beta_0,k}(\mathcal{U},\hat{\mathcal{S}})$ is the value function under BT action $\hat{\mathcal{S}} \in \mathcal{A}_{\mathrm{BT},k}(\mathcal{U})$, defined as

$$\begin{aligned}\hat{V}^{(\mathrm{BT})}_{\beta_0,k}(\mathcal{U},\hat{\mathcal{S}}) &\triangleq \mathbb{E}\left[\hat{V}_{\beta_0,k+|\hat{\mathcal{S}}|+1}(\mathcal{U})|\mathcal{U},\hat{\mathcal{S}}\right] \\ &= \frac{1}{\sum_{j\in\mathcal{U}}\beta_0(j)}\left[\sum_{s\in\hat{\mathcal{S}}}\beta_0(s)\hat{V}_{\beta_0,k+|\hat{\mathcal{S}}|+1}(\{s\}) + \sum_{s\in\mathcal{U}\setminus\hat{\mathcal{S}}}\beta_0(s)\hat{V}_{\beta_0,k+|\hat{\mathcal{S}}|+1}(\mathcal{U}\setminus\hat{\mathcal{S}})\right], \end{aligned} \tag{6.33}$$

where the first and second terms in (6.33) correspond to receiving $Y = s \in \hat{\mathcal{S}}$ and $Y = \emptyset$, respectively. Note that $\hat{V}_{\beta_0,K}(\mathcal{U}) = 0, \forall \mathcal{U} \subseteq \mathcal{S}$ trivially since the frame ends at $k = K$. Moreover, for singleton support, no BT action is optimal, yielding $\hat{V}_{\beta_0,k}(\mathcal{U}) = K - k, \forall |\mathcal{U}| = 1, \forall k \in \mathcal{K}$.

6.2.4 Structural Properties and Value Iteration

This section presents the structural properties for the MDP and provides an optimal value iteration to solve the MDP. In the following theorem, we prove that the optimal BT action should scan the most likely BPI according to the belief. This result may not hold for the POMDP since the uncertainty cannot be completely removed after scanning a beam.

Theorem 6.1. *Let β_0 be any prior belief with support $\mathcal{U}_0 \subseteq \mathcal{S}$. Then, in each state $\mathcal{U} \subseteq \mathcal{U}_0$, the optimal BT action contains m most likely BPIs from $\mathcal{U}$, given as*

$$\hat{\mathcal{S}}[m] = \arg\max_{\hat{\mathcal{S}}\subset\mathcal{U}:|\hat{\mathcal{S}}|=m} \sum_{s\in\hat{\mathcal{S}}}\beta_0(s), \tag{6.34}$$

where m^* is selected to maximize the BT value function, i.e.,

$$m^* = \arg\max_{m\in\mathcal{M}_k(|\mathcal{U}|)} \hat{V}^{(\mathrm{BT})}_{\beta_0,k}(\mathcal{U},\hat{\mathcal{S}}[m]), \tag{6.35}$$

and $\mathcal{M}_k(|\mathcal{U}|) \triangleq \{1,\ldots,\min\{|\mathcal{U}|-1, K-k-1\}\}$ is the set of feasible number of BPIs. In other words, the optimal BT action $\hat{\mathcal{S}}^* = \hat{\mathcal{S}}[m^*]$.

Proof. The proof is provided in Appendix 8.G. □

The above theorem implies that optimal BT is performed by scanning the $m \in \mathcal{M}_k(|\mathcal{U}_k|)$ most likely BPIs from the support $\mathcal{U}_k$ based on the prior belief shape β_0, where the optimal m is chosen to maximize the BT value function. Therefore, we can restrict the action-space for a given prior belief shape and support as

$$\begin{aligned}\mathcal{A}_{\mathrm{BT},k}(\beta_0,\mathcal{U}) &= \left\{\hat{\mathcal{S}}[m] : \hat{\mathcal{S}}[m] = \arg\max_{\hat{\mathcal{S}}\subset\mathcal{U}:|\hat{\mathcal{S}}|=m} \sum_{s\in\hat{\mathcal{S}}} \beta_0(s), \forall m \in \mathcal{M}_k(|\mathcal{U}|)\right\},\\ \mathcal{A}_{\mathrm{DT}}(\beta_0,\mathcal{U}) &= \left\{\hat{s}_{\mathrm{DT}} : \hat{s}_{\mathrm{DT}} = \arg\max_{s\in\mathcal{U}} \beta_0(s)\right\}\end{aligned} \tag{6.36}$$

Based upon the restricted action-space in (6.36), we can simplify the state-space. To this end, we show that the state for a given prior belief can be represented by a 2D vector in the following Corollary.

Corollary 6.1. *Let β_0 be prior belief with prior support $\mathcal{U}_0 = \{u_0, \ldots, u_0 + w_0 - 1\}$ of cardinality w_0, so that $\beta_0(u_0) \geq \beta_0(u_0+1) \geq \cdots \geq \beta_0(u_0+w_0-1) > 0$. Then, under any BT action $\mathrm{BT}, \hat{\mathcal{S}}[m] \in \mathcal{A}_{\beta_0,k}(\mathcal{U}_k)$, the support can be expressed as $\mathcal{U}_k = \{u_k, \ldots, u_k + w_k - 1\}$, where $u_k \in \mathcal{U}_0$ is the most likely BPI (ML-BPI) in $\mathcal{U}_k$ and $w_k \triangleq |\mathcal{U}_k|$. After executing $\hat{\mathcal{S}}[m] \in \mathcal{A}_{\mathrm{BT},k}(\beta_0,\mathcal{U})$ in slots $k, \ldots, k+L-1$, (u_k, w_k) is updated upon receiving the feedback Y_{k+L-1} as follows:*

$$(u_{k+L}, w_{k+L}) = \begin{cases} (u_{k+L-1} + m, u_0 + w_0 - u_{k+L-1} - m), & Y_{k+L-1} = \emptyset \\ (s^*, 1), & Y_{k+L-1} = s^* \in \hat{\mathcal{S}}[m], \end{cases} \tag{6.37}$$

where $(u_\ell, w_\ell) = (u_k, w_k), \forall \ell \in \{k, \ldots, k+L-1\}$.

Proof. The proof is provided in Appendix 8.H. □

The above corollary implies that (u_k, w_k) are the sufficient statistics for $\mathcal{U}_k$. Therefore, the state-space can be reduced to

$$\tilde{\mathcal{Z}} = \{(u, w) : u_0 \leq u \leq u_0 + w_0 - 1, w \in \{1, u_0 - u + w_0\}\}, \tag{6.38}$$

and the action space can be expressed as a function of the simplified state definition, as

$$\mathcal{A}_{\mathrm{BT},k}(u,w)=\{\hat{\mathcal{S}}[m]=\{u,\ldots,u+m-1\}:m\in\mathcal{M}_k(w)\},\mathcal{A}_{\mathrm{DT},k}(u,w)=\{u\}$$

Then, the value function in (6.32) can be expressed as a function of (u,w) instead of $\mathcal{U}$ as

$$\hat{V}_{\beta_0,k}(u,w)=\max\left\{(K-k)\frac{\beta_0(u)}{\sum_{j=u}^{u+w-1}\beta_0(j)},\max_{\hat{\mathcal{S}}[m]\in\mathcal{A}_{\mathrm{BT},k}(\mathcal{U})}\hat{V}^{(\mathrm{BT})}_{\beta_0,k}(u,w,\hat{\mathcal{S}}[m]),\right\} \tag{6.39}$$

where

$$\begin{aligned}&\hat{V}^{(\mathrm{BT})}_{\beta_0,k}(u,w,\hat{\mathcal{S}}[m])\\&=\frac{1}{\sum_{j=u}^{u+w-1}\beta_0(j)}\left[\sum_{s=u}^{u+m-1}\beta_0(s)\hat{V}_{\beta_0,k+m+1}(s,1)+\sum_{s=u+m}^{u_0+w_0-1}\beta_0(s)\hat{V}_{\beta_0,k+m+1}(u+m,u_0+w_0-u-m)\right].\end{aligned} \tag{6.40}$$

Finally, starting from $V_{\beta_0,K}(\mathcal{U})=0,\forall\mathcal{U}$, the value function can be computed for each k using backward induction of (6.39).

Both POMDP and MDP-based policies require the prior belief at the start of each frame to select the first action. Until now, we assumed that an estimate of the mobility model to compute the prior belief updates (see (6.18)) is available. In the following section, we will propose an R-VAE-based learning framework to obtain such an estimate.

6.3 Long Timescale: Mobility Learning via Recurrent Variational Autoencoders

In this section, we present the mobility learning module, aiming to learn a mobility model based on past sequences of observations and actions. Let $\mathbb{A}_{0:T}$ be a given encoded action sequence of length $T+1$, generated by following any arbitrary policy π, and $o_{0:T}$ be the corresponding sequence of encoded observations. Note that $\mathbb{A}_t$ and o_t encode all the BT actions and observations of frame t, respectively. Let $f(o_t|S_t,\mathbb{A}_t)$ be the known observation model; $\mathcal{P}_\psi(s_{0:T})\triangleq\beta_{0,0}(s_0)\prod_{t=1}^{T}p_\psi(s_t|s_{t-1})$ be the joint probability of state sequence $S_{0:T}$, where $\beta_{0,0}$ is the prior belief over S_0 at $t=0$ and $p_\psi(s'|s)$ is the unknown mobility model parameterized by ψ. Then, we aim to learn a mobility model $p_\psi(s'|s)$, which maximizes the

marginal likelihood of the observation and action sequences $f(o_{0:T}, \mathbb{A}_{0:T}|\beta_{0,0}, \psi)$, as stated in the following optimization problem:

$$\begin{aligned}\max_{\psi} f(o_{0:T}, \mathbb{A}_{0:T}|\beta_{0,0}, \psi) &\overset{(a)}{=} \max_{\psi} \mathbb{E}_{\mathcal{P}_\psi}\left[f(o_{0:T}, \mathbb{A}_{0:T}|S_{0:T}, \beta_{0,0})\right] \\ &\overset{(b)}{=} \max_{\psi} \sum_{s_{0:T}} \left[\mathcal{P}_\psi(s_{0:T}) \prod_{t=0}^{T} f(o_t|s_t, \mathbb{A}_t) f\left(\mathbb{A}_t \mid \beta_{0,0}, o_{0:t-1}, \mathbb{A}_{0:t-1}\right)\right],\end{aligned} \tag{6.41}$$

where the expectation in (a) is with respect to the unknown state sequence $S_{0:T} \sim \mathcal{P}_\psi(s_{0:T})$; (b) follows from using the following

$$\begin{aligned}f\left(o_{0:T}, \mathbb{A}_{0:T} \mid s_{0:T}, \beta_{0,0}\right) &\overset{(c)}{=} \prod_{t=0}^{T} f\left(o_t, \mathbb{A}_t \mid s_{0:T}, \beta_{0,0}, o_{0:t-1}, \mathbb{A}_{0:t-1}\right) \\ &\overset{(d)}{=} \prod_{t=0}^{T} f\left(o_t \mid s_t, \mathbb{A}_t\right) f\left(\mathbb{A}_t \mid s_{0:T}, \beta_{0,0}, o_{0:t-1}, \mathbb{A}_{0:t-1}\right) \\ &\overset{(e)}{=} \prod_{t=0}^{T} f\left(o_t \mid s_t, \mathbb{A}_t\right) f\left(\mathbb{A}_t \mid \beta_{0,0}, o_{0:t-1}, \mathbb{A}_{0:t-1}\right),\end{aligned} \tag{6.42}$$

where (c) follows from the law of conditional probability; (d) follows from the fact that given $(S_t, \mathbb{A}_t)$, o_t is independent of $(\beta_{0,0}, \mathbb{A}_{0:t-1}, o_{0:t-1}, s_{0:t-1})$; (e) follows from the fact that given $(\beta_{0,0}, o_{0:t-1}, \mathbb{A}_{0:t-1})$, $\mathbb{A}_t$ is independent of $s_{0:T}$ since $\mathbb{A}_t$ is obtained from policy π, which determines the actions based on history of actions and observations only. Notice that the term $f\left(\mathbb{A}_t \mid \beta_{0,0}, o_{0:t-1}, \mathbb{A}_{0:t-1}\right)$ is a functional of the policy used and is independent of the mobility model $p_\psi(s'|s)$, and can be determined in closed form. However, the marginal likelihood is intractable in general due to the lack of closed-form. In the latent variable learning literature, several approximate learning techniques [50], [74] are proposed to overcome this challenge, where a surrogate metric is used instead of the marginal likelihood. These techniques include the expectation-maximization (EM)-based algorithms such as the Baum-Welch algorithm [50], and variational techniques such as variational autoencoders [74]. The variational techniques jointly learn a separate posterior and prior state transition model, whereas the EM-based techniques perform an alternating optimization of a non-convex variational objective. Therefore, due to the joint optimization procedure, the variational techniques are

expected to perform better than the EM-based techniques and will be adopted in this chapter.

The variational autoencoder is one of the most powerful tools to learn latent variable models [74]. The VAE comprises two coupled but independently parameterized models: the encoder or inference model and the decoder or generative model. The encoder goal is to provide the posterior distribution over the latent state variable conditioned on the observations associated with the latent variable; the decoder measures the representation quality of the latent state variable produced by the encoder via the observation model and the prior distribution over the latent variable, thereby forcing the encoder to learn a meaningful representation of the latent variable from the observations. The R-VAE is an extension of the VAE for temporally correlated observations, such as the one obtained through the sampling of POMDP following a policy [49]. For this reason, we choose the R-VAE to learn the mobility model from noisy observations.

The goal of R-VAE is to learn the SBPI transition probability $p(s_t|s_{t-1}) \triangleq \mathbb{P}(S_t = s_t|S_{t-1} = s_{t-1}), \forall s_{t-1} \in \mathcal{S}, \forall s_t \in \bar{\mathcal{S}}$. However, since S_t is not observable if $S_t \neq \bar{s}$, it has to be inferred based on the history of actions and observations up to frame t. Let $\tilde{S}_t$ denotes the SBPI inferred from the action and observation history with its realization denoted by $\tilde{s}_t \in \bar{\mathcal{S}}$. In R-VAE setting, the latent variable $\tilde{S}_t$ is inferred by leveraging an encoder, which provides the posterior transition model $q_\nu(\tilde{s}_t|\tilde{s}_{t-1}, o_t, \mathbb{A}_t) \triangleq \mathbb{P}(\tilde{S}_t = \tilde{s}_t|\tilde{S}_{t-1}, O_t = o_t, \mathbb{A}_t)$, parameterized by ν. The decoder is composed of the prior transition model $p_\psi(\tilde{s}_t|\tilde{s}_{t-1})$, parameterized by ψ and a known observation model $f(o_t|\tilde{s}_t, \mathbb{A}_t)$.[2] The goal of the encoder is of the inferring $\tilde{S}_t \in \bar{\mathcal{S}}$ given $\tilde{S}_{t-1} = \tilde{s}_{t-1}$, encoded actions $\mathbb{A}_t$ of the current frame t and their corresponding encoded observations o_t; the decoder provides the prior transition $p_\psi(\tilde{s}_t|\tilde{s}_{t-1})$ and measures the likelihood of $\tilde{s}_t$ based on the a given observation o_t under a given observation model $f(o_t|\tilde{s}_t, \mathbb{A}_t)$. Let $\mathcal{P}_\psi(\tilde{s}_{0:T}) \triangleq \beta_{0,0}(\tilde{s}_0) \prod_{t=1}^{T} p(\tilde{s}_t|\tilde{s}_{t-1})$ denotes the joint probability of $\tilde{S}_{0:T}$ based on prior transition model p_ψ, where $\beta_{0,0}$ is prior belief over $\tilde{S}_0$. In R-VAE settings, the posterior dis-

[2]↑The observation models can also be learned under the R-VAE framework. However, enforcing a known accurate observation model reduces the dimensionality of the search space leading to better learning. Moreover, it leads to learning a state representation by R-VAE, which can be easily interpreted in light of the observation model [74]. In this chapter, an accurate observation model is obtained based on the distribution of the received signal based sectored antenna approximation.

tribution is approximated by $Q_\nu(\tilde{s}_{0:T}|o_{0:T}, \mathbb{A}_{0:T}, \beta_{0,0}) \triangleq \beta_{\text{post},0}(\tilde{s}_0) \prod_{t=1}^{T} q_\nu(\tilde{s}_t|\tilde{s}_{t-1}, o_t, \mathbb{A}_t)$ [49], where $\beta_{\text{post},0}$ is the posterior belief over $\tilde{S}_o$ after observing o_0, computed as

$$\beta_{\text{post},0}(\tilde{s}_0) = \frac{f(o_0|\tilde{s}_0, \mathbb{A}_0)\beta_{0,0}(\tilde{s}_0)}{\sum_{\tilde{s}\in\mathcal{S}} f(o_0|\tilde{s}, \mathbb{A}_0)\beta_{0,0}(\tilde{s})} \tag{6.43}$$

Then, the encoder and decoder are jointly designed to maximize the evidence lower bound (ELBO) for a given sequence of observation $o_{0:T}$ and $\mathbb{A}_{0:T}$ [74], defined as

$$\begin{aligned}\text{ELBO}(\nu, \psi, o_{0:T}, \mathbb{A}_{0:T}) &\triangleq \mathbb{E}_{Q_\nu}\left[\log \frac{\mathcal{P}_\psi(\tilde{S}_{0:T}) f(o_{0:T}, \mathbb{A}_{0:T}|\tilde{S}_{0:T}, \beta_{0,0})}{Q_\nu(\tilde{S}_{0:T}|o_{0:T}, \mathbb{A}_{0:T}, \beta_{0,0})}\right] \\ &\overset{(a)}{\leq} \log\left(\mathbb{E}_{\mathcal{P}_\psi}\left[f(o_{0:T}, \mathbb{A}_{0:T}|\tilde{S}_{0:T}, \beta_{0,0})\right]\right) \triangleq \log\left(f(o_{0:T}, \mathbb{A}_{0:T}|\beta_{0,0}, \psi)\right)\end{aligned} \tag{6.44}$$

where (a) follows by using the Jensen's inequality and the concavity of $\log(\cdot)$ function. Hence, the ELBO metric provides a tractable lower-bound to the log marginal likelihood $\log(f(o_{0:T}, \mathbb{A}_{0:T}|\beta_{0,0}, \psi))$ (see (6.41)), which is intractable in general. Therefore, an increase in the ELBO leads to a monotonic improvement in the marginal likelihood $f(o_{0:T}, \mathbb{A}_{0:T}|\beta_{0,0}, \psi)$. For gradient-based learning, the ELBO metric in (6.44) is expressed in more tractable form as follows:

$$\begin{aligned}\text{ELBO}(\nu, \psi, o_{0:T}, \mathbb{A}_{0:T}) = \mathbb{E}_{Q_\nu}&\left[\sum_{t=1}^{T} \log f(o_t|\tilde{S}_t, \mathbb{A}_t) - \log \frac{q_\nu(\tilde{S}_t|\tilde{S}_{t-1}, o_t, \mathbb{A}_t)}{p_\psi(\tilde{S}_t|\tilde{S}_{t-1})}\right] \\ &+ \sum_{t=0}^{T} \log f\left(\mathbb{A}_t \mid \beta_{0,0}, o_{0:t-1}, \mathbb{A}_{0:t-1}\right) + \log f(o_0|\beta_{0,0}, \mathbb{A}_0),\end{aligned} \tag{6.45}$$

where we have used (6.42). Notice that the last term in (6.45) is equal to $\log f(o_0|\beta_{0,0}, \mathbb{A}_0) = \log\left(\sum_{\tilde{s}\in\bar{\mathcal{S}}} f(o_0|\tilde{s}, \mathbb{A}_0)\beta_{0,0}(\tilde{s})\right)$. Since the last two terms in (6.45) are independent of the learnable parameters (ν, ψ), we neglect these two terms in the R-VAE's training and train the R-VAE based on the following modified ELBO

$$\widehat{\text{ELBO}}(\nu, \psi, o_{1:T}, \mathbb{A}_{1:T}) = \mathbb{E}_{Q_\nu}\left[\sum_{t=1}^{T} \log f(o_t|\tilde{S}_t, \mathbb{A}_t) - \log \frac{q_\nu(\tilde{S}_t|\tilde{S}_{t-1}, o_t, \mathbb{A}_t)}{p_\psi(\tilde{S}_t|\tilde{S}_{t-1})}\right]. \tag{6.46}$$

The overall design of the R-VAE is carried out by the maximization of the ELBO averaged over $N \geq 1$ episodes of form $(o_{0:T}, \mathbb{A}_{0:T}, T)$ as follows:

$$\max_{\nu,\psi} \overline{\text{ELBO}}(\nu, \psi) = \max_{\nu,\psi} \frac{1}{N} \sum_{n=1}^{N} \widehat{\text{ELBO}}(\nu, \psi, o^{(n)}_{1:T^{(n)}}, \mathbb{A}^{(n)}_{1:T^{(n)}}) \tag{6.47}$$

We now provide the concrete details of the R-VAE framework.

6.3.1 R-VAE framework

Actions encoding and observation model: We now describe the how the observations and actions are encoded. For each BPI $j \in \mathcal{S}$ and for each frame t, let $\{\Gamma^{(j)}_{t,1}, \ldots, \Gamma^{(j)}_{t,n^{(t)}_j}\}$ be the sequence of SNR measurements collected after execution all BT actions, where $n^{(t)}_j$ is total number of times BPI j is scanned during all BT actions in frame t. Then, the BT actions of the frame t are encoded as tuple $\mathbb{A}_t \triangleq (n^{(t)}_1, \ldots, n^{(t)}_{|\mathcal{S}|})$ containing the number of times each beam is scanned during all BT actions in the frame t, which is a sufficient statistics for the BT actions. For each frame t, the observation is denoted by $o_t \triangleq (o^{(t)}_1, \ldots, o^{(t)}_{|\mathcal{S}|})$, where $o^{(t)}_j$ is the average SNR measurement corresponding BPI j, defined as

$$o^{(t)}_j = \begin{cases} \frac{1}{n^{(t)}_j} \sum_{i=1}^{n^{(t)}_j} \Gamma^{(j)}_{t,i}, & n^{(t)}_j \geq 1 \\ 0, & n^{(t)}_j = 0. \end{cases} \tag{6.48}$$

If $\tilde{s}_t \neq \bar{s}$, the observation $o^{(t)}_j$ has the following distribution

$$f(o^{(t)}_j = o_j | \tilde{S}_t = s, n^{(t)}_j = n_j) = \begin{cases} n_j \text{Erlang}(n_j o_j | \lambda_{s,j}), & n_j \geq 1 \\ 1[o_j = 0], & n_j = 0 \end{cases} \tag{6.49}$$

where $\text{Erlang}(\cdot|\lambda)$ is the probability density function (pdf) of the Erlang distribution with rate parameter

$$\lambda_{s,j} = \frac{1}{1 + \text{SNR}\rho^{1[s \neq j]}}, \tag{6.50}$$

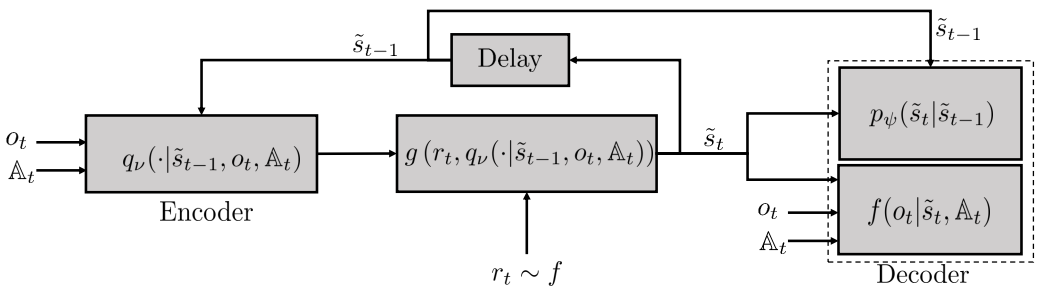

Figure 6.3. VAE training framework.

where SNR is the target SNR Since $\{o_j^{(t)}\}_{j\in\mathcal{S}}$ are conditionally independent given $\tilde{S}_t$ and $\mathbb{A}_t$, the overall distribution of o_t is given as

$$\begin{aligned} &f\left(o_t = (o_1, \cdots, o_{|\mathcal{S}|})|\tilde{S}_t = s, \mathbb{A}_t = (n_1^{(t)}, \ldots, n_{|\mathcal{S}|}^{(t)})\right) \\ &= \prod_{j\in\mathcal{S}:n_j\geq 1} f(o_j^{(t)} = o_j|\tilde{S}_t = s, n_j^{(t)} = n_j), \end{aligned} \tag{6.51}$$

On the other hand if the UE exits the coverage area of the BS, $o_t = \bar{s}$ and $\mathbb{A}_t = \mathbf{0}$ (since no BT is performed), we enforce $f(o_t = \bar{s}|\tilde{s}_t = \bar{s}, \mathbb{A}_t) = 1$.

Decoder: The decoder (generative model) models the observation distributions $f(o_t|\tilde{s}_t, \mathbb{A}_t)$ and the transition distribution $p_\psi(\tilde{s}_t|\tilde{s}_{t-1})$, parametrized by ψ. Since neural networks are universal function approximators and are well suited for gradient-based learning, we choose $p_\psi(\tilde{s}_t|\tilde{s}_{t-1})$ to be a feedforward neural network with learnable parameters ψ.

Encoder: Like previous works [49], we choose $q_\nu(\tilde{s}_t|\tilde{s}_{t-1}, o_t, \mathbb{A}_t)$ to be a recurrent neural network with weights and biases denoted by ν. The output of the neural network is produced by the softmax activation. The encoder models the posterior transition from $\tilde{s}_{t-1}$ to $\tilde{s}_t$ after observing o_t corresponding to $\mathbb{A}_t$. Note that, if $o_t = \bar{s}$ (the UE has exited the BS's coverage), we enforce $q_\nu(\tilde{s}_t = \bar{s}|\tilde{s}_{t-1}, o_t = \bar{s}, \mathbb{A}_t)$=1. On the other hand, if $o_t \neq \bar{s}$, we enforce $q_\nu(\tilde{s}_t = \bar{s}|\tilde{s}_{t-1}, o_t = \bar{s}, \mathbb{A}_t)$=0.

6.3.2 Optimization Algorithm

The VAE auto-encoder is trained using episodes, each of form $e_n \triangleq (o_{0:T^{(n)}}^{(n)}, \mathbb{A}_{0:T^{(n)}}^{(n)}, T^{(n)})$, where $T^{(n)} \geq 1$ denotes the total number of frames associated with the partial episode e_n.

The calculation of gradient of the ELBO objective in (6.46) with respect to (ν, ψ) is not tractable since the expectation therein is taken over the latent variables $\tilde{S}_{0:T} \sim Q_\nu$, whose joint distribution depends on ν. In the VAE literature, latent variable reparameterization techniques are proposed to overcome this problem, where a stochastic gradient estimate of the ELBO [74] is calculated. This is achieved by choosing a random variable $R_t \sim f_R$ and a differentialble function $g(\cdot, q_\nu(\cdot|\tilde{s}_{t-1}, o_t, \mathbb{A}_t))$ so that $g(R_t, q_\nu(\cdot|\tilde{s}_{t-1}, o_t, \mathbb{A}_t)) = \tilde{S}_t$. Then, expectation in (6.45) can be taken with respect to R_t instead of $\tilde{S}_t$, which enables tractable estimation of stochastic gradient estimates of the ELBO. We use Gumbel-softmax reparameterization technique (proposed in [75] for the non-recurrent VAE) to sample $\tilde{S}_t$. To the best of our knowledge, this is the first paper to adopt the Gumbel-softmax reparameterization with R-VAE. The Gumbel-softmax provides a simple yet efficient way to draw samples from any categorical distribution. We are interested in drawing samples from categorical distribution q_ν. To this end, for each $t \leq T$, we first draw $r_t \triangleq (r_{t,i})_{i \in \bar{\mathcal{S}}}$, where $r_{t,i} \sim \text{Gumbel}(0, 1), \forall i \in \bar{\mathcal{S}}$, where Gumbel$(0, 1)$ is standard Gumbel's distribution. Given $(\tilde{s}_{t-1}, o_t, \mathbb{A}_t)$ and r_t, the next $\tilde{s}_t$ is generated as

$$\tilde{s}_t = \arg\max_{i \in \bar{\mathcal{S}}} [r_{t,i} + \log q_\nu(\tilde{s}_t = i|\tilde{s}_{t-1}, o_t, \mathbb{A}_t)] \triangleq g(r_t, q_\nu(\cdot|\tilde{s}_{t-1}, o_t, \mathbb{A}_t)), \tag{6.52}$$

However, arg max in (6.52) is not differentiable, thereby making it unsuitable for gradient-based learning. Similar to [75], we adopt a hybrid strategy, where for forward-propagation, we generate $\tilde{s}_t$ based on (6.52) and for gradient calcuation via back-propagation, we use a differentiable approximation of (6.52) through the softmax function.

The overall training of R-VAE is shown in Fig. 6.3. The algorithm to train the VAE is given in Algorithm 5. It takes as input the encoder q_ν and transition model p_ψ and batch of episodes $\mathcal{E} = \left\{e_n = (o^{(n)}_{0:T^{(n)}}, \mathbb{A}^{(n)}_{0:T^{(n)}}, T_n) : n = 1, 2, \ldots, |\mathcal{E}|\right\}$; the algorithm returns the encoder and transition model trained on the batch $\mathcal{E}$. For each partial episode $e_n \in \mathcal{E}$, N_{trg} trajectories of $\tilde{s}_{0:T^{(n)}-1}$ are generated. Final gradient estimate is obtained by averaging over $|\mathcal{E}|$ episodes and N_{trg} trajectories of $\tilde{s}_{0:T^{(n)}-1}$ for each episodes. For each episode e_n and each trajectory ℓ, in lines 4, the $\tilde{s}_0$ is generated based on posterior belief over $\tilde{s}_0$. Then, $\tilde{s}_t$ is sequentially generated in lines 6–7 by drawing $r_{t,i} \sim \text{Gumbel}(0, 1) \forall i \in \bar{\mathcal{S}}$; in line 7 in line 6,

Algorithm 5: train-vae

input : q_ν, p_ψ, $\mathcal{E} = \left\{(o^{(n)}_{0:T^{(n)}}, \mathbb{A}^{(n)}_{0:T^{(n)}}, T_n) : n = 1, 2, \ldots, |\mathcal{E}|\right\}$

1 **for** *each episode* $e_n \in \mathcal{E}$ **do**

2 compute $\tilde{\beta}_{\text{post}}$ using $(o^{(n)}_0, \mathbb{A}^{(n)}_0)$

3 **for** $\ell = 1, \ldots, N_{\text{trg}}$ **do**

4 Sample $\tilde{s}^{(n)}_0 \sim \tilde{\beta}_{\text{post},0}$

5 **for** $t = 1, 2, \ldots, T$ **do**

6 Sample i.i.d $r_t = (r_{t,i})_{\in \bar{\mathcal{S}}} \sim \text{Gumbel}(0, 1)$

7 Generate one hot representation of $\tilde{s}^{(n)}_t$ as $\tilde{s}^{(n)}_t = g(r_t, q_\nu(\cdot|\tilde{s}^{(n)}_{t-1}, o^{(n)}_t, \mathbb{A}^{(n)}_t))$

8 $z_{n,\ell,t}(\nu, \psi) = \log f(o^{(n)}_t|\tilde{s}^{(n)}_t, \mathbb{A}^{(n)}_t) - \log \frac{q_\nu(\tilde{s}^{(n)}_t|\tilde{s}^{(n)}_{t-1}, o^{(n)}_t, \mathbb{A}^{(n)}_t)}{p_\psi(\tilde{s}^{(n)}_t|\tilde{s}^{(n)}_{t-1})}$

9 $\widehat{\text{ELBO}}_{n,\ell}(\nu, \psi) = \sum_{t=1}^{T^{(n)}} z_{n,\ell,t}(\nu, \psi)$

10 compute sample gradient $\nabla_{\nu,\psi}\widehat{\text{ELBO}}_{n,\ell}(\nu, \psi)$ by back-propagation through time

11 Compute batch gradient $\hat{\mathbf{g}}(\nu, \psi) = \frac{1}{|\mathcal{E}|N_{\text{trg}}} \sum_{n=1}^{|\mathcal{E}|} \sum_{\ell=1}^{N_{\text{trg}}} \nabla_{\nu,\psi}\widehat{\text{ELBO}}_{n,\ell}(\nu, \psi)$

12 $(\nu, \psi) \leftarrow (\nu, \psi) + \gamma\hat{\mathbf{g}}(\nu, \psi)$

13 **return** q_ν, p_ψ

followed by using (6.52) in 7. In line 8, we compute the partial ELBO for n^{th} partial episode, ℓ^{th} trajectory and t^{th} frame; in line 9, we compute the ELBO for n^{th} partial episode and ℓ^{th} trajectory; in line 10, the gradient is calculated by using back propagation through time (BPTT) for each e_n and ℓ. After traversing through each episode $e \in \mathcal{E}$ and N_{trg} trajectories for each partial episode, in line 11, the expected gradient estimate is obtained by averaging over all partial episodes $e \in \mathcal{E}$ and N_{trg} trajectories for each episode. Finally, the stochastic gradient ascent with step-size γ, is performed in line 12.

We can use the learned mobility model to perform the prior belief updates as follows. Let (ν^*, ψ^*) denote the parameters of the encoder and decoder, respectively, after the ELBO has converged. Then, the prior belief is computed as follows:

$$\beta_{t+1,0}(s') = \sum_{s \in \mathcal{S}} p_{\psi^*}(s'|s)\beta_{t,K}(s), \tag{6.53}$$

where $\beta_{t,K}(s)$ is the posterior belief at the end of the frame t.

Table 6.1. Simulation parameters.

Parameter	Symbol	Value
Number of BS antennas	M_{tx}	$128 = (16 \times 8)$
Number of UE antennas	$M_{\text{rx}}^{(I)}$	$32 = (8 \times 4)$
Number of BS beam	$\lvert\mathcal{C}\rvert$	32
Number of UE beams	$\lvert\mathcal{F}\rvert$	16
Slot duration	T_{slot}	400μs
Frame duration	K	50 [slots]
Distance of BS to Rd center	D	22m
Lane separation	Δ_{lane}	3.7m
BS height	h_{BS}	10m
Bandwidth	W_{tot}	100MHz
Carrier frequency	f	30GHz
Noise psd	N_0	-174dBm/Hz
Noise figure	F	10dB
Sidelobe/mainlobe SNR ratio	ρ	-10dB
Fraction of DT slot for channel estimation	κ	0.01
UE average speed	μ_v	30m/s
UE speed st. dev.	σ_v	10
UE mobility memory param.	γ	0.2
UE lane change prob.	$q_{1\to 2} = q_{2\to 1}$	0.01

6.4 Numerical Results

In this section, we present the numerical results illustrating the performance of the two proposed policies, namely PBVI policy and MDP-based policy. The simulation setup is described as follows.

6.4.1 Simulation Setup

We consider 2D mobility for a two-lane straight highway, similar to the one depicted in Fig. 6.1. The two-lane are separated by $3.7m$. The UE changes the lanes with probability $q_{l\to l} = 0.01$. The UE position along y-axis (along the direction of the road) evolves according to a Gauss-Markov mobility process. The speed $V_{y,t}$ and position $X_{y,t}$ of the UE along the road (y-axis) evolves as follows

$$\begin{cases} V_t = \gamma V_{t-1} + (1-\gamma)\mu_v + \sigma_v\sqrt{1-\gamma^2}\tilde{V}_{t-1}, \\ X_{y,t} = X_{y,t-1} + T_{\text{frame}} V_{t-1}, \end{cases} \tag{6.54}$$

where, unless otherwise stated, $\mu_v = 30$m/s is the average speed; $\sigma_v = 10$m/s is the standard deviation of speed; $\gamma = 0.2$ is the memory parameter; $\tilde{V}_{t-1} \sim \mathcal{N}(0,1)$, i.i.d. over slots.

The BS and UE use uniform planar arrays with analog 3D beamforming using codebooks $\mathcal{C}$ and $\mathcal{F}$, respectively with $|\mathcal{C}|= 32$ and $|\mathcal{F}|= 16$.

The encoder q_ϕ and decoder p_ψ are both neural networks. In particular, the encoder is a recurrent neural network with one fully-connected hidden layer having 100 units, each with the relu ($\max(0,x)$) activation function. A softmax layer produces the encoder's output with $|\bar{\mathcal{S}}|$ output units. Similarly, the decoder is a fully-connected feed-forward neural network with two hidden layers, each with 100 units and relu activation. Similar to the encoder, a softmax layer produces the output of the decoder with $|\bar{\mathcal{S}}|$ output units. We use the KL divergence between the ground truth mobility model $p^*(s'|s)$ (learned via error-free feedback) and the learned mobility model $\hat{p}(s'|s)$ to measure the accuracy of the learned model. In particular, we use KL divergence averaged over current SBPI s, defined as

$$\overline{\mathrm{KL}}(p^* \| \hat{p}) \triangleq \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \sum_{s' \in \bar{\mathcal{S}}} p^*(s'|s) \log \frac{p^*(s'|s)}{\hat{p}(s'|s)}. \tag{6.55}$$

We compare the two proposed policies (PBVI and MDP-based policies) with an exhaustive search restricted to SBPIs (EXOS) [16]. We evaluate the performance of the policies through the average spectral efficiency $\bar{T}^\pi_{\mathrm{tot}}$ [bp/s]. Using Little's law, the average spectral under a policy π is expressed as a function of the average number of bits successfully delivered to the UE $\bar{B}^\pi_{\mathrm{tot}}$ and the average duration of UE stay in the BS coverage $\bar{D}_{\mathrm{tot}}$[s] as

$$\bar{T}^\pi_{\mathrm{tot}} = \frac{\bar{B}^\pi_{\mathrm{tot}}}{\bar{D}_{\mathrm{tot}} W_{\mathrm{tot}}}. \tag{6.56}$$

The $\bar{D}_{\mathrm{tot}}$ is independent of policy and is a function of the ground truth mobility model p^*. $\bar{B}^\pi$ is expressed as

$$\bar{B}^\pi_{\mathrm{tot}} \triangleq \mathbb{E}_{p^*} \left[\sum_{t=0}^{\infty} T^\pi_{\mathrm{fr}}(S_t) 1[S_t \neq \bar{s}] \Big| \beta_{0,0} = \beta_{\mathrm{init}} \right] \cdot W_{\mathrm{tot}} T_{\mathrm{frame}}, \tag{6.57}$$

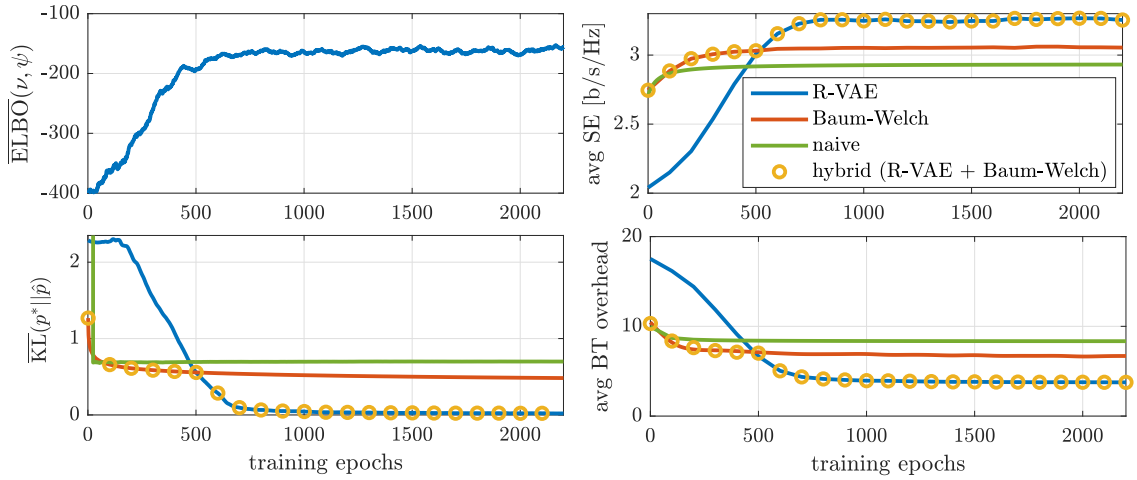

Figure 6.4. The training progress of R-VAE; SNR = 20dB, $\rho = -10$dB.

where expectation is with respect to the ground truth mobility model $p^*(s'|s)$ and $\beta_{\text{init}}(s) \triangleq 1[s = s_{\text{init}}]$ is a given prior belief at the start of frame t; s_{init} is initial state at $t = 0$. For the simulations, we use the system parameters values given in Table 6.1 unless stated otherwise.

6.4.2 Performance Evaluation

In Fig. 6.4, we show the training progress of R-VAE in terms of the average ELBO, accuracy of the learned mobility model in terms of average KL divergence $\overline{\text{KL}}(p^*\|\hat{p})$ between the ground truth mobility model $p^*(s'|s)$ and the learned $\hat{p}(s'|s)$. We measure the overall performance via the average spectral efficiency $\bar{T}^\pi$ and average BT overhead (percentage of frame duration used for BT). We compare the accuracy of the mobility model learned using the R-VAE with the accuracy of the mobility models learned using the naive approach (see (6.8)) and the Baum-Welch algorithm. The actions and observations are obtained following the PBVI-based policy (Algorithm 4). It can be seen that as training of R-VAE progresses, the ELBO increases and $\overline{\text{KL}}(p^*\|\hat{p})$ decreases simultaneously, indicating an improvement in the accuracy of the learned mobility model. Moreover, as the $\overline{\text{KL}}(p^*\|\hat{p})$ decreases, the average spectral efficiency achieved under R-VAE increases and the BT overhead reduces. The Baum-Welch and the naive approach converge faster than the R-VAE. However, despite a more gradual initial decline in $\overline{\text{KL}}(p^*\|\hat{p})$ of the R-VAE compared to that of the other two techniques, the R-VAE outperforms the other two techniques after the convergence. For

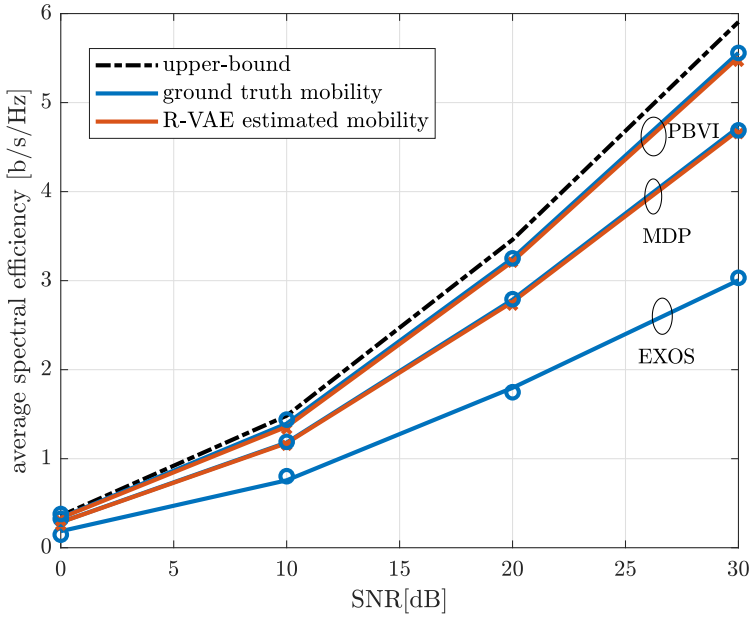

Figure 6.5. Average spectral efficiency versus SNR; $\rho = -10$dB. Solid lines correspond to the sectored antenna and Markovian mobility; markers correspond to the simulation with the 3D analog beam-forming and the 2D Gauss-Markov mobility.

instance, we can see that the R-VAE offers 92% and 86% reduction in $\overline{\mathrm{KL}}(p^*\|\hat{p})$ compared to the naive approach and the Baul-Welch algorithm, respectively. The improved accuracy of mobility learning by R-VAE translates to spectral efficiency gain and reduction of the BT overhead, as depicted in the figure. In particular, after the convergence, the R-VAE outperforms the Baum-Welch algorithm and the naive approach with a spectral efficiency gain of 8% and 12.6%, respectively. Similarly, after convergence, the R-VAE yields 43% and 60% reduced BT overhead compared to the Baum-Welch algorithm and the naive approach, respectively. However, before the ELBO converges, the Baum-Welch performs the best, followed by the naive approach. So, to achieve the best performance, we propose a hybrid approach, where before the convergence of the ELBO for the R-VAE, Baum-Welch is used, and the R-VAE is used after the ELBO has converged. Such a scheme is possible due to the decoupling of policy design and mobility learning via the dual timescale approach. Because of such a design, we can train multiple mobility models simultaneously and use prior belief based on any one of the mobility models. The performance of the hybrid scheme is shown

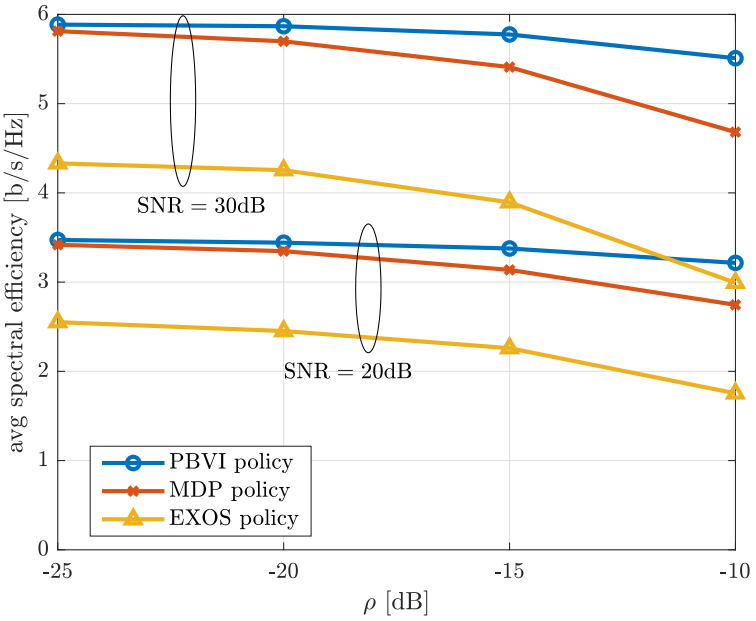

Figure 6.6. Average spectral efficiency versus ρ.

in the figure with the markers, where we depict its fastest convergence as well as its best performance among all schemes.

In Fig. 6.5, the spectral efficiency of the proposed policies is depicted. We also show a spectral efficiency upper-bound, attained by a gene-aided policy using the error-free BT feedback and the ground-truth mobility model. Note that this upper-bound is not attainable in practice because the feedback is erroneous due to noise and beam imperfections. The solid lines represent the performance using the analytical model based on the sector antenna gain and Markovian approximation of the non-Markovian Gauss-Markov mobility of UE. The markers represent the performance evaluated using Monte-Carlo simulation with the analog beamforming, and the non-Markovian Gauss-Markov mobility of the UE, where each simulated point is obtained by the sample mean of 10^5 episodes. It can be seen that both the analytical and simulated spectral efficiency values match, thereby verifying the accuracy of the sectored antenna-based gain approximation and the Markovian approximation of the Gauss-Markov mobility. For both PBVI and MDP-based policies, the R-VAE provides performance very close to the ground-truth mobility model. It can be observed that the PBVI policy coupled with R-VAE yields the best performance very close to gene-aided

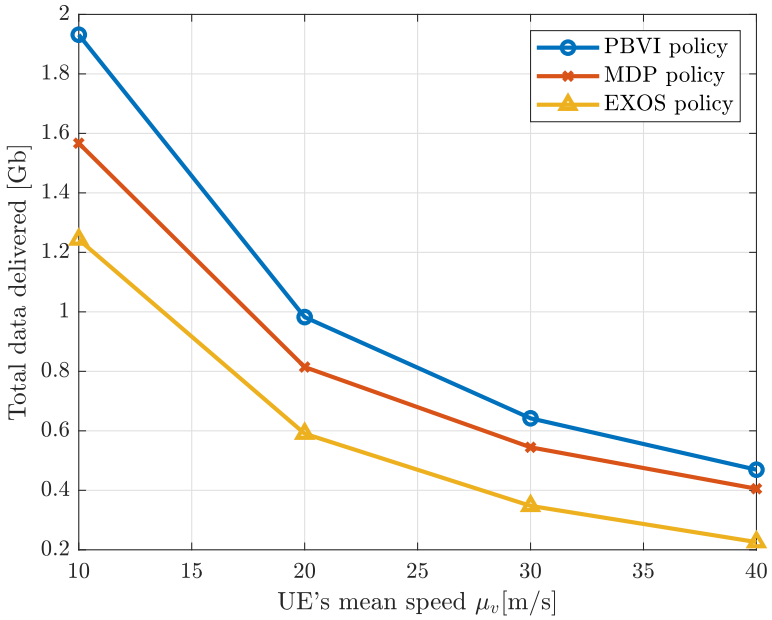

Figure 6.7. Throughput vs the mean speed μ_v; SNR = 20dB, $\rho = -10$dB.

upper-bound. The PBVI policy outperforms the MDP-based policy (coupled with the R-VAE) and EXOS policy, with spectral efficiency gains of up to 16% and 46%, respectively. This is attributed to the enhanced robustness of the PBVI policy, incorporated via the BT feedback distribution. On the other hand, the MDP-based policy shows a spectral efficiency gain of 37% over the EXOS policy. This performance gain is because the MDP-based policy can reduce the BT overhead by adaptively scanning a few most likely beam pairs based on each frame's prior belief and feedback.

In Fig. 6.6, we depict the behavior of spectral efficiency as the aligned SNR-to-misaligned SNR ratio ρ is varied. It can be noticed that the performance of the three policies degrades as ρ increases since the errors in feedback become more frequent at higher ρ. Notably, the PBVI policy's performance degrades the least by an increase in ρ, whereas the EXOS's performance degrades the most by an increase in ρ. Moreover, at lower values of ρ, the MDP-based policy performs very close to the PBVI policy. Furthermore, the total optimization and execution time of MDP-based policy is 4.7 times smaller than the PBVI policy. Therefore, the MDP-based policy offers a low-complexity alternative for the PBVI, having negligible performance degradation when ρ is small.

In Fig. 6.7, we depict the average total data delivered to the UE successfully as a function of mean speed μ_v. The total data delivered follows a monotonically decreasing trend with the mean speed μ_v. This trend is attributed to the shorter average episode duration at the higher speed and the exacerbated overhead of beam-training since SBPI prediction become less and less accurate at higher speeds. As observed before, the PBVI policy outperforms the MDP-based policy, and MDP policy outperforms the EXOS policy.

7. CONCLUSION

Millimeter wave communications use large antenna arrays with narrow beams to overcome the huge frequency-dependent path loss. However, the use of large antenna arrays with beamforming demands precise beam-alignment between transmitter and receiver, and may entail huge overhead, especially in highly mobile environments. This thesis addresses the challenges in the design of beam-alignment and data transmission by proposing various schemes that exploit different timescales.

In Chapter 2, we have designed an optimal interactive beam-alignment scheme, with the goal of minimizing power consumption under a rate constraint. For the case of perfect detection and uniform prior on AoD/AoA, we proved that the optimal beam-alignment protocol has fixed beam-alignment duration, and that a *decoupled fractional search* method is optimal. Inspired by this scheme, we have proposed a heuristic policy for the case of a non-uniform prior, and showed that the uniform prior is the worst-case scenario. Furthermore, we have investigated the impact of beam-alignment errors on the average throughput and power consumption. The numerical results depicted the superior performance of our proposed scheme, with up to 4dB, 7.5dB, and 14dB gain compared to a state-of-the-art bisection search, conventional exhaustive search and interactive exhaustive search policies, respectively, and robustness against multi-cluster channels.

In Chapter 3, we have designed a coded energy-efficient beam-alignment. The scheme minimizes power consumption and uses an error correction code to recover from detection errors introduced during beam-alignment. We compare our proposed scheme with energy-efficient uncoded beam-alignment and exhaustive search, demonstrating its superior performance.

In Chapter 4, we have formulated the beam-alignment problem as a Bayesian multiarm bandits problem. For the optimal control design, we have identified a simplified sufficient statistic referred to as the preference of beam pairs. Based on the preference and bounding of the value function, we have proposed a heuristic policy, which selects the beam pair with the second best-preference to scan. We have shown numerically that the proposed scheme

outperforms the first-best, linear Thompson samping, and upper confidence bound based beam-alignment schemes proposed in the literature.

In Chapter 5, we have investigated the design of beam-training/data-transmission/handover strategies for mm-wave vehicular networks. The mobility and blockage dynamics have been leveraged to obtain the approximately optimal policy via a partially observable Markov decision process (POMDP) formulation and its solution via a point-based value iteration (PBVI) algorithm based on PERSEUS [8]. Moreover, we have proposed two heuristic policies, namely adaptive heuristic (A-HEU) and belief-based heuristic (B-HEU), which provide low computational alternatives to PBVI and exhibit performance comparable to the optimal policy obtained via PBVI. We have also compared the performance of the proposed policies to a baseline algorithm with periodic beam training. Our numerical results demonstrate the importance of an adaptive design to tackle the highly dynamic environments caused by mobility and blockages in vehicular networks. This is demonstrated by the superiority of the PERSEUS-based and heuristic schemes compared to a baseline scheme with periodic beam training (up to 2$\times$ improvement in spectral efficiency). Additionally, our results depict a complexity-performance tradeoff: while the PERSEUS-based policy achieves near-optimal performance, the heuristic schemes A-HEU and B-HEU achieve a small performance degradation ($\sim 10\%$), at a fraction of the computational cost of PERSEUS-based.

In Chapter 6, we have proposed a dual timescale approach, which exploits mobility dynamics to mitigate the beam-training overhead. We have developed a POMDP framework for the short-timescale to design an approximately optimal policy. We have also designed a recurrent variational autoencoder-based mobility learning framework, which uses noisy observations collected under the policy to learn the Markovian dynamics of the UE mobility. Our performance evaluation have demonstrated that the proposed policy, coupled with the mobility learning framework, yields approximately optimal performance, showing spectral efficiency gain of up to 46%, compared to an exhaustive search variant. We have also demonstrated the superior learning performance of the mobility learning framework, yielding spectral efficiency gains of 8% and 12.6%, compared to the Baum-Welch algorithm and a naive mobility learning approach.

8. APPENDICES

8.A Proof of Lemma 2.1

Proof. We need the following lemma.

Lemma 8.1. *Given $f_k, \mathbf{a}_k, C_k$, the belief f_{k+1} is computed as*

$$f_{k+1}(\boldsymbol{\theta}) = \begin{cases} \frac{\chi_{\mathcal{B}_k}(\boldsymbol{\theta})}{\int_{\mathcal{B}_k} f_k(\tilde{\boldsymbol{\theta}})\mathrm{d}\tilde{\boldsymbol{\theta}}} f_k(\boldsymbol{\theta}), & k \in \mathcal{I}_s, C_k = \text{ACK}, \\ \frac{1-\chi_{\mathcal{B}_k}(\boldsymbol{\theta})}{1-\int_{\mathcal{B}_k} f_k(\tilde{\boldsymbol{\theta}})\mathrm{d}\tilde{\boldsymbol{\theta}}} f_k(\boldsymbol{\theta}), & k \in \mathcal{I}_s, C_k = \text{NACK}, \\ f_k(\boldsymbol{\theta}), & k \in \mathcal{I}_d, C_k = \text{NULL}. \end{cases} \tag{8.1}$$

Proof. We denote AoD/AoA random variables pair by $\boldsymbol{\Theta} \triangleq (\Theta_{\mathrm{t}}, \Theta_{\mathrm{r}})$ and its realization by $\boldsymbol{\theta} \triangleq (\theta_{\mathrm{t}}, \theta_{\mathrm{r}})$. First note that for $0 \leq k \leq N-1$, we have

$$\begin{aligned} f_{k+1}(\boldsymbol{\theta}) &= f(\boldsymbol{\Theta} = \boldsymbol{\theta} | \mathbf{a}^k, C^{k-1}, C_k = c_k) \\ &\overset{(a)}{=} \frac{\mathbb{P}(C_k = c_k | A^k, C^{k-1}, \boldsymbol{\Theta} = \boldsymbol{\theta}) f(\boldsymbol{\Theta} = \boldsymbol{\theta} | \mathbf{a}^k, C^{k-1})}{\int_{-\pi}^{\pi} \mathbb{P}(C_k = c_k | A^k, C^{k-1}, \boldsymbol{\Theta} = \tilde{\boldsymbol{\theta}}) f(\boldsymbol{\Theta} = \tilde{\boldsymbol{\theta}} | \mathbf{a}^k, C^{k-1}) d\tilde{\boldsymbol{\theta}}} \\ &\overset{(b)}{=} \frac{\mathbb{P}(C_k = c_k | \mathbf{a}_k, \boldsymbol{\Theta} = \boldsymbol{\theta}) f_k(\boldsymbol{\theta})}{\int_{-\pi}^{\pi} \mathbb{P}(C_k = c_k | \mathbf{a}_k, \boldsymbol{\Theta} = \tilde{\boldsymbol{\theta}}) f_k(\tilde{\boldsymbol{\theta}}) d\tilde{\boldsymbol{\theta}}} \end{aligned} \tag{8.2}$$

where we have used Bayes' rule in step (a); (b) is obtained by using the fact that, given $\boldsymbol{\Theta} = \boldsymbol{\theta}$, C_k is a deterministic function of $(\mathbf{a}_k, \boldsymbol{\theta})$, independent of $\mathbf{a}^{k-1}, C^{k-1}$; additionally, we used the fact that $f_k(\boldsymbol{\theta}) = f(\boldsymbol{\Theta} = \boldsymbol{\theta} | \mathbf{a}^k, C^{k-1})$ since $\boldsymbol{\Theta}$ is independent of $\mathbf{a}_k$ given $(\mathbf{a}_{k-1}, C^{k-1})$. Now consider the case $k \in \mathcal{I}_s$, i.e., $\xi_k = 1$ and $C_k = \text{ACK}$. Then, we can use (8.2) to get

$$\begin{aligned} f_{k+1}(\boldsymbol{\theta}) &= \frac{\mathbb{P}(C_k = \text{ACK} | \mathcal{B}_{\mathrm{t},k}, \mathcal{B}_{\mathrm{r},k}, \xi_k = 1, \boldsymbol{\Theta} = \boldsymbol{\theta}) f_k(\boldsymbol{\theta})}{\int_{-\pi}^{\pi} \mathbb{P}(C_k = \text{ACK} | \mathcal{B}_{\mathrm{t},k}, \mathcal{B}_{\mathrm{r},k}, \xi_k = 1, \boldsymbol{\Theta} = \tilde{\boldsymbol{\theta}}) f_k(\tilde{\boldsymbol{\theta}}) d\tilde{\boldsymbol{\theta}}} \\ &= \frac{\chi_{\mathcal{B}_k}(\boldsymbol{\theta})}{\int_{\mathcal{B}_k} f_k(\tilde{\boldsymbol{\theta}}) \mathrm{d}\tilde{\boldsymbol{\theta}}} f_k(\boldsymbol{\theta}), \end{aligned} \tag{8.3}$$

where $\mathcal{B}_k \triangleq \mathcal{B}_{\mathrm{t},k} \times \mathcal{B}_{\mathrm{r},k}$. Similarly, for $k \in \mathcal{I}_s$ and $C_k = \text{NACK}$, (8.2) can be used to get

$$f_{k+1}(\boldsymbol{\theta}) = \frac{1 - \chi_{\mathcal{B}_k}(\boldsymbol{\theta})}{1 - \int_{\mathcal{B}_k} f_k(\tilde{\boldsymbol{\theta}}) \mathrm{d}\tilde{\boldsymbol{\theta}}} f_k(\boldsymbol{\theta}). \tag{8.4}$$

For $k \in \mathcal{I}_d$, $\mathbb{P}(C_k = \text{NULL}|\mathcal{B}_{\text{t},k}, \mathcal{B}_{\text{r},k}, \xi_k = 0, \boldsymbol{\Theta} = \boldsymbol{\theta}) = 1$. Therefore, we use (8.2) to get

$$f_{k+1}(\boldsymbol{\theta}) = f_k(\boldsymbol{\theta}). \tag{8.5}$$

Thus we have proved the Lemma. □

We prove the lemma by induction. The hypothesis holds trivially for $k = 0$. Let us assume that it holds in slot $k \geq 0$, we show that it holds in slot $k+1$ as well. First, let us consider the case when $k \in \mathcal{I}_\text{s}$ and $C_k = \text{ACK}$. By using (8.1) along with the induction hypothesis, we get

$$f_{k+1}(\boldsymbol{\theta}) = \frac{f_0(\boldsymbol{\theta})}{\int_{-\pi}^{\pi} \chi_{\mathcal{U}_k \cap \mathcal{B}_k}(\tilde{\boldsymbol{\theta}}) f_0(\tilde{\boldsymbol{\theta}}) d\tilde{\boldsymbol{\theta}}} \chi_{\mathcal{U}_k \cap \mathcal{B}_k}(\boldsymbol{\theta}). \tag{8.6}$$

By substituting $\mathcal{U}_{k+1} \equiv \mathcal{U}_k \cap \mathcal{B}_k$, we get (2.31).

Next, we focus on the case when $k \in \mathcal{I}_\text{s}$ and $C_k = \text{NACK}$. In this case, (8.1) yields

$$f_{k+1}(\theta_\text{t}, \theta_\text{r}) = \frac{f_0(\boldsymbol{\theta})}{\int_{-\pi}^{\pi} \chi_{\mathcal{U}_k \setminus \mathcal{B}_k}(\tilde{\boldsymbol{\theta}}) f_0(\tilde{\boldsymbol{\theta}}) d\tilde{\boldsymbol{\theta}}} \chi_{\mathcal{U}_k \setminus \mathcal{B}_k}(\boldsymbol{\theta}), \tag{8.7}$$

where we used the fact that $\chi_{[-\pi,\pi]^2 \setminus \mathcal{A}}(x) \equiv 1 - \chi_{\mathcal{A}}(x)$. By observing that $\mathcal{U}_{k+1} \equiv \mathcal{U}_k \setminus \mathcal{B}_k$, we get the expression for $f_{k+1}(\boldsymbol{\theta})$, as given in (2.31).

Finally, for $k \in \mathcal{I}_d$, (8.1) yields $f_{k+1}(\boldsymbol{\theta}) = f_k(\boldsymbol{\theta})$. Therefore, from the induction hypothesis it follows that $f_{k+1}(\boldsymbol{\theta})$ is given by (2.31) with $\mathcal{U}_{k+1} = \mathcal{U}_k$. Hence, the lemma is proved. □

8.B Supplementary Lemma 8.2

Lemma 8.2. *The optimal 2D beam satisfies*

$$\begin{cases} \mathcal{B}_k \subset \mathcal{U}_k, \forall k \in \mathcal{I}_\text{s} \\ \mathcal{B}_k \subseteq \mathcal{U}_k, \forall k \in \mathcal{I}_\text{d}. \end{cases} \tag{8.8}$$

Proof. We prove this lemma by contradiction. First, we consider the beam-alignment action $\mathbf{a}_k = (1, \mathcal{B}_k, 0)$ such that $\mathcal{B}_k \setminus \mathcal{U}_k \neq \emptyset$, i.e., $\mathcal{B}_k$ has non-empty support outside of $\mathcal{U}_k$. Let

$\tilde{\mathbf{a}}_k = (1, \tilde{\mathcal{B}}_k, 0)$ be new beam-alignment action such that $\tilde{\mathcal{B}}_k = \mathcal{U}_k \cap \mathcal{B}_k$, i.e., $\tilde{\mathcal{B}}_k$ is constructed by restricting $\mathcal{B}_k$ within the belief support $\mathcal{U}_k$. Using (2.37) , we get

$$\begin{aligned}\hat{V}_k(\mathbf{a}_k; \mathcal{U}_k, D_k) =& \phi_s \left|\mathcal{B}_k\right| + \mathbb{P}(C_k = \text{ACK}|\mathcal{U}_k, \mathcal{B}_k)\hat{V}^*_{k+1}(\mathcal{U}_k \cap \mathcal{B}_k, D_k) \\ &+ \mathbb{P}(C_k = \text{NACK}|\mathcal{U}_k, \mathcal{B}_k)\hat{V}^*_{k+1}(\mathcal{U}_k \setminus \mathcal{B}_k, D_k). \end{aligned} \tag{8.9}$$

Using the fact that $\tilde{\mathcal{B}}_k = \mathcal{U}_k \cap \mathcal{B}_k$, hence $\mathcal{U}_k \setminus \mathcal{B}_k = \mathcal{U}_k \setminus \tilde{\mathcal{B}}_k$, it follows that $\mathbb{P}(C_k = c|\mathcal{U}_k, \mathcal{B}_k) = \mathbb{P}(C_k = c|\mathcal{U}_k, \tilde{\mathcal{B}}_k),\ \forall c \in \{\text{ACK}, \text{NACK}\}$. Therefore, we rewrite (8.9) as

$$\begin{aligned}&\hat{V}_k(\mathbf{a}_k; \mathcal{U}_k, D_k) = \phi_s \left|\tilde{\mathcal{B}}_k\right| + \phi_s \left|\mathcal{U}_k \setminus \mathcal{B}_k\right| \\ &+ \mathbb{P}(C_k = \text{ACK}|\mathcal{U}_k, \tilde{\mathcal{B}}_k)\hat{V}^*_{k+1}(\tilde{\mathcal{B}}_k, D_k) + \mathbb{P}(C_k = \text{NACK}|\mathcal{U}_k, \tilde{\mathcal{B}}_k)\hat{V}^*_{k+1}(\mathcal{U}_k \setminus \tilde{\mathcal{B}}_k, D_k) \\ &> \hat{V}_k(\tilde{\mathbf{a}}_k; \mathcal{U}_k, D_k), \end{aligned} \tag{8.10}$$

where we have used $|\mathcal{U}_k \setminus \mathcal{B}_k| > 0$. Thus $\mathbf{a}_k$ is suboptimal, implying that optimal beam-alignment beam satisfy $\mathcal{B}_k \subseteq \mathcal{U}_k$. Now, let $\mathcal{B}_k = \mathcal{U}_k$, and consider a new action with beam $\tilde{\mathcal{B}}_k = \emptyset$. Using a similar approach, it can be shown that $\mathcal{B}_k = \mathcal{U}_k$ is suboptimal with respect to $\tilde{\mathcal{B}}_k$, hence we must have $\mathcal{B}_k \subset \mathcal{U}_k$.

To prove the lemma for $k \in \mathcal{I}_d$, consider the action $\mathbf{a}_k = (0, \mathcal{B}_k, R_k)$ such that $\mathcal{B}_k \setminus \mathcal{U}_k \neq \emptyset$. Now consider a new action $\tilde{\mathbf{a}}_k = (0, \tilde{\mathcal{B}}_k, R_k)$ such that $\tilde{\mathcal{B}}_k = \mathcal{B}_k \cap \mathcal{U}_k$. It can be observed that $\mathbb{P}(\boldsymbol{\theta} \in \mathcal{B}_k|\mathcal{U}_k) = \mathbb{P}(\boldsymbol{\theta} \in \tilde{\mathcal{B}}_k|\mathcal{U}_k)$. The cost-to-function for the action $\mathbf{a}_k$ is given as

$$\begin{aligned}\hat{V}_k(\mathbf{a}_k; \mathcal{U}_k, D_k) &= \frac{\psi_d(R_k)}{\bar{F}^{-1}_\gamma\left(\frac{1-\epsilon}{\mathbb{P}(\boldsymbol{\theta} \in \mathcal{B}_k|\mathcal{U}_k)}|\hat{\gamma}\right)} \left|\mathcal{B}_k\right| + \hat{V}^*_{k+1}(\mathcal{U}_k, D_k - TR_k) \\ &> \frac{\psi_d(R_k)}{\bar{F}^{-1}_\gamma\left(\frac{1-\epsilon}{\mathbb{P}(\boldsymbol{\theta} \in \tilde{\mathcal{B}}_k|\mathcal{U}_k)}|\hat{\gamma}\right)} \left|\tilde{\mathcal{B}}_k\right| + \hat{V}^*_{k+1}(\mathcal{U}_k, D_k - TR_k) \\ &= \hat{V}_k(\tilde{\mathbf{a}}_k; \mathcal{U}_k, D_k), \end{aligned}$$

hence we must have $\mathcal{B}_k \subseteq \mathcal{U}_k$. The lemma is thus proved. □

8.C Proof of Theorem 2.1

Proof. For a data communication action $\mathbf{a}_k \in \mathcal{A}_{\text{ext}}(\mathcal{U}, D)$, the state transition is independent of $\mathcal{B}_k$ since $\mathcal{U}_{k+1}=\mathcal{U}_k$ and $D_{k+1}=D_k-R_kT$. Hence, the optimal beam given R_k is obtained by minimizing $c(\mathbf{a}_k;\mathcal{U}_k, D_k)$ in (2.36), yielding

$$\begin{aligned} c(\mathbf{a}_k;\mathcal{U}_k, D_k) &\overset{(a)}{=} \psi_d(R_k)\frac{|\mathcal{B}_k|}{\bar{F}_\gamma^{-1}\left(\frac{(1-\epsilon)|\mathcal{U}_k|}{|\mathcal{B}_k|}|\hat{\gamma}\right)} \\ &\overset{(b)}{\geq} \psi_d(R_k)(1-\epsilon)\,|\mathcal{U}_k|\,\frac{1}{q^*\bar{F}_\gamma^{-1}(q^*|\hat{\gamma})}, \end{aligned} \tag{8.11}$$

where (a) follows from $\mathbb{P}(\boldsymbol{\theta}\in\mathcal{B}_k|\mathcal{U}_k,\mathbf{a}_k)=\frac{|\mathcal{B}_k|}{|\mathcal{U}_k|}$, with $q\triangleq(1-\epsilon)\frac{|\mathcal{U}_k|}{|\mathcal{B}_k|}\leq 1$ to enforce the ϵ-outage constraint; (b) follows by maximizing $q\bar{F}_\gamma^{-1}(q|\hat{\gamma})$ over $q\in[1-\epsilon,1]$. Equality holds in (b) if $|\mathcal{B}_k|=\vartheta\,|\mathcal{U}_k|$, with $\vartheta=(1-\epsilon)/q^*$ and q^* as in the statement. The theorem is thus proved. □

8.D Proof of Theorem 2.4

Proof. Note that, if this policy satisfies $\mathcal{B}_k\equiv\mathcal{B}_{\text{t},k}\times\mathcal{B}_{\text{r},k}\subseteq\mathcal{U}_k \equiv \text{supp}(f_k)$, along with the appropriate fractional values $|\mathcal{B}_k|/|\mathcal{U}_k|$, then it is optimal since it satisfies all the conditions of Theorems 2.1 and 2.3. We now verify these conditions. Since $\mathcal{B}_{\text{t},k}\subseteq\mathcal{U}_{\text{t},k}$ and $\mathcal{B}_{\text{r},k}\subseteq\mathcal{U}_{\text{r},k}$, it is sufficient to prove that $\mathcal{U}_{\text{t},k}\times\mathcal{U}_{\text{r},k}\equiv\mathcal{U}_k, \forall k$. Indeed, $\mathcal{U}_0\equiv\mathcal{U}_{\text{t},0}\times\mathcal{U}_{\text{r},0}$. By induction, assume that $\mathcal{U}_k\equiv\mathcal{U}_{\text{t},k}\times\mathcal{U}_{\text{r},k}$. Then, for $\beta_k=1$ (a similar result holds for $\beta_k=2$), using (2.32) we get

$$\mathcal{U}_{k+1}=\begin{cases}(\mathcal{U}_{\text{t},k}\cap\mathcal{B}_{t,k})\times\mathcal{U}_{r,k}, & \text{if } C_k=\text{ACK},\\ (\mathcal{U}_{\text{t},k}\setminus\mathcal{B}_{t,k})\times\mathcal{U}_{r,k}, & \text{if } C_k=\text{NACK}.\end{cases} \tag{8.12}$$

By letting $\mathcal{U}_{\text{r},k}\equiv\mathcal{U}_{\text{r},k-1}$, $\mathcal{U}_{\text{t},k}\equiv\mathcal{U}_{\text{t},k-1}\cap\mathcal{B}_{t,k-1}$ if C_k=ACK and $\mathcal{U}_{\text{t},k}\equiv\mathcal{U}_{\text{t},k-1}\setminus\mathcal{B}_{t,k-1}$ if C_k=NACK, we obtain $\mathcal{U}_k\equiv\mathcal{U}_{\text{t},k}\times\mathcal{U}_{\text{r},k}$. This policy is then optimal. Finally, (2.65) is obtained by using the relation between power consumption and value function. Thus, we have proved the theorem. □

8.E Proof of Theorem 2.6

Proof. We prove it by induction using the DP updates. Let $\bar{T}_k(\mathcal{U}_k, e_k)$ be the *throughput-to-go* function from state $(\mathcal{U}_k, e_k)$ in slot $k \leq L^*$. We prove by induction that

$$\bar{T}_k(\mathcal{U}_k, e_k) = (1-e_k)(1-\epsilon)R_{\min} \prod_{j=k}^{L^*-1} \left[(1-\rho_k)\left(1-p_{\text{fa}}\right) + \rho_k(1-p_{\text{md}})\right]. \tag{8.13}$$

Then, (2.77) follows from $\bar{T}_{\text{err}}=\bar{T}_0(\mathcal{U}_0, 0)$. The induction hypothesis holds at $k=L^*$, since $\bar{T}_{L^*}(\mathcal{U}_{L^*}, e_{L^*}) = (1-e_{L^*})(1-\epsilon)R_{\min}$, see (2.76). Now, assume it holds for some $k+1 \leq L^*$. Using the transition probabilities from state $(\mathcal{U}_k, 1)$ and the induction hypothesis, we obtain $\bar{T}_k(\mathcal{U}_k, 1) = 0$. Instead, from state $(\mathcal{U}_k, 0)$ we obtain

$$\begin{aligned}\bar{T}_k(\mathcal{U}_k, 0) &= \rho_k(1-p_{\text{md}})\bar{T}_{k+1}(\mathcal{B}_k, 0) \\ &\quad + (1-\rho_k)\left(1-p_{\text{fa}}\right)\bar{T}_{k+1}(\mathcal{U}_k \setminus \mathcal{B}_k, 0) \\ &= (1-\epsilon)R_{\min} \prod_{j=k}^{L^*-1} \Big[(1-\rho_k)\left(1-p_{\text{fa}}\right) + \rho_k(1-p_{\text{md}})\Big],\end{aligned}$$

which readily follows by applying the induction hypothesis. The induction step is thus proved.

Let $\bar{E}_k(\mathcal{U}_k, e_k)$ be the *energy-to-go* from state $(\mathcal{U}_k, e_k)$ in slot $k \leq L^*$. We prove that

$$\bar{E}_k(\mathcal{U}_k, e_k) = \left[v_k^{(L^*)} + h_k + u_k(1-e_k)\right] |\mathcal{U}_k|. \tag{8.14}$$

Then, (2.77) follows from $\bar{P}_{\text{err}} = \frac{1}{T_{\text{fr}}}\bar{E}_0(\mathcal{U}_0, 0)$, and by noticing that $v_0^{(L^*)}/T_{\text{fr}}$ is the power consumption in the error-free case, given in Theorem 2.4. The induction hypothesis holds at $k=L^*$, since $\bar{E}_{L^*}(\mathcal{U}_{L^*}, e_{L^*})=(N-L^*)\phi_d\left(\frac{NR_{\min}}{N-L^*}, \epsilon\right)|\mathcal{U}_{L^*}| = v_{L^*}^{(L^*)} + h_{L^*} + u_{L^*}(1-e_{L^*})$, with $v_{L^*}^{(L^*)}$ given by (2.58), $h_{L^*} = u_{L^*} = 0$, see (2.76). Now, assume it holds for some $k+1 \leq L^*$.

Using the transition probabilities from state $(\mathcal{U}_k, e_k)$, the induction hypothesis, and the fact that $|\mathcal{B}_k| = \rho_k |\mathcal{U}_k|$ and $|\mathcal{U}_k \setminus \mathcal{B}_k| = (1-\rho_k)|\mathcal{U}_k|$, we obtain

$$\begin{aligned}\bar{E}_k(\mathcal{U}_k, 1) &= \phi_s \rho_k |\mathcal{U}_k| + p_{\text{fa}} \bar{E}_{k+1}(\mathcal{B}_k, 1) + (1-p_{\text{fa}}) \bar{E}_{k+1}(\mathcal{U}_k \setminus \mathcal{B}_k, 1) \\ &= \Big\{ \phi_s \rho_k + \left(v_{k+1}^{(L^*)} + h_{k+1} \right) \left[p_{\text{fa}} \rho_k + (1-p_{\text{fa}})(1-\rho_k) \right] \Big\} |\mathcal{U}_k| \,;\end{aligned}$$

$$\begin{aligned}\bar{E}_k(\mathcal{U}_k, 0) =& \phi_s \rho_k |\mathcal{U}_k| + \rho_k (1-p_{\text{md}}) \bar{E}_{k+1}(\mathcal{B}_k, 0) + (1-\rho_k) p_{\text{fa}} \bar{E}_{k+1}(\mathcal{B}_k, 1) \\ &+ (1-\rho_k)(1-p_{\text{fa}}) \bar{E}_{k+1}(\mathcal{U}_k \setminus \mathcal{B}_k, 0) + \rho_k p_{\text{md}} \bar{E}_{k+1}(\mathcal{U}_k \setminus \mathcal{B}_k, 1) \\ =& \Big\{ \phi_s \rho_k + \left(v_{k+1}^{(L^*)} + h_{k+1} + u_{k+1} \right) \left[\rho_k^2 (1-p_{\text{md}}) + (1-\rho_k)^2 (1-p_{\text{fa}}) \right] \\ &+ \left(v_{k+1}^{(L^*)} + h_{k+1} \right) \rho_k (1-\rho_k)(p_{\text{fa}} + p_{\text{md}}) \Big\} |\mathcal{U}_k| \,.\end{aligned}$$

The induction step $\bar{E}_k(\mathcal{U}_k, e_k) = (v_k^{(L^*)} + h_k + u_k(1-e_k))|\mathcal{U}_k|$ can be finally proved by expressing $v_k^{(L^*)} = g_k(\rho_k)$ and $\rho_k = \frac{1}{2} - \frac{\phi_s}{4 v_{k+1}^{(L^*)}}$ using (2.62), and using (2.79)-(2.80). □

8.F Proof of Theorem 4.1

Proof. We prove the theorem using induction. Notice that from the definition of Q-function (4.18) and the optimal value function expression (4.19) for $k = L$, we get

$$q_{L-1}(\boldsymbol{m}, a_{\text{s}}) = \int_0^\infty \frac{e^{\max_{\hat{a}} m'[\hat{a}|\boldsymbol{m}, a_{\text{s}}, y]}}{\sum_{l \in \mathcal{I}} e^{m'[l|\boldsymbol{m}, a_{\text{s}}, y]}} f(y|\boldsymbol{m}, a_{\text{s}}) dy,$$

where we have defined the preference update (4.15) as

$$m'[x|\boldsymbol{m}, a_{\text{s}}, y] = m[x] + J(y)\delta[a_{\text{s}}, x].$$

Moreover, using (4.14) and (4.11) we note that

$$\sum_{l \in \mathcal{I}} e^{m'[l|\boldsymbol{m}, a_{\text{s}}, y]} = \sum_{l \in \mathcal{I}} e^{m[l] + J(y)\delta[a_{\text{s}}, l]} = e^y f(y|\boldsymbol{m}, a_{\text{s}}) \sum_{l \in \mathcal{I}} e^{m[l]}. \tag{8.15}$$

This yields

$$q_{L-1}(\boldsymbol{m}, a_{\mathrm{s}}) \overset{(a)}{=} \frac{1}{\sum_{l\in\mathcal{I}} e^{m[l]}} \int_0^{\infty} e^{\max_{\hat{a}} m[\hat{a}]+J(y)\delta[a_{\mathrm{s}},\hat{a}]} e^{-y} \mathrm{d}y$$
$$\overset{(b)}{=} \frac{1}{\sum_{l\in\mathcal{I}} e^{m[l]}} \xi(a_{\mathrm{s}}; \boldsymbol{m}) \tag{8.16}$$

where (b) follows by evaluating the integral in (a) for the two cases in (4.22), and noting that it is given by $\xi(a_{\mathrm{s}}; \boldsymbol{m})$. Using Lemma 8.3 and (8.16)(b), the optimal value function becomes

$$V^*_{L-1}(\boldsymbol{m}) = \frac{1}{\sum_{l\in\mathcal{I}} e^{m[l]}} \left[e^{m[x_{[1]}]} + h(\nu) e^{\frac{m[x_{[2]}]-\nu m[x_{[1]}]}{1-\nu}} \right].$$

Thus, the theorem statement holds for $k{=}L-1$ with equality. Assume it holds for $k{+}1$. Using Lemma 8.3, we can bound

$$\max_{\hat{a}} \xi(\hat{a}; \boldsymbol{m}'[n|\boldsymbol{m}, a_{\mathrm{s}}, y]) \geq \exp\{\max_{\hat{a}} m'[\hat{a}|\boldsymbol{m}, a_{\mathrm{s}}, y]\}$$
$$+h(\nu) e^{\frac{\min_{x_i\neq x_j} m'[x_i|\boldsymbol{m},a_{\mathrm{s}},y]-\nu m'[x_j|\boldsymbol{m},a_{\mathrm{s}},y]}{1-\nu}}.$$

Using (4.18), the induction hypothesis (4.25) for $k{+}1$ and the above bound, we obtain $q_k(\boldsymbol{m}, a_{\mathrm{s}})$

$$\geq \int_0^{\infty} \left\{ \frac{e^{\max_{\hat{a}} m'[\hat{a}|\boldsymbol{m},a_{\mathrm{s}},y]}}{\sum_{l\in\mathcal{I}} e^{m'[l|\boldsymbol{m},a_{\mathrm{s}},y]}} + \frac{e^{\frac{\min_{x_i\neq x_j} m'[x_i|\boldsymbol{m},a_{\mathrm{s}},y]-\nu m'[x_j|\boldsymbol{m},a_{\mathrm{s}},y]}{1-\nu}}}{\sum_{l\in\mathcal{I}} e^{m'[l|\boldsymbol{m},a_{\mathrm{s}},y]}} \right.$$
$$\left. \times\, h(\nu) \frac{1-[g(\nu)]^{L-k-1}}{1-g(\nu)} \right\} f(y|\boldsymbol{m}, a_{\mathrm{s}})\, \mathrm{d}y. \tag{8.17}$$

Moreover, we note that

$$\min_{x_i\neq x_j} m'[x_i|\boldsymbol{m}, a_{\mathrm{s}}, y] - \nu m'[x_j|\boldsymbol{m}, a_{\mathrm{s}}, y]$$
$$\geq \min_{x_i\neq x_j} [m[x_i]-\nu m[x_j]] + \min_{x_i\neq x_j} J(y)\{\delta[a_{\mathrm{s}}, x_i]-\nu\delta[a_{\mathrm{s}}, x_j]\}$$
$$= \min_{x_i\neq x_j} [m[x_i] - \nu m[x_j]] + \min\{J(y), -\nu J(y)\}. \tag{8.18}$$

By substituting (8.18) and (8.15) into (8.17), yields

$$q_k(\boldsymbol{m}, a_{\rm s}) \geq \frac{1}{\sum\limits_{l\in\mathcal{I}} e^{m[l]}} \Bigg[\int_0^\infty e^{\max_{\hat{a}} m[\hat{a}] + J(y)\delta[a_{\rm s},\hat{a}]} e^{-y} {\rm d}y + e^{\frac{\min_{x_i\neq x_j} m[x_i] - \nu m[x_j]}{1-\nu}} \int_0^\infty e^{\frac{\min\{J(y), -\nu J(y)\}}{1-\nu}} e^{-y} {\rm d}y \times h(\nu) \frac{1-[g(\nu)]^{L-k-1}}{1-g(\nu)} \Bigg]. \tag{8.19}$$

The first integral in (8.19) is equal to $\xi(a_{\rm s};\boldsymbol{m})$ and the second integral is found to be equal to

$$\int_0^\infty e^{\frac{\min\{J(y), -\nu J(y)\}}{1-\nu}} e^{-y} {\rm d}y = e^{\frac{\ln\nu}{1-\nu}} \left[\frac{1}{\nu+1} - \frac{\ln\nu}{1-\nu} \right] = g(\nu) > 0.$$

Upon substituting these integrals into (8.19) yields the following lower-bound to the Q-function,

$$q_k(\boldsymbol{m}, a_{\rm s}) \geq \frac{\xi(a_{\rm s};\boldsymbol{m}) + e^{\frac{\min_{x_i\neq x_j} m[x_i] - \nu m[x_j]}{1-\nu}} h(\nu) \frac{g(\nu) - [g(\nu)]^{L-k}}{1-g(\nu)}}{\sum\limits_{l\in\mathcal{I}} e^{m[l]}},$$

which proves the induction step (4.20), and whose maximization (see Lemma 8.3) yields (4.25).

Similarly, using the induction hypothesis (4.26) for $k+1$ and the upper-bound

$$\max_{\hat{a}} \xi(\hat{a}; \boldsymbol{m}'[\boldsymbol{m}, a_{\rm s}, y]) \leq (1 + h(\nu) \exp\{\max_{\hat{a}} m'[\hat{a}|\boldsymbol{m}, a_{\rm s}, y],$$

we obtain the following upper-bound to the Q-function,

$$q_k(\boldsymbol{m}, a_{\rm s}) \leq \frac{[1+h(\nu)]^{L-k-1}}{\sum_{l\in\mathcal{I}} e^{m[l]}} \int_0^\infty e^{\max_{\hat{a}} m[\hat{a}] + J(y)\delta[a_{\rm s},\hat{a}]} e^{-y} {\rm d}y.$$

The integral above is equal to $\xi(a_{\rm s};\boldsymbol{m})$, which proves the induction step (4.21), hence

$$V_k^*(\boldsymbol{m}) = \max_{a_{\rm s}\in\mathcal{I}} q_k(\boldsymbol{m}, a_{\rm s}) \leq \frac{[1+h(\nu)]^{L-k-1}}{\sum_{l\in\mathcal{I}} e^{m[l]}} \max_{a_{\rm s}\in\mathcal{I}} \xi(a_{\rm s};\boldsymbol{m}). \tag{8.20}$$

Noting that $\max_{a_s} \xi(a_s;\boldsymbol{m}) = \xi(x_{[2]};\boldsymbol{m})$ (see Lemma 8.3), and upon substitution in (8.20) yields (4.26). □

Lemma 8.3. *We have that* $\arg\max_{a_s\in\mathcal{I}} \xi(a_s;\boldsymbol{m}) = x_{[2]}$ *and*

$$\max_{a_s\in\mathcal{I}} \xi(a_s;\boldsymbol{m}) = e^{m[x_{[1]}]} + h(\nu) e^{\frac{m[x_{[2]}]-\nu m[x_{[1]}]}{1-\nu}}. \tag{8.21}$$

Proof. To show that $\arg\max_{a_s\in\mathcal{I}} \xi(a_s;\boldsymbol{m}) = x_{[2]}$, we proceed as follows. Clearly, if $a_s \in \{x_{[2]}, x_{[3]}, \ldots, x_{[|\mathcal{I}|]}\}$, then $\max_{\hat{a}\neq a_s} m[\hat{a}] - m[a_s] = m[x_{[1]}] - m[a_s] \geq 0 > \ln(\nu)$, hence

$$\xi(a_s;\boldsymbol{m}) = e^{m[x_{[1]}]} + h(\nu) e^{\frac{m[a_s]-\nu m[x_{[1]}]}{1-\nu}},$$

maximized at $a_s = x_{[2]}$. Therefore, we restrict $a_s \in \{x_{[1]}, x_{[2]}\}$ without loss in performance. Next, we show that $\xi(x_{[2]};\boldsymbol{m}) \geq \xi(x_{[1]};\boldsymbol{m})$. Let $\Delta \triangleq m[x_{[1]}] - m[x_{[2]}]$. If $\Delta > -\ln\nu$, then $\xi(x_{[1]};\boldsymbol{m}) = e^{m[x_{[1]}]}$ and $\xi(x_{[2]};\boldsymbol{m}) > \xi(x_{[1]};\boldsymbol{m})$. Otherwise,

$$\xi(x_{[2]};\boldsymbol{m}) - \xi(x_{[1]};\boldsymbol{m}) \propto \frac{e^{\Delta}-1}{e^{\frac{\Delta}{1-\nu}} - e^{-\nu\frac{\Delta}{1-\nu}}} - h(\nu) \triangleq u(\Delta,\nu).$$

Note that $u(\Delta,\nu)$ is decreasing in $\Delta \in (0, -\ln\nu]$, $\forall \nu \in (0,1)$, minimized at $\Delta = -\ln\nu$, yielding, after algebraic steps,

$$\xi(x_{[2]};\boldsymbol{m}) - \xi(x_{[1]};\boldsymbol{m}) \propto u(\Delta,\nu) \geq \frac{h(\nu)}{e^{-\frac{1+\nu}{1-\nu}\ln\nu} - 1} > 0.$$

In both cases, $\max_{a_s} \xi(a_s) = \xi(x_{[2]})$. Upon substitution of $a_s = x_{[2]}$ in (4.22), yields (8.21). □

8.G Proof of Theorem 6.1

Proof. We will prove the theorem using induction. It can be easily verified that BT is suboptimal for $k \in \{K-1, K-2\}$ with the optimal value function given under DT action, given as

$$\hat{V}_{\beta_0,k}(\mathcal{U}) = (K-k)\frac{\max_{\hat{s}_{\mathrm{DT}}\in\mathcal{U}} \beta_0(\hat{s}_{\mathrm{DT}})}{\sum_{j\in\mathcal{U}} \beta_0(j)}, \forall k \in \{K-1, K-2\}, \mathcal{U} \subseteq \mathcal{S} \tag{8.22}$$

For $k = K-3$, $|\hat{\mathcal{S}}| \geq 2$ is suboptimal since the $\hat{V}^{(\mathrm{BT})}_{\beta_0,K-3}(\mathcal{U}, \hat{\mathcal{S}}) = 0, \forall |\hat{\mathcal{S}}| \geq 2$. The BT value function under the only feasible BT action $\hat{\mathcal{S}} = \{\hat{s}\}$ is given as

$$\hat{V}^{(\mathrm{BT})}_{\beta_0,K-3}(\mathcal{U}, \{\hat{s}\}) = \frac{1}{\sum_{j\in\mathcal{U}} \beta_0(j)} \left[\beta_0(\hat{s}) + \max_{\hat{s}_{\mathrm{DT}}\in\mathcal{U}\setminus\{\hat{s}\}} \beta_0(\hat{s}_{\mathrm{DT}})\right] \tag{8.23}$$

$$\leq \frac{\beta_0(1) + \beta_0(2)}{\sum_{j\in\mathcal{U}} \beta_0(j)}, \tag{8.24}$$

with equality if $\hat{s} = \arg\max_{s\in\mathcal{U}} \beta_0(s)$, i.e., $\hat{\mathcal{S}}^* = \hat{\mathcal{S}}^{(1)}$. Let the induction hypothesis hold for all $k \in \{K-1, K-2, \ldots, k_0+1\}$, then we will show that it holds for $k_0 \geq 0$. To this end, we will use a contradiction argument. Without loss of generality, let $m \in \mathcal{M}_{k_0}(|\mathcal{U}|)$ be a constant and let $\mathcal{U}$ be ordered as $\mathcal{U} = \{s_1, s_2, \ldots, s_{|\mathcal{U}|}\}$, so that $\beta_0(s_1) \geq \beta_0(s_{m+1}) \geq \beta_0(s_{m+2}) \geq \cdots \geq \beta_0(s_{|\mathcal{U}|})$ and $\beta_0(s_2), \ldots, \beta_0(s_m)$ can follow any arbitrary order. Consider two BT actions $\hat{\mathcal{S}}_0 = \{s_1, s_2, s_3, \ldots, s_m\} \subset \mathcal{U}$ and $\hat{\mathcal{S}}_1 = \{s_2, s_3, \ldots, s_m, s_{m+1}\} \subset \mathcal{U}$. Note that $\hat{\mathcal{S}}_0$ is obtained by replacing the least likely BPI in $\hat{\mathcal{S}}_1$ (s_{m+1}) with the most likely beam in $\mathcal{U} \setminus \hat{\mathcal{S}}_1(s_1)$. Let $\delta_0 \triangleq K - k_0$ be the number of remaining slots in the frame. Then, the BT value function under the BT action $\hat{\mathcal{S}}_0$ is given as

$$\begin{aligned}
&\hat{V}^{(\mathrm{BT})}_{\beta_0,k_0}(\mathcal{U}, \hat{\mathcal{S}}_0) \\
&= \frac{1}{\sum_{j\in\mathcal{U}} \beta_0(j)} \Bigg[\sum_{i=1}^{m} \beta_0(s_i)(\delta_0 - m - 1) + \max\Bigg\{(\delta_0 - m - 1)\beta_0(s_{m+1}), \\
&\max_{m'\in\mathcal{M}_{k_0+m+1}(|\mathcal{U}|-m)} (\delta_0 - m - m' - 2) \sum_{i=m+1}^{m+m'} \beta_0(s_i) + \left[\sum_{i=m+m'+1}^{|\mathcal{U}|} \beta_0(s_i)\right] \hat{V}_{\beta_0,k_0+m+m'+2}(\mathcal{U}[m'])\Bigg\}\Bigg],
\end{aligned} \tag{8.25}$$

where we have used the induction hypothesis for $k = k_0 + m + 1$ and value function update (6.32) and (6.33) and $\mathcal{U}[m'] \triangleq \{m + m' + 1, \ldots, |\mathcal{U}|\}$ is support after performing the next BT round with BPI $\hat{\mathcal{S}}_0[m'] \equiv \{s_{m+1}, \ldots, s_{m+m'}\}$ and under $Y = \emptyset$. Under action $\hat{\mathcal{S}}_1$, the BT value function is given as

$$\begin{aligned}
&\hat{V}^{(\mathrm{BT})}_{\beta_0,k_0}(\mathcal{U}, \hat{\mathcal{S}}_1) \\
&= \frac{1}{\sum_{j\in\mathcal{U}} \beta_0(j)} \Bigg[\sum_{i=2}^{m+1} \beta_0(s_i)(\delta_0 - m - 1) + \max \Bigg\{ (\delta_0 - m - 1)\beta_0(s_1), \\
&\max_{m' \in \mathcal{M}_{k_0+m+1}(|\mathcal{U}|-m)} (\delta_0 - m - m' - 2) \left[\beta_0(s_1) + \sum_{i=m+2}^{m+m'} \beta_0(s_i) \right] + \sum_{i=m+m'+1}^{|\mathcal{U}|} \beta_0(s_i) \hat{V}_{\beta_0,k_0+m+m'+2}(\mathcal{U}[m']) \Bigg\} \Bigg],
\end{aligned} \tag{8.26}$$

where we have used the induction hypothesis for $k = k_0 + m + 1$, where the BPIs in $\hat{\mathcal{S}}_1[m'] \equiv \{s_1, s_{m+2}, \ldots, s_{m+m'}\}$ are scanned in the next action if BT is selected. Using the above two BT value functions, we get

$$\begin{aligned}
&\hat{V}^{(\mathrm{BT})}_{\beta_0,k_0}(\mathcal{U}, \hat{\mathcal{S}}_0) - \hat{V}^{(\mathrm{BT})}_{\beta_0,k_0}(\mathcal{U}, \hat{\mathcal{S}}_1) \\
&\propto \max \Bigg\{ 0, \max_{m' \in \mathcal{M}_{k_0+m+1}(|\mathcal{U}|-m)} \sum_{i=m+2}^{m+m'} \beta_0(s_i)(\delta_0 - m - m' - 2) - (m'+1)\beta_0(s_{m+1}) \\
&\qquad\qquad + \sum_{i=m+m'+1}^{|\mathcal{U}|} \beta_0(s_i) \hat{V}_{\beta_0,k_0+m+m'+2}(\mathcal{U}[m']) \Bigg\} \\
&\quad - \max \Bigg\{ 0, \max_{m' \in \mathcal{M}_{k_0+m+1}(|\mathcal{U}|-m)} \sum_{i=m+2}^{m+m'} \beta_0(s_i)(\delta_0 - m - m' - 2) - (m'+1)\beta_0(s_1) \\
&\qquad\qquad + \sum_{i=m+m'+1}^{|\mathcal{U}|} \beta_0(s_i) \hat{V}_{\beta_0,k_0+m+m'+2}(\mathcal{U}[m']) \Bigg\} \\
&\geq 0,
\end{aligned} \tag{8.27}$$

since $\beta_0(s_1) \geq \beta_0(s_{m+1})$. Therefore, the BT value function improves by using $\hat{\mathcal{S}}_0$ instead of $\hat{\mathcal{S}}_1$, i.e., BT value function improves by replacing the least likely BPI from $\hat{\mathcal{S}}_1$ with most-likely BPI from remaining beam indices, $\mathcal{U} \setminus \hat{\mathcal{S}}_1$. Therefore, we can use the same argument recursively to improve the value function by replacing least likely BPI in any BT BPI set $\hat{\mathcal{S}}$

until no BPI is more-likely in $\mathcal{U} \setminus \hat{\mathcal{S}}$ than any BPI in $\hat{\mathcal{S}}$. Finally, the optimal m can be found by solving

$$m^* = \arg \max_{m \in \mathcal{M}_k(|\mathcal{U}|)} \hat{V}^{(\mathrm{BT})}_{\beta_0,k}(\mathcal{U}, \hat{\mathcal{S}}[m]) \tag{8.28}$$

□

8.H Proof of Corollary 6.1

Proof. We prove the Corollary by induction. The statement of Corollary holds for $k = 0$ by definition of $\mathcal{U}_0$. Let the statement hold for $k \geq 0$; then we will show that statement holds for $k + L$. Since the statement holds for $k + L - 1$, the support is either $\mathcal{U}_{k+L-1} = \{u_{k+L-1}\}$ or $\mathcal{U}_k = \{u_{k+L-1}, \ldots, u_0 + w_0 - 1\}$ with $u_{k+L-1} < u_0 + w_0 - 1$. In the case $\mathcal{U}_{k+L-1} = \{u_{k+L-1}\}$ (singleton support), no BT action is feasible. In the second case, if (non-singleton $\mathcal{U}_{k+L-1}$), let $\hat{\mathcal{S}}_k[m] \in \mathcal{A}_{\beta_0,k}(\mathcal{U}_{k+L-1})$ be any BT action selected. Then, the support is updated following (6.30), yielding

$$\mathcal{U}_{k+L} = \begin{cases} \{u_{k+L-1} + m, \ldots, u_0 + w_0 - 1\}, & Y_k = \emptyset \\ \{s^*\}, & Y_k = s^* \in \hat{\mathcal{S}}_k[m] \end{cases} \tag{8.29}$$

By letting $u_{k+L} = u_{k+L-1} + m$ and $w_{k+L} = u_0 + w_0 - u_{k+L}$ if $Y_k = \emptyset$ and $u_{k+L} = s^*$ and $w_{k+L} = 1$ if $Y_k = s^* \in \hat{\mathcal{S}}_k[m]$, we have shown the statement to hold for $k + L$. □

REFERENCES

[1] CISCO, *Cisco visual networking index: Global mobile data traffic forecast update, 2016–2021 white paper*. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html.

[2] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter Wave Channel Modeling and Cellular Capacity Evaluation," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1164–1179, Jun. 2014, ISSN: 0733-8716. DOI: 10.1109/JSAC.2014.2328154.

[3] T. S. Rappaport, R. W. Heath, R. C. Daniels, and J. N. Murdock, *Millimeter wave wireless communications*. Prentice Hall, 2015.

[4] M. Hussain and N. Michelusi, "Energy-efficient interactive beam alignment for millimeter-wave networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 2, pp. 838–851, Feb. 2019, ISSN: 1536-1276. DOI: 10.1109/TWC.2018.2885041.

[5] M. Hussain and N. Michelusi, "Coded energy-efficient beam-alignment for millimeter-wave networks," in *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2018, pp. 407–412. DOI: 10.1109/ALLERTON.2018.8635944.

[6] M. Hussain and N. Michelusi, "Second-best beam-alignment via bayesian multi-armed bandits," in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6. DOI: 10.1109/GLOBECOM38437.2019.9013578.

[7] M. Hussain, M. Scalabrin, M. Rossi, and N. Michelusi, "Mobility and blockage-aware communications in millimeter-wave vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 11, pp. 13 072–13 086, 2020. DOI: 10.1109/TVT.2020.3020898.

[8] M. T. J. Spaan and N. Vlassis, "Perseus: Randomized point-based value iteration for pomdps," *J. Artif. Int. Res.*, vol. 24, no. 1, pp. 195–220, Aug. 2005, ISSN: 1076-9757. [Online]. Available: http://dl.acm.org/citation.cfm?id=1622519.1622525.

[9] N. Michelusi and M. Hussain, "Optimal beam-sweeping and communication in mobile millimeter-wave networks," in *2018 IEEE International Conference on Communications (ICC)*, May 2018, pp. 1–6. DOI: 10.1109/ICC.2018.8422675.

[10] M. Hussain and N. Michelusi, "Throughput optimal beam alignment in millimeter wave networks," in *2017 Information Theory and Applications Workshop (ITA)*, Feb. 2017, pp. 1–6. DOI: 10.1109/ITA.2017.8023460.

[11] R. A. Hassan and N. Michelusi, "Multi-user beam-alignment for millimeter-wave networks," in *2018 Information Theory and Applications Workshop (ITA)*, Feb. 2018, pp. 1–6.

[12] M. Hussain and N. Michelusi, "Energy efficient beam-alignment in millimeter wave networks," in *2017 51st Asilomar Conference on Signals, Systems, and Computers*, Oct. 2017, pp. 1219–1223. DOI: 10.1109/ACSSC.2017.8335545.

[13] J. Zhang, Y. Huang, Q. Shi, J. Wang, and L. Yang, "Codebook design for beam alignment in millimeter wave communication systems," *IEEE Transactions on Communications*, vol. 65, no. 11, pp. 4980–4995, Nov. 2017, ISSN: 0090-6778. DOI: 10.1109/TCOMM.2017.2730878.

[14] M. Hussain and N. Michelusi, "Optimal interactive energy efficient beam-alignment for millimeter-wave networks," in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, 2018, pp. 577–581. DOI: 10.1109/ACSSC.2018.8645195.

[15] S. Haghighatshoar and G. Caire, "The beam alignment problem in mmwave wireless networks," in *2016 50th Asilomar Conference on Signals, Systems and Computers*, Nov. 2016, pp. 741–745. DOI: 10.1109/ACSSC.2016.7869144.

[16] C. Jeong, J. Park, and H. Yu, "Random access in millimeter-wave beamforming cellular networks: issues and approaches," *IEEE Communications Magazine*, vol. 53, no. 1, pp. 180–185, Jan. 2015, ISSN: 0163-6804. DOI: 10.1109/MCOM.2015.7010532.

[17] V. Desai, L. Krzymien, P. Sartori, W. Xiao, A. Soong, and A. Alkhateeb, "Initial beamforming for mmWave communications," in *48th Asilomar Conference on Signals, Systems and Computers*, Nov. 2014. DOI: 10.1109/ACSSC.2014.7094805.

[18] J. Seo, Y. Sung, G. Lee, and D. Kim, "Training beam sequence design for millimeter-wave mimo systems: A pomdp framework," *IEEE Transactions on Signal Processing*, vol. 64, no. 5, pp. 1228–1242, Mar. 2016, ISSN: 1053-587X. DOI: 10.1109/TSP.2015.2496241.

[19] N. Gonzalez-Prelcic, R. Mendez-Rial, and R. W. Heath, "Radar aided beam alignment in MmWave V2I communications supporting antenna diversity," in *2016 Information Theory and Applications Workshop (ITA)*, Jan. 2016, pp. 1–7. DOI: 10.1109/ITA.2016.7888145.

[20] T. Nitsche, A. B. Flores, E. W. Knightly, and J. Widmer, "Steering with eyes closed: Mm-wave beam steering without in-band measurement," in *2015 IEEE Conference on Computer Communications (INFOCOM)*, Apr. 2015, pp. 2416–2424. DOI: 10.1109/INFOCOM.2015.7218630.

[21] V. Va, T. Shimizu, G. Bansal, and R. W. Heath, "Beam design for beam switching based millimeter wave vehicle-to-infrastructure communications," in *2016 IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–6. DOI: 10.1109/ICC.2016.7511414.

[22] 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*, vol. 67, no. 5, pp. 4042–4058, May 2018, ISSN: 0018-9545. DOI: 10.1109/TVT.2017.2787627.

[23] A. Alkhateeb, O. E. Ayach, G. Leus, and R. W. Heath, "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 831–846, Oct. 2014, ISSN: 1932-4553. DOI: 10.1109/JSTSP.2014.2334278.

[24] Z. Marzi, D. Ramasamy, and U. Madhow, "Compressive channel estimation and tracking for large arrays in mm-wave picocells," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 514–527, Apr. 2016, ISSN: 1932-4553. DOI: 10.1109/JSTSP.2016.2520899.

[25] "IEEE Std 802.15.3c-2009," *IEEE Standard*, pp. 1–200, Oct. 2009. DOI: 10.1109/IEEESTD.2009.5284444.

[26] "IEEE Std 802.11ad-2012," *IEEE Standard*, pp. 1–628, Dec. 2012. DOI: 10.1109/IEEESTD.2012.6392842.

[27] S. Noh, M. D. Zoltowski, and D. J. Love, "Multi-resolution codebook and adaptive beamforming sequence design for millimeter wave beam alignment," *IEEE Transactions on Wireless Communications*, vol. 16, no. 9, pp. 5689–5701, 2017, ISSN: 1536-1276. DOI: 10.1109/TWC.2017.2713357.

[28] V. Raghavan, J. Cezanne, S. Subramanian, A. Sampath, and O. Koymen, "Beamforming tradeoffs for initial ue discovery in millimeter-wave mimo systems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 3, pp. 543–559, Apr. 2016, ISSN: 1932-4553. DOI: 10.1109/JSTSP.2016.2523442.

[29] T. Bai and R. W. Heath, "Coverage and Rate Analysis for Millimeter-Wave Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 2, pp. 1100–1114, Feb. 2015, ISSN: 1536-1276. DOI: 10.1109/TWC.2014.2364267.

[30] J. Zhang, Y. Huang, Q. Shi, J. Wang, and L. Yang, "Codebook design for beam alignment in millimeter wave communication systems," *IEEE Transactions on Communications*, vol. 65, no. 11, pp. 4980–4995, 2017.

[31] V. Suresh and D. J. Love, "Error Control Sounding Strategies for Millimeter Wave Beam Alignment," in *Information Theory and Applications Workshop (ITA)*, Feb. 2018.

[32] Y. Shabara, C. E. Koksal, and E. Ekici, "Linear block coding for efficient beam discovery in millimeter wave communication networks," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, 2018, pp. 2285–2293. DOI: 10.1109/INFOCOM.2018.8486302.

[33] M. Hashemi, A. Sabharwal, C. Emre Koksal, and N. B. Shroff, "Efficient beam alignment in millimeter wave systems using contextual bandits," in *IEEE INFOCOM 2018*, Apr. 2018, pp. 2393–2401. DOI: 10.1109/INFOCOM.2018.8486279.

[34] S. Chiu, N. Ronquillo, and T. Javidi, "Active learning and csi acquisition for mmwave initial alignment," *IEEE Journal on Selected Areas in Communications*, pp. 1–1, 2019. DOI: 10.1109/JSAC.2019.2933967.

[35] A. G. Sutton Richard S. Barto, *Reinforcement learning: An introduction*. MIT Press, 2018.

[36] J. Pineau, G. Gordon, and S. Thrun, "Anytime point-based approximations for large pomdps," *J. Artif. Int. Res.*, vol. 27, no. 1, pp. 335–380, Nov. 2006, ISSN: 1076-9757. [Online]. Available: http://dl.acm.org/citation.cfm?id=1622572.1622582.

[37] M. T. J. Spaan and N. Vlassis, "Perseus: Randomized point-based value iteration for pomdps," *J. Artif. Int. Res.*, vol. 24, no. 1, pp. 195–220, Aug. 2005, ISSN: 1076-9757.

[38] M. Gapeyenko, A. Samuylov, M. Gerasimenko, D. Moltchanov, S. Singh, M. R. Akdeniz, E. Aryafar, N. Himayat, S. Andreev, and Y. Koucheryavy, "On the temporal effects of mobile blockers in urban millimeter-wave cellular scenarios," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 11, pp. 10 124–10 138, 2017.

[39] N. Michelusi and M. Hussain, "Optimal beam-sweeping and communication in mobile millimeter-wave networks," in *2018 IEEE International Conference on Communications (ICC)*, May 2018, pp. 1–6. DOI: 10.1109/ICC.2018.8422675.

[40] 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.

[41] 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*, IEEE, 2016, pp. 1–6.

[42] M. Scalabrin, N. Michelusi, and M. Rossi, "Beam training and data transmission optimization in millimeter-wave vehicular networks," in *2018 IEEE Global Communications Conference (GLOBECOM)*, Dec. 2018, pp. 1–7. DOI: 10.1109/GLOCOM.2018.8647890.

[43] M. Mezzavilla, S. Goyal, S. Panwar, S. Rangan, and M. Zorzi, "An mdp model for optimal handover decisions in mmwave cellular networks," in *Networks and Communications (EuCNC), 2016 European Conference on*, IEEE, 2016, pp. 100–105.

[44] J. Pan and W. Zhang, "An mdp-based handover decision algorithm in hierarchical lte networks," in *Vehicular Technology Conference (VTC Fall), 2012 IEEE*, IEEE, 2012, pp. 1–5.

[45] E. Stevens-Navarro, Y. Lin, and V. W. Wong, "An mdp-based vertical handoff decision algorithm for heterogeneous wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 2, pp. 1243–1254, 2008.

[46] A. Alkhateeb, S. Alex, P. Varkey, Y. Li, Q. Qu, and D. Tujkovic, "Deep learning coordinated beamforming for highly-mobile millimeter wave systems," *IEEE Access*, vol. 6, pp. 37 328–37 348, 2018. DOI: 10.1109/ACCESS.2018.2850226.

[47] V. Va, T. Shimizu, G. Bansal, and R. W. Heath, "Online learning for position-aided millimeter wave beam training," *IEEE Access*, vol. 7, pp. 30 507–30 526, 2019. DOI: 10.1109/ACCESS.2019.2902372.

[48] 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*, IEEE, 2017, pp. 1–6.

[49] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," in *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[50] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989. DOI: 10.1109/5.18626.

[51] A. Ali, N. González-Prelcic, and R. W. Heath, "Millimeter wave beam-selection using out-of-band spatial information," *IEEE Transactions on Wireless Communications*, vol. 17, no. 2, pp. 1038–1052, 2018. DOI: 10.1109/TWC.2017.2773532.

[52] T. .-H. Chou, N. Michelusi, D. Love, and J. V. Krogmeier, "Fast position-aided mimo beam training via noisy tensor completion," *IEEE Journal of Selected Topics in Signal Processing*, pp. 1–1, 2021. DOI: 10.1109/JSTSP.2021.3063837.

[53] Y. Wang, N. J. Myers, N. González-Prelcic, and R. W. H. J. au2, *Deep learning-based compressive beam alignment in mmwave vehicular systems*, 2021. arXiv: 2103.00125 [eess.SP].

[54] Y. Heng and J. G. Andrews, "Machine learning-assisted beam alignment for mmwave systems," in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6. DOI: 10.1109/GLOBECOM38437.2019.9013296.

[55] A. A. M. Saleh and R. Valenzuela, "A statistical model for indoor multipath propagation," *IEEE Journal on Selected Areas in Communications*, vol. 5, no. 2, pp. 128–137, Feb. 1987, ISSN: 0733-8716. DOI: 10.1109/JSAC.1987.1146527.

[56] C. N. Barati, S. A. Hosseini, M. Mezzavilla, T. Korakis, S. S. Panwar, S. Rangan, and M. Zorzi, "Initial Access in Millimeter Wave Cellular Systems," *IEEE Transactions on Wireless Communications*, vol. 15, no. 12, pp. 7926–7940, Dec. 2016, ISSN: 1536-1276. DOI: 10.1109/TWC.2016.2609384.

[57] Y. Li, J. G. Andrews, F. Baccelli, T. D. Novlan, and C. J. Zhang, "Design and analysis of initial access in millimeter wave cellular networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 10, pp. 6409–6425, Oct. 2017, ISSN: 1536-1276. DOI: 10.1109/TWC.2017.2723468.

[58] V. Va, J. Choi, and R. W. Heath, "The impact of beamwidth on temporal channel variation in vehicular channels and its implications," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 6, pp. 5014–5029, Jun. 2017, ISSN: 0018-9545. DOI: 10.1109/TVT.2016.2622164.

[59] M. Gapeyenko, A. Samuylov, M. Gerasimenko, D. Moltchanov, S. Singh, M. R. Akdeniz, E. Aryafar, N. Himayat, S. Andreev, and Y. Koucheryavy, "On the temporal effects of mobile blockers in urban millimeter-wave cellular scenarios," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 11, pp. 10 124–10 138, Nov. 2017, ISSN: 0018-9545. DOI: 10.1109/TVT.2017.2754543.

[60] M. Hussain, D. J. Love, and N. Michelusi, "Neyman-Pearson Codebook Design for Beam Alignment in Millimeter-Wave Networks," in *the 1st ACM Workshop on Millimeter-Wave Networks and Sensing Systems*, ser. mmNets '17, Snowbird, Utah, USA, 2017, ISBN: 978-1-4503-5143-0. DOI: 10.1145/3130242.3130247.

[61] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 37, 1948.

[62] 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, Mar. 2014, ISSN: 0018-9219. DOI: 10.1109/JPROC.2014.2299397.

[63] M. K. Simon, *Probability Distributions Involving Gaussian Random Variables*. Springer Pr., 2002.

[64] D. P. Bertsekas, *Dynamic programming and optimal control*. Athena Scientific, 2005.

[65] J. Song, J. Choi, and D. J. Love, "Codebook design for hybrid beamforming in millimeter wave systems," in *2015 IEEE International Conference on Communications (ICC)*, Jun. 2015, pp. 1298–1303.

[66] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE journal on selected areas in communications*, vol. 32, no. 6, pp. 1164–1179, 2014.

[67] M. Hussain, M. Scalabrin, M. Rossi, and N. Michelusi, "Adaptive millimeter-wave communications exploiting mobility and blockage dynamics," in *IEEE International Conference on Communications (ICC)*, 2020, pp. 1–6.

[68] M. Baghani, S. Parsaeefard, M. Derakhshani, and W. Saad, "Dynamic non-orthogonal multiple access and orthogonal multiple access in 5g wireless networks," *IEEE Transactions on Communications*, vol. 67, no. 9, pp. 6360–6373, 2019.

[69] N. Michelusi, U. Mitra, A. F. Molisch, and M. Zorzi, "Uwb sparse/diffuse channels, part i: Channel models and bayesian estimators," *IEEE Transactions on Signal Processing*, vol. 60, no. 10, pp. 5307–5319, Oct. 2012. DOI: 10.1109/TSP.2012.2205681.

[70] D.-S. Shim, C.-K. Yang, J. Kim, J. Han, and Y. Cho, "Application of motion sensors for beam-tracking of mobile stations in mmwave communication systems," *Sensors*, vol. 14, no. 10, pp. 19 622–19 638, Oct. 2014, ISSN: 1424-8220. DOI: 10.3390/s141019622.

[71] J. D. C. Little and S. Graves, "Little's law," in. Jul. 2008, pp. 81–100. DOI: .

[72] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge Univ. Pr., 2011.

[73] C. A. Balanis, *Antenna theory: analysis and design*. Wiley, 2016.

[74] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019, ISSN: 1935-8237. DOI: 10.1561/2200000056. [Online]. Available: http://dx.doi.org/10.1561/2200000056.

[75] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," in *ICLR*, 2017. [Online]. Available: https://arxiv.org/abs/1611.01144.