Physics-Informed Parametric Bandits for Beam Alignment in mmWave Communications

Hao Qin [*1] **Thang Duong** [*2] **Ming Li** [3] **Chicheng Zhang** [2]

Abstract

In millimeter wave (mmWave) communications, aligning the transmitter and receiver beams is crucial to reduce the significant path loss. As scanning the entire directional space is inefficient, designing an efficient and robust method to identify the correct optimal beamforming direction is essential. Many existing works use bandit algorithms for beam alignment but rely on unimodality or multimodality assumptions on the reward structure, as well as assuming the horizon is sufficiently long. However, such assumptions may not hold in practice and cause such algorithms to converge to choosing suboptimal beams.

In this work, we propose the physics-informed algorithms PR-ETC and PR-GREEDY that exploit the existence of a dominant path (e.g., LoS path), an assumption that is perhaps more realistic in practice, which has a connection to the Phase Retrieval Bandit problem. Through simulated experiments using the DeepMIMO dataset (Alkhateeb, 2019), we demonstrate that both algorithms outperform existing approaches across 4,952 bandit instances.

1. Introduction

We study beam alignment in Millimeter-wave (mmWave) wireless communications. mmWave communication leverages reconfigurable high-gain antenna arrays to achieve high data rates. For example, when a Base Station (BS) and a user equipment (UE) would like to communicate with each other, BS performs *beam steering*, i.e., it dynamically reconfigures its phased array antenna by adjusting the phase delay of the signal fed to each antenna element, allowing the signal to be sent or received in the desired directions. The *beam pattern* of an antenna array configuration characterizes the received signal strength of users at different directions.

In modern applications, one chooses from a predefined codebook of antenna configurations, whose beam patterns are so narrow that even a slight misalignment of the beam to the user location can result in a significant loss in signal strength. For example, Nitsche et al. (2015) notes that with a 7-degree beam width, an 18-degree misalignment can reduce the link signal strength by 17 dB. Aligning the beam accurately with the direction that provides the highest gain is crucial to fully harness the potential of next-generation communication hardware.

Offline training-based methods (such as beam scanning) can be adopted; however, they are inefficient and incur high overhead due to the large beam space. To fulfill real-time communication requirements, there is a critical need to develop an algorithm to select the best beam in a sample-efficient and online manner. In this paper, we investigate beam alignment in a short-horizon setting.

An effective strategy for this beam alignment problem is to cast it as a Multi-Armed Bandit (MAB) problem. Specifically, each possible beam can be viewed as an arm of a slot machine, and the corresponding channel gain serves as the reward feedback. In each round, a learning agent selects an arm (i.e., a beam) and receives feedback drawn from the distribution associated with that arm. The goal is for the agent to adaptively choose the arms online that yield the highest expected reward through repeated interactions.

Previously, effective algorithms for the MAB problem have been proposed, such as Upper Confidence Bound (UCB) (Lai, 1987; Auer et al., 2002) and Explore-Then-Commit (ETC) (Langford & Zhang, 2007; Lattimore & Szepesvári, 2020). To improve sample efficiency, Yu & Mannor (2011); Cutkosky et al. (2023) studies the setting where the expected reward is a unimodal function of the arm (i.e., it has only one peak), a framework adopted by several prior works in beam alignment (Hashemi et al., 2018; Ghosh et al., 2024). Subsequent works (Saber & Maillard, 2024) extend this to the setting where the reward is a multimodal

[1]Department of Mathematics, University of Arizona, Tucson, AZ, US [2]Department of Computer Science, University of Arizona, Tucson, AZ, US [3]Department of Electrical and Computer Engineering, University of Arizona, Tucson, AZ, US. Correspondence to: Thang Duong <thangduong@arizona.edu>, Hao Qin <hqin@arizona.edu>, Ming Li <lim@arizona.edu>, Chicheng Zhang <chichengz@arizona.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

function of the arm, however, they typically have restrictive assumptions about a known (or a bound) number of peaks.

Nevertheless, multi-armed bandit approaches and their extensions are insufficient for mmWave beam alignment problems:

1. The real-world mmWave channel induces reward functions that are often not unimodal. Even with only a Line-of-Sight (LoS) path, because of the antenna sidelobes, the reward function contains many local peaks – see Figure 1. In addition, the mmWave signal propagation is generally modeled using a geometric-based statistical model which includes multiple reflection paths (or clusters) (Rappaport et al., 2017; Gustafson et al., 2013), which is more complex than the LoS path-loss model in (Samimi & Rappaport, 2014). In reality, the channel model parameters, such as the number of paths, reflection coefficients, and the path-loss exponents, all depend on the environment, which cannot be exactly known in advance. Previous methods that assume unimodality or multimodality (Yu & Mannor, 2011; Hashemi et al., 2018; Cutkosky et al., 2023) fail to generalize under such realistic channel settings.

2. Real-world applications of beam alignment oftentimes require low latency for real-time applications, and requires selecting good beams in a short time horizon (Hassanieh et al., 2018; Mazaheri et al., 2019).

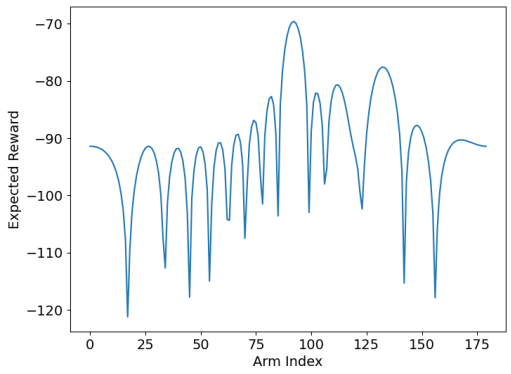

Figure 1: The expected reward function of a bandit instance from the DeepMIMO simulated dataset. It's challenging to generalize bandit algorithms that focus on exploiting the unimodal property, especially in a sample efficient manner.

In this paper, inspired by works on parametric bandits (Filippi et al., 2010), we propose physics-informed parametric bandit algorithms, PR-ETC and PR-GREEDY, for beam selection by exploiting the underlying reward function structure similar to the Phase Retrieval (PR) Bandit problem. We found that the electromagnetic wave propagation model commonly used in far-field mmWave channel modeling (Samimi & Rappaport, 2014) can be viewed as a phase retrieval model and solved using a similar approach for Generalized Linear Model (GLM) bandits in Filippi et al. (2010) , while taking into consideration the properties of phase-array antennas. Empirically, we demonstrate that leveraging the unimodal or multimodal property is not as efficient as leveraging the underlying wave propagation structure. Our experiments also show that PR-GREEDY is robust under model misspecification. Furthermore, the experiment results show the adaptability and robustness of our algorithm across 4,952 bandit instances from the DeepMIMO simulated environment (Alkhateeb, 2019), in a parameter-free manner , without imposing any extra assumptions on the underlying structure of the reward function (such as the number of paths, the number of peaks, etc.). We also proposed the PR-ETC as a more computationally efficient approximation of PR-GREEDY to trade off between performance and computational latency.

2. Related work

Many works have studied beam alignment in mmWave wireless communications, using the online learning regret as the evaluation metric (Aykin et al., 2020). Hashemi et al. (2018)'s algorithm can be seen as a mixture of local search and greedy action, with an early stopping condition based on the ratio between the peak and the average estimated channel gain. Aykin et al. (2020) addressed the beam-tracking problem in dynamic environments using a model-free approach and adapted the Thompson Sampling method to solve it. While their experimental results are impressive and were conducted with real hardware, their approach is based on strong assumptions about the shape of the reward distribution (categorical distribution), which limits its applicability in more general or practical settings.

Yu & Mannor (2011); Cutkosky et al. (2023) focus on bandit with a unimodal structure of the reward function. Yu & Mannor (2011) proposes the Line Search Elimination Algorithm (LSE) to efficiently select the action with the highest expected reward by leveraging the unimodal property in general, not specifically designed for beam alignment. The noiseless and noisy 1-d convex bandit algorithms in Lattimore (2024) can also be used to exploit the unimodal structure. Cutkosky et al. (2023) proposes a more sample-efficient approach, but it maintains expected reward estimates with a large collection of arms, which limits its application in real-time deployment.

For the bandit setting with a multimodal reward function, Saber & Maillard (2024) proposes the IMED-MB algorithm, which uses the IMED index from Honda & Takemura (2015).

Saber & Maillard (2024) shows that IMED-MB is asymptotically optimal in this setting and demonstrates impressive empirical results.

The difficulty of short-horizon problems motivates us to investigate solving beam alignment by viewing it as a parametric bandit problem. Parametric bandits with GLM structure have been previously studied by Filippi et al. (2010). Here, the authors model the reward distribution using a canonical exponential family, whose canonical parameter is a linear function of the arm features. Then, they propose the GLM-UCB algorithm that utilizes confidence bounds for exploration.

In contrast, the physics-informed mmWave propagation model we employ has a phase retrieval aspect (Kotłowski & Neu, 2019; Lattimore & Hao, 2021) as the reward feedback is only related to the signal strength but not phase shift. Our PR-GREEDY algorithm simply uses Maximum Likelihood Estimation (MLE) to estimate the parameters of the environment and act greedily.

3. Problem formulation

3.1. Far-Field Channel Strength

We model the channel using the far-field channel from a uniform linear array (ULA) antenna with a dominant path. Such a dominant path may be the Line-of-Sight (LoS) path or other paths. Given an array of antennas, the total channel received by the receiver is the sum of the channels provided by each antenna. Suppose that we have $2\bar{N}+1$ antennas and give the $\bar{N}+1$-th antenna element as index 0 , as in Figure 2. Therefore, the **array response vector** of this ULA array can be represented as a complex vector, H:

$$H := \left(\zeta e^{-j\frac{2\pi}{\lambda}\mathbf{r}_n}\right)_{n=-\bar{N}}^{\bar{N}}$$

where λ denotes the wavelength of the carrier, j is the imaginary unit, $\mathbf{r}_n$ is the distance from the n-th antenna to the receiver, and ζ denotes the channel gain (amplitude) for the antennas. Now, consider the layout shown in Figure 2: $\mathbf{r}$ is the distance from the center of the ULA antenna to the receiver, $\mathbf{s}_n$ is the position vector of the n-th antenna element, θ and ϕ denote the azimuth and elevation angles of the receiver with respect to the x–z plane. The elevation angle ϕ does not affect the channel and can be assumed to be 0 without loss of generality. Therefore, $\mathbf{r}_n \approx ||r|| - nd\cos(\theta)$ and the array response vector H is equal to:

$$H = \zeta e^{-j\frac{2\pi||\mathbf{r}||}{\lambda}}\left(e^{j\frac{2\pi}{\lambda}(nd\cos\theta)}\right)_{n=-\bar{N}}^{\bar{N}}.$$

To be more convenience in optimization, we define:

$$h(\theta) := \left(e^{j\frac{2\pi}{\lambda}(nd\cos\theta)}\right)_{n=-\bar{N}}^{\bar{N}}$$
$$H(\beta,\theta) := \beta h(\theta),$$

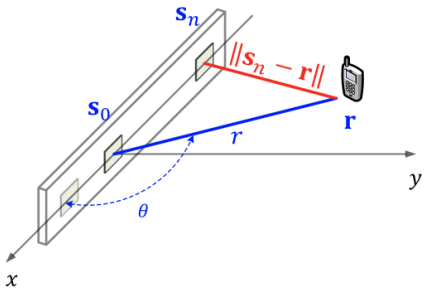

Figure 2: System layout of MISO system with ULA antenna. Image from (Liu et al., 2023).

where $\beta := \zeta e^{-2\pi j||\mathbf{r}||/\lambda}$ such that H can be represented as a function of (β,θ), denoted as $H(\beta,\theta)$. The parameters β,θ come from a space Θ that depends on the environment and the receiver's location.

Next, for the antenna in the receiver, we can control the phase shift of each antenna and each configuration corresponds to a different steering vector $\mathbf{f}$.

We assume that we have a codebook of K different antenna configurations (namely, steering vectors) that correspond to different beam directions, forming an action set $\{\mathbf{f}_a : a \in [K]\}$. The steering vector $\mathbf{f}_a$ is generated in the form:

$$\mathbf{f}_a := \left(e^{j\frac{2\pi}{\lambda}(nd\cos\pi\frac{a}{K})}\right)_{n=\bar{N}}^{\bar{N}}.$$

Each steering vector corresponds to a set of phase shifts on each antenna; therefore, it can be considered as an arm in the MAB problem.

Then the received signal strength after applying a phase shift characterized by the steering vector $\mathbf{f}$ is expressed as Equation (1):

$$r = 30 + 10\log_{10}\left(\left|\beta\mathbf{f}^\top h(\theta)\right|^2\right) + \eta \tag{1}$$

where η a zero-mean Gaussian noise with standard deviation $\sigma = 3.6$ (Samimi & Rappaport, 2014) (in dB). We define the expected signal strength as a function of steering vector $\mathbf{f}$, and receiver's location parameter (θ,β) as

$$R(\mathbf{f},\theta,\beta) := 30 + 10\log_{10}\left(\left|\beta\mathbf{f}^\top h(\theta)\right|^2\right) \tag{2}$$

We remark that the above modeling assumptions (Equations (1) and (2)) have only two unknown free parameters to be learned – channel gain parameter β and angle of departure θ. This makes our method directly applicable to different path loss exponents and carrier frequencies.

3.2. Beam Alignment problem in the MAB Problem Setting

The online learning for beam alignment protocol. For each time step $t = 1, \ldots, T$, the base station chooses beam index $a_t \in [K]$, and receives signal strength feedback $r_t = R(\mathbf{f}_{a_t}, \theta^*, \beta^*) + \eta_t$, where $\eta_t \sim N(0, \sigma^2)$, where θ^* and β^* are the ground truth parameters. We denote the maximum expected return under the parameter θ and β and the expected return as $R^*(\theta, \beta) \coloneqq \max_{a \in [K]} R(\mathbf{f}_a, \theta, \beta)$.

Our goal is to minimize the *regret*, which is the cumulative gap of signal strength between the perfectly aligned and the steering vector selected by the algorithm up to time step T, defined as:

$$\text{Regret}_T \coloneqq \sum_{t=1}^{T} R^*(\theta^*, \beta^*) - R(\mathbf{f}_{a_t}, \theta^*, \beta^*)$$

3.3. Channel Estimation using Maximum Likelihood

Since we do not have access to the ground truth parameters θ^* and β^*, we rely on the Maximum Likelihood Estimation (MLE) $\hat{\theta}, \hat{\beta}$ to approximate these two parameters. We expect the reconstructed reward function $R(\cdot, \hat{\theta}, \hat{\beta})$ to be sufficiently close to $R(\cdot, \theta^*, \beta^*)$ after the estimation. Suppose that we have a dataset of T action-reward pairs $S_T \coloneqq \{a_t, r_t\}_{t=1}^{T}$. We aim to find the estimates $\hat{\theta}_T, \hat{\beta}_T$ by maximizing the likelihood function defined as:

$$\mathcal{L}(\theta, \beta \mid S_T) \coloneqq \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(r_t - R(\mathbf{f}_{a_t}, \theta, \beta))^2}{2\sigma^2}\right).$$

Note that maximizing the likelihood function $\mathcal{L}(\theta, \beta \mid S_T)$ over the parameter space Θ is equivalent to minimizing the square loss:

$$\operatorname*{argmax}_{(\theta,\beta)\in\Theta} \mathcal{L}(\theta, \beta \mid S_T) = \operatorname*{argmin}_{(\theta,\beta)\in\Theta} \sum_{t=1}^{T} (r_t - R(\mathbf{f}_{a_t}, \theta, \beta))^2 \tag{3}$$

Therefore, at each time step, when the historical dataset S_t is updated (e.g., by collecting one more pair (a_{t+1}, r_{t+1}), which updates the dataset to S_{t+1}, we solve the optimization problem (Equation (3)) to obtain the MLE $\hat{\theta}_{t+1}, \hat{\beta}_{t+1}$. We expect that selecting the arm that maximizes $R(\mathbf{f}_a, \hat{\theta}_t, \hat{\beta}_t)$ will minimize the instantaneous regret most effectively.

4. Algorithm

We introduce two versions of physics-informed parametric bandit algorithms: PR-ETC (Algorithm 1) and PR-GREEDY (Algorithm 2), which are the Explore-then-commit (ETC) and Greedy policies for phase retrieval bandit. PR-ETC follows an Explore-Then-Commit (Lattimore & Szepesvári, 2020) style via randomly selecting arms and collecting feedback to estimate θ, β. PR-GREEDY focuses on exploitation and takes action greedily in every time step, which can still perform well in some contextual bandit problems as previously shown in Kannan et al. (2018).

For both algorithms, we use MLE to estimate (θ^*, β^*) by solving optimization problem (3), and choose the action that maximizes the predicted expected reward.

Algorithm 1 PR-ETC

Input: $K \geq 2$, the configuration matrix F, the initial θ_0 and β_0, exploration parameter M.
for $t = 1, 2, \cdots, T$ **do**
  if $t < M$ **then**
    Pull arm $a_t \in [K]$ uniform randomly
    Receive reward r_t
  else if $t = M$ **then**
    Estimate parameters $\hat{\theta}_T, \hat{\beta}_T$ by solving Equation (3).
    Pull arm $a_t := \operatorname{argmax}_{a \in [K]} R(\mathbf{f}_a, \hat{\theta}_t, \hat{\beta}_t)$
  else
    Pull arm $a_t = a_M$
  end if
end for

Algorithm 2 PR-GREEDY

Input: $K \geq 2$, the configuration matrix F, the initial θ_0 and β_0,
for $t = 1, 2, \cdots, T$ **do**
  Estimate parameters $\hat{\theta}_t, \hat{\beta}_t$ by solving Equation (3).
  Pull arm $a_t := \operatorname{argmax}_{a \in [K]} R(\mathbf{f}_a, \hat{\theta}_t, \hat{\beta}_t)$
  Receive reward r_t
end for

Note that PR-ETC only solves the optimization problem once, which is at the last time step M of the exploration phase, and we need to balance exploration and exploitation by choosing M appropriately, while PR-GREEDY solves the optimization problem T times. The computation cost of PR-ETC will be much smaller than PR-GREEDY. Though both algorithms collect T samples, we found the empirical performance of PR-GREEDY is also better than PR-ETC by a margin.

5. Experiment

In this section, we evaluate PR-ETC and PR-GREEDY in environments with complex channel strength reward functions. To this end, we utilize the DeepMIMO (Alkhateeb, 2019), which is a generic ray-tracing-based dataset for benchmarking beam alignment algorithms. With DeepMIMO, we can vary the User Equipment (UE) location,

allowing us to evaluate the algorithm across a wide range of locations.

For our experiment, we utilize scenario 4 from an area in Phoenix, which includes 3 base stations and 6,794 users. For each combination of base station and user, we evaluate all algorithms' performance. In total, we have 20,382 transmitter-receiver pairs (i.e., tasks), including 4,952 pairs with a channel connection. For each transmitter-receiver pair, we construct a bandit instance and run every algorithm, recording the arm chosen and the returned reward. Since our analysis focuses on the ULA antenna, we configure 16 antennas in an array for the base station and set 1 antenna for each user. The frequency of the mmWave in this experiment is 28GHz, thus, the wavelength is $\lambda = \frac{c}{f} = 0.011$ meters. The spacing of the antenna is $d = 0.5\lambda = 0.005$. The size of the beam codebook is $K = 180$. We set the parameter space $\Theta = \left\{(\theta, \beta) : \theta \in [0, 360],\ \beta \in [10^{-50}, 1]\right\}$. Currently, we use a grid search method to solve the optimization problem (Equation (3)). We divide the parameter space Θ by splitting β into 50 grids and θ into 180 grids, then search for the optimal pair across all grid points. Because the objective function (Equation (3)) is non-convex, a coarse grid may cause the algorithm to miss the global optimum.

5.1. Evaluation metric: normalized regret

Since we evaluate across 4,952 tasks, to make the regrets of all tasks comparable, we shift and scale the regret of all tasks to a common range. For each bandit instance, we rescale the regret per step as follows:

$$\text{regret}_t = \frac{R^*(\theta^*, \beta^*) - R(\mathbf{f}_{a_t}, \theta^*, \beta^*)}{R^*(\theta^*, \beta^*) - \mathbb{E}_{a\sim U[K]}[R(\mathbf{f}_a, \theta^*, \beta^*)]}$$

where $a \sim U[K]$ means that a is drawn from the uniform distribution over $[K]$. The total cumulative regret becomes

$$\text{N-Regret}_T = \frac{1}{T}\sum_{t=1}^{T} \text{regret}_t$$

With this normalization, the normalized regret for any reasonable algorithm better than random guessing is likely to be $[0, 1]$, with the best action still having zero regret.

We compare our PR-ETC and PR-GREEDY algorithms with several baseline MAB algorithms.

- UCB (Auer et al., 2002): a basic version of the Upper Confidence Bound algorithm, where the confidence bound is given by $\sqrt{2\ln(T)/N_{t,a}}$ with $N_{t,a}$ representing the number of arm a has being pulled up to time t.
- LSE (Yu & Mannor, 2011): Line Search Elimination Algorithm that selects the action with the highest expected reward by leveraging the unimodal property.
- BISECTION (Lattimore, 2024): the noisy bisection-based algorithm.
- IMED-MB (Saber & Maillard, 2024), where we set the number of assumed peaks in the reward function to be $M' = 10$.

In addition, for PR-ETC, we set M, the number of random exploration rounds, to be 20.

5.2. Performance on the synthetic dataset

5.2.1. DISTRIBUTION OF CUMULATIVE REGRET

Figures 3 and 4 show the distribution of N-Regret$_T$ over 4,952 environments for time horizons $T = 50$ and 200, respectively. For each environment, the N-Regret$_T$ is computed by averaging over 10 runs.

We find that PR-GREEDY consistently outperforms other algorithms for a time horizon of $T = 50$. At time step 50, the expected mean regret of PR-GREEDY is 0.47, while PR-ETC has a mean regret of 0.64 (Figure 3). Increasing the time horizon from 50 to 200 significantly improves the performance of both PR-GREEDY and PR-ETC (Figure 4), although PR-GREEDY requires more time to solve the optimization problem. This improvement occurs because, after the initial 50 steps, PR-GREEDY can identify a near-optimal beam with high probability, resulting in near-zero regret over the subsequent 150 steps. We show this in the next section (Section 5.2.2).

For the other unimodal bandit algorithms, we empirically observe that they do not perform as well as our physics-informed algorithms. LSE's performance improves as the horizon T increases from 50 to 200, approaching that of PR-ETC, but it still underperforms compared to PR-GREEDY. BISECTION, designed for unimodal settings, struggles in both the 50-step and 200-step horizons, as it is not well-suited for the multimodal nature of the problem. Although IMED-MB is designed for multimodal settings, it exhibits a bimodal distribution in the histogram plots. The experiment shows that IMED-MB struggles to select the best beam with only a very small number of observations. This behavior occurs because they can identify the best beam in some environments, but in others, their overall performance is similar to a random policy. UCB seems to have a clearly inferior performance due to its need in choosing every beam once to begin with; with the total number of beams being 180 and a time horizon of 50 or 200, it performs almost identical to a random policy.

Due to space limitations, we present in Appendix A.3.1 the histograms of the normalized simple regret across dif-

ferent algorithms. The results show a similar trend, with PR-GREEDY generally achieving lower normalized simple regret.

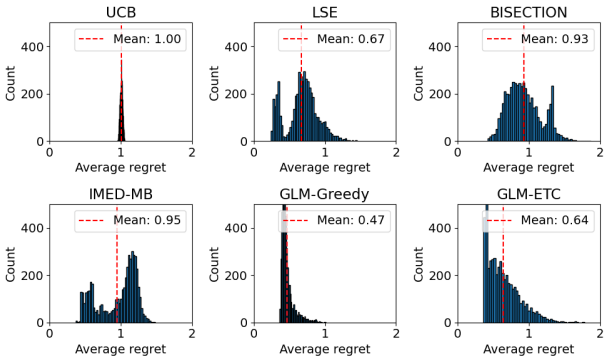

Figure 3: Distribution of N-Regret$_T$ over all transmitter-receiver pairs at the first 50-th step

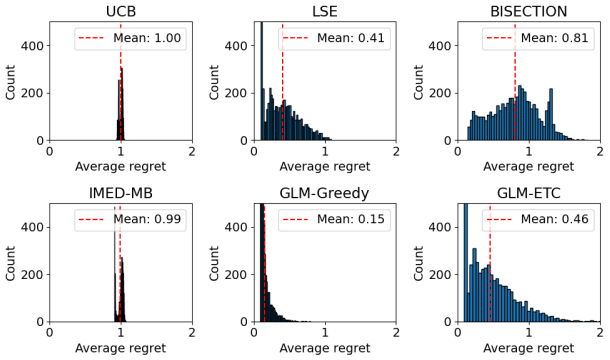

Figure 4: Distribution of N-Regret$_T$ over all transmitter-receiver pairs at the first 200-th step

5.2.2. REWARD FUNCTION AND ITS ESTIMATION

To further understand the behavior of the algorithms and baselines, we select one bandit instance from all user locations as an example (Figures 5 and 6). Both PR-GREEDY and PR-ETC achieve lower average regret, which continues to decrease as time progresses and more data is collected, and PR-GREEDY performs better than PR-ETC. BISECTION and LSE learn some patterns, but their performance remains inferior compared to our methods. UCB and IMED-MB perform the worst, with UCB particularly affected by the insufficient number of samples to effectively shrink the confidence interval—on average, we have no more than two samples per arm.

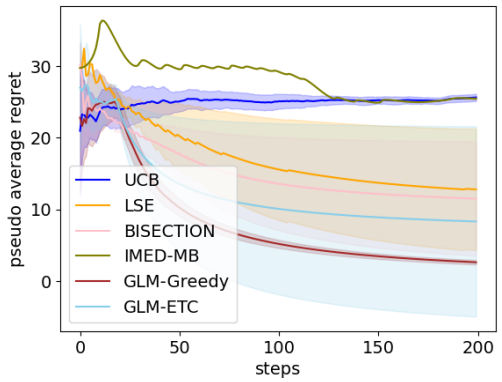

Figure 5: Average regret vs time step for different algorithms. Because of the finite-horizon setting with $T = 200$ steps and $K = 180$ arms, UCB and IMED-MB have a flat regret from the initial uniform exploration of each arm.

We also plot in Figure 6 the reward function and estimated reward function rebuilt by using $\hat{\theta}_T$ and $\hat{\beta}_T$ by PR-GREEDY and PR-ETC. From the plot, we can see that the beams chosen by PR-GREEDY are centered around the peak of the reward function rather than uniformly distributed on the action space.

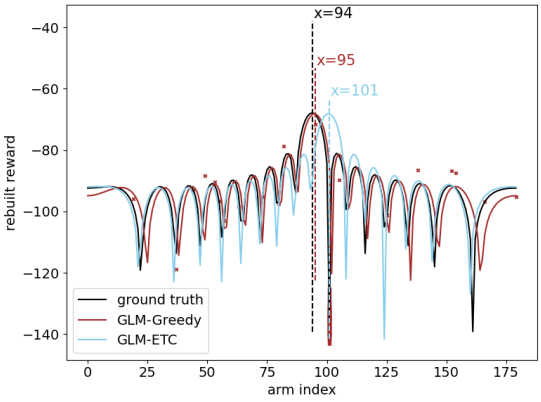

Figure 6: Reward function estimated by PR-GREEDY and PR-ETC at time step 200. Each cross represents the arm selected by PR-GREEDY and PR-ETC. For each color, the vertical lines represent the best beam that maximizes the respective reward function.

5.2.3. THE REGRET IN THE MAP

We also plot the spatial heatmap of the N-Regret$_T$ in Figures 7 to 9. Additional comparisons with other algorithms are provided in Figure 11 in Appendix A.2. We observe that the performance of PR-GREEDY aligns well with the histogram distribution, indicating generally strong perfor-

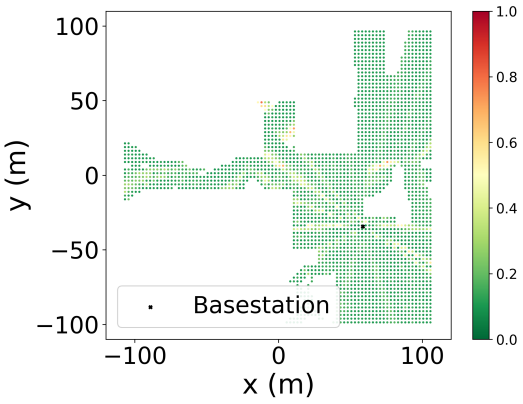

Figure 7: Spatial heat map of N-Regret$_T$ for PR-GREEDY across the coverage area of Base Station 1

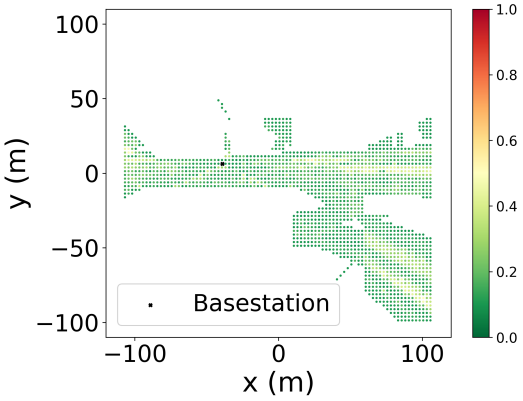

Figure 8: Spatial heat map of N-Regret$_T$ for PR-GREEDY across the coverage area of Base Station 2

mance. Only a few areas on the map exhibit relatively poorer results, which we suspect stem from computational precision limitations.

6. Conclusion

We develop the physics-informed parametric bandit algorithms, PR-ETC and PR-GREEDY, to address the beam alignment problem in millimeter-wave communications. Unlike other works that rely on unimodality or multimodality assumptions on the reward function, our approach is grounded in the fundamental model of far-field mmWave propagation with one dominant path. Our algorithms demonstrate good performance on synthetic datasets compared to existing multi-armed bandit-based methods. However, we observe that the computational cost of solving the optimization problem remains non-negligible (See Appendix A.1 for detailed comparison). There is significant potential to extend this work. A natural next step is to consider dynamic environments where the user is moving, as explored in Aykin et al. (2020). Additionally, real-world experiments will be important to validate the practical effectiveness and robustness of our algorithms under realistic conditions.

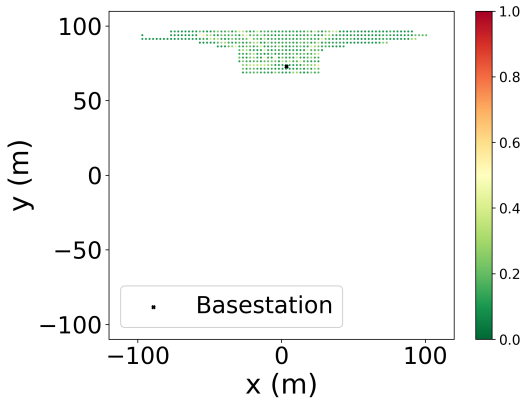

Figure 9: Spatial heat map of N-Regret$_T$ for PR-GREEDY across the coverage area of Base Station 3

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning and Wireless Communications. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Alkhateeb, A. Deepmimo: A generic deep learning dataset for millimeter wave and massive mimo applications. *arXiv preprint arXiv:1902.06435*, 2019.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

Aykin, I., Akgun, B., Feng, M., and Krunz, M. Mamba: A multi-armed bandit framework for beam tracking in millimeter-wave systems. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, pp. 1469–1478. IEEE, 2020.

Cutkosky, A., Das, A., Kong, W., Lee, C., and Sen, R. Black-box optimization of unimodal functions. In *Uncertainty in Artificial Intelligence*, pp. 476–484. PMLR, 2023.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. Parametric bandits: The generalized linear case. *Advances in neural information processing systems*, 23, 2010.

Ghosh, D., Hanawal, M. K., and Zlatanov, N. Ub3: Fixed budget best beam identification in mmwave massive miso

via pure exploration unimodal bandits. *IEEE Transactions on Wireless Communications*, 2024.

Gustafson, C., Haneda, K., Wyne, S., and Tufvesson, F. On mm-wave multipath clustering and channel modeling. *IEEE transactions on antennas and propagation*, 62(3): 1445–1455, 2013.

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

Hassanieh, H., Abari, O., Rodriguez, M., Abdelghany, M., Katabi, D., and Indyk, P. Fast millimeter wave beam alignment. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, pp. 432–445, 2018.

Honda, J. and Takemura, A. Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards. *J. Mach. Learn. Res.*, 16:3721–3756, 2015.

Kannan, S., Morgenstern, J. H., Roth, A., Waggoner, B., and Wu, Z. S. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. *Advances in neural information processing systems*, 31, 2018.

Kotłowski, W. and Neu, G. Bandit principal component analysis. In *Conference On Learning Theory*, pp. 1994–2024. PMLR, 2019.

Lai, T. L. Adaptive treatment allocation and the multi-armed bandit problem. *The annals of statistics*, pp. 1091–1114, 1987.

Langford, J. and Zhang, T. The epoch-greedy algorithm for multi-armed bandits with side information. *Advances in neural information processing systems*, 20, 2007.

Lattimore, T. Bandit convex optimisation. *arXiv preprint arXiv:2402.06535*, 2024.

Lattimore, T. and Hao, B. Bandit phase retrieval. *Advances in Neural Information Processing Systems*, 34:18801–18811, 2021.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Liu, Y., Wang, Z., Xu, J., Ouyang, C., Mu, X., and Schober, R. Near-field communications: A tutorial review. *IEEE Open Journal of the Communications Society*, 4:1999–2049, 2023.

Mazaheri, M. H., Ameli, S., Abedi, A., and Abari, O. A millimeter wave network for billions of things. In *Proceedings of the ACM Special Interest Group on Data Communication*, pp. 174–186. 2019.

Nitsche, T., Flores, A. B., Knightly, E. W., and Widmer, J. Steering with eyes closed: mm-wave beam steering without in-band measurement. In *2015 IEEE Conference on Computer Communications (INFOCOM)*, pp. 2416–2424. IEEE, 2015.

Rappaport, T. S., Xing, Y., MacCartney, G. R., Molisch, A. F., Mellios, E., and Zhang, J. Overview of millimeter wave communications for fifth-generation (5g) wireless networks—with a focus on propagation models. *IEEE Transactions on antennas and propagation*, 65(12):6213–6230, 2017.

Saber, H. and Maillard, O.-A. Bandits with multimodal structure. *Reinforcement Learning Journal*, 5:2400–2439, 2024.

Samimi, M. K. and Rappaport, T. S. Characterization of the 28 ghz millimeter-wave dense urban channel for future 5g mobile cellular. *NYU Wireless TR*, 1:1–322, 2014.

Yu, J. Y. and Mannor, S. Unimodal bandits. In *ICML*, pp. 41–48, 2011.

A. Appendix

We include additional experimental results of the bandit algorithms to provide a more comprehensive evaluation. In Appendix A.1, we present the computational time required by each algorithm when running for 200 steps on the synthetic dataset. In Appendix A.2, we provide regret maps for other bandit algorithms. In Appendix A.3, we measure the performance of all algorithms using the normalized simple regret instead of N-Regret$_T$ This analysis highlights the performance of each algorithm after a fixed number of time steps, effectively removing the influence of the initial training phase.

A.1. Time cost

We compare the computational time required by each algorithm for a time horizon of $T = 200$ on the DeepMIMO dataset. Notably, PR-GREEDY has the highest computational cost, requiring approximately 75,346 ms per transmitter-receiver pair . , which is several orders of magnitude greater than most other methods. This is primarily due to the grid search approach used in our implementation to solve the non-convex optimization problem (Equation (3)), as built-in solvers have proven unreliable for this task. Similarly, PR-ETC is less expensive, taking about 1,581 ms, since it only solves the optimization problem once during the whole training phase.

In contrast, classical algorithms such as UCB, LSE, and BISECTION have very low computational overhead, each completing within 3 ms, making them highly efficient, though potentially less accurate in complex environments. IMED-MB exhibits a moderate cost of around 100 ms.

We leave it as an important open problem to develop a faster and more reliable method to solve the optimization problem (Equation (3)) without sacrificing accuracy.

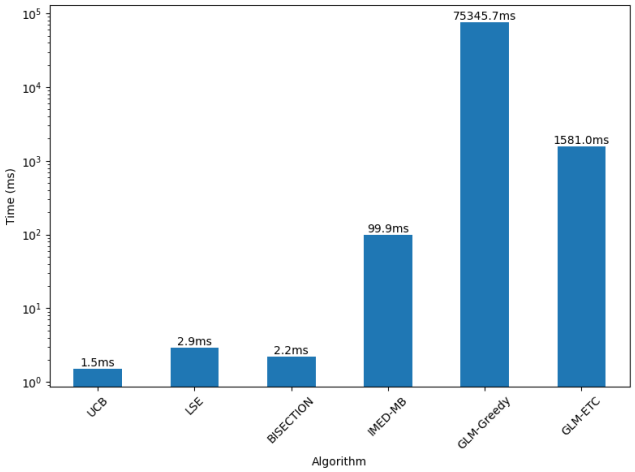

Figure 10: Time cost of each algorithm, averaged over 40 different transmitter-receiver pairs with 5 repeats.

A.2. Spatial heat maps of N-Regret$_T$

We include the spatial heat maps of N-Regret$_T$ of other algorithms over the synthetic dataset in Figure 11. From the figures, it is clear that PR-GREEDY achieves the best overall performance, followed by PR-ETC and LSE.

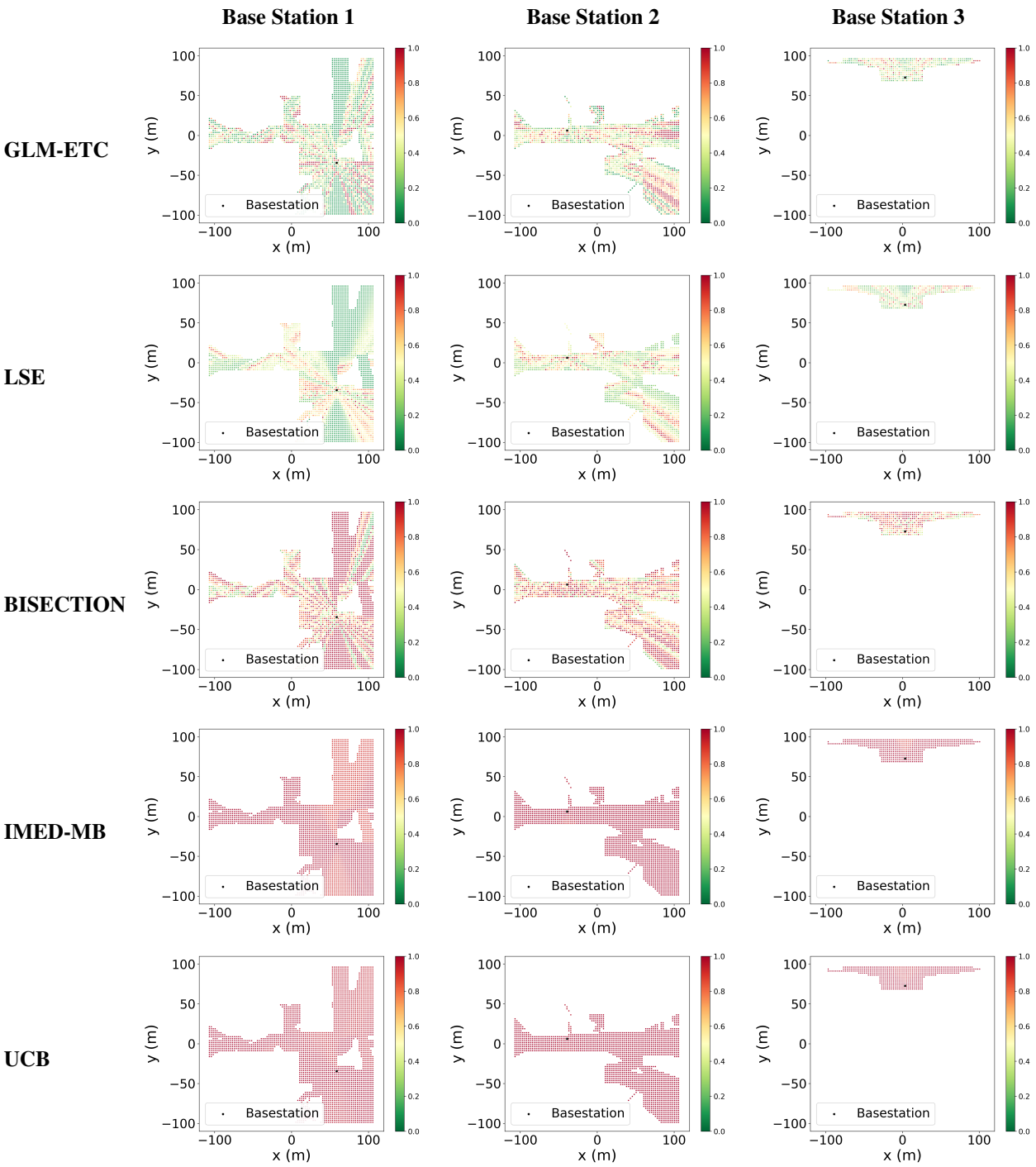

Figure 11: Spatial heat maps of N-Regret$_T$ of different algorithms (rows) across Base Stations 1–3 (columns).

A.3. Experiment result measured by simple regret

We present experimental results using *simple regret*, which is defined as the reward gap at the final time step, which is the performance of the action committed by the algorithm after the training phase. Specifically, the *simple regret* is given by

$$\text{regret}_T := R^*(\theta^*, \beta^*) - R(\mathbf{f}_{a_T}, \theta^*, \beta^*). \tag{4}$$

and we apply the same normalization as described in Section 5.1. To remind us, the normalized simple regret is defined as

$$\text{regret}_t = \frac{R^*(\theta^*, \beta^*) - R(\mathbf{f}_{a_t}, \theta^*, \beta^*)}{R^*(\theta^*, \beta^*) - \mathbb{E}_{a \sim U[K]}[R(\mathbf{f}_a, \theta^*, \beta^*)]}$$

where $a \sim U[K]$ means that a is drawn from the uniform distribution over $[K]$. As a result, the regrets are mostly within the range $[0, 1]$, with the optimal policy yielding a regret of 0, and a random policy having a regret of 1.

In Appendix A.3.1, we show the distribution of normalized simple regret to illustrate the performance across different BS-user location pairs. Additionally, in Appendix A.3.2 we include spatial heatmaps from three base stations for our algorithms as well as our baselines for comparison.

A.3.1. Simple regret distribution

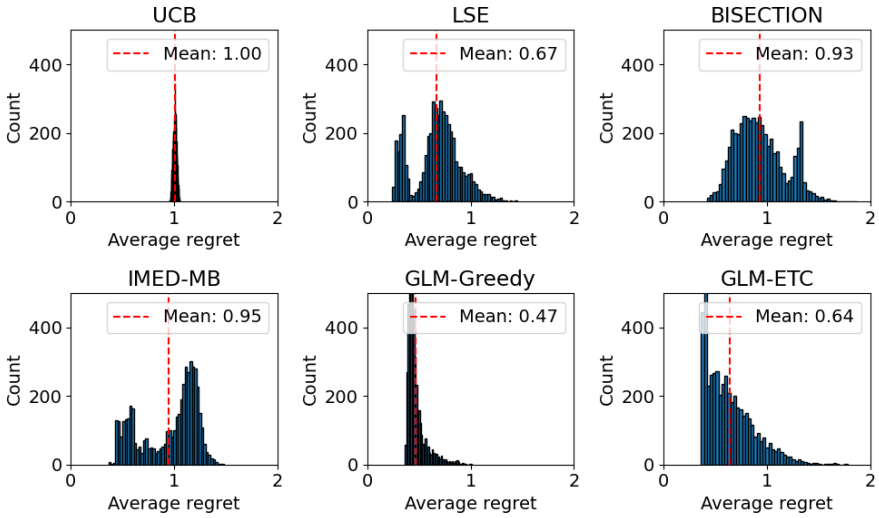

Figure 12: Simple regret at the 50-th step

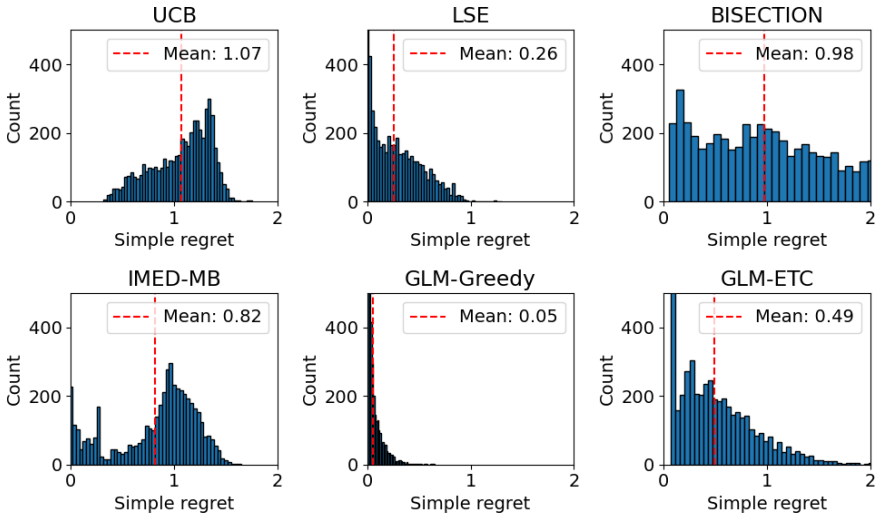

Figure 13: Simple regret at the 200-th step

A.3.2. SIMPLE REGRET MAPS

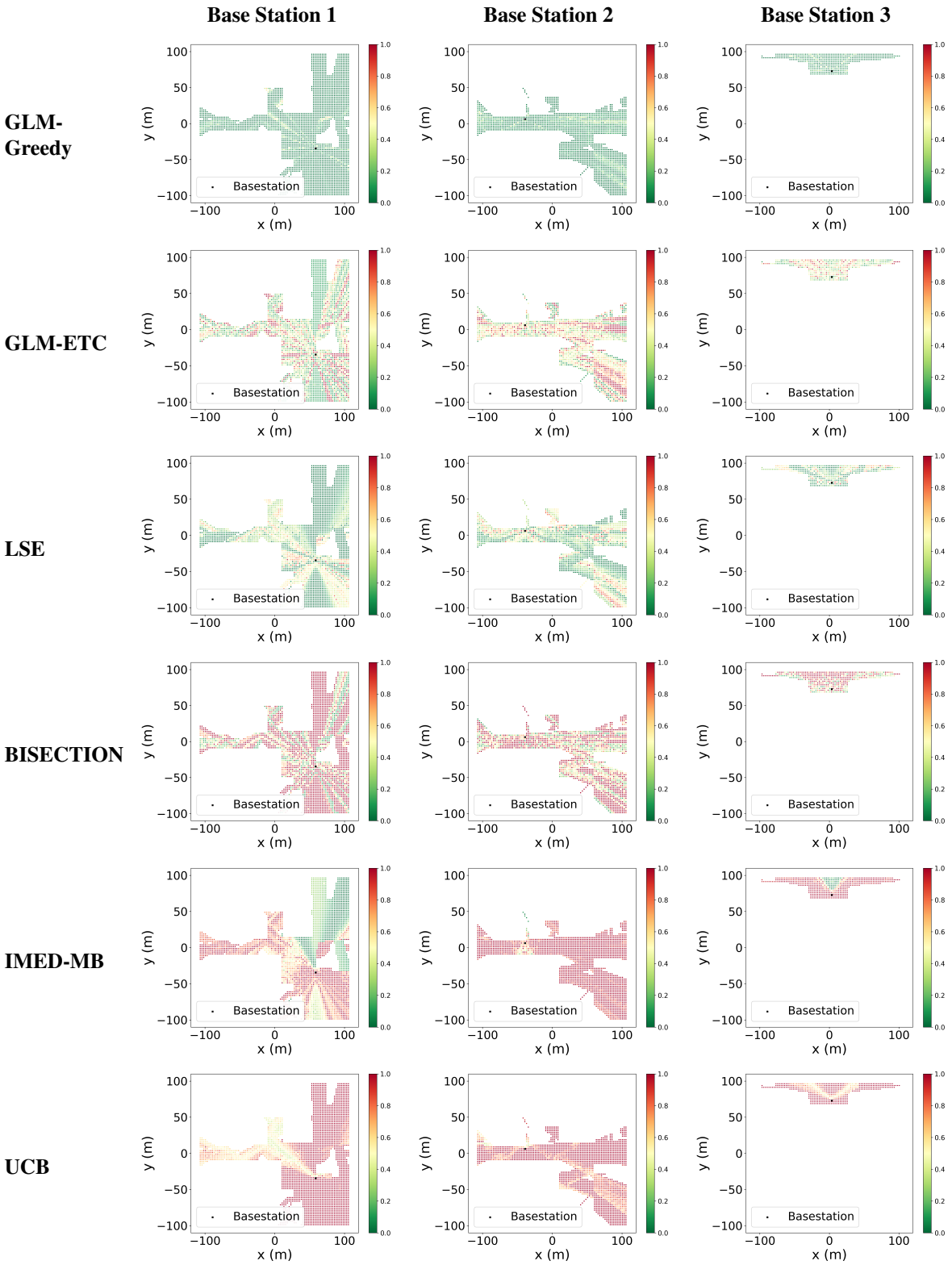

Figure 14: Spatial heat maps of the normalized simple regret of different algorithms (rows) across Base Stations 1–3 (columns).