Physics-Informed Parametric Bandits for Beam Alignment in mmWave Communications

Hao Qin* Thang Duong* Ming F. Li Chicheng Zhang

University of Arizona

*Equal contribution

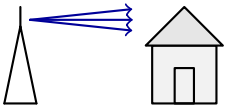
WiOpt 2026

THE UNIVERSITY OF ARIZONA

mmWave beam alignment: why it matters

mmWave (30 to 300 GHz): the spectrum for next-gen 5G/6G links

28 GHz



5G Fixed Wireless Access

Verizon, T-Mobile

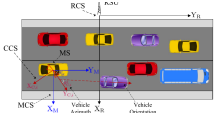
(Qualcomm Dragonwing '25)

28 GHz

XR & cloud gaming

Qualcomm *Boundless XR*

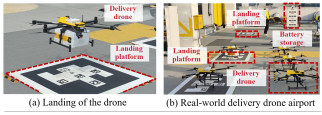
28 GHz



V2X / autonomous

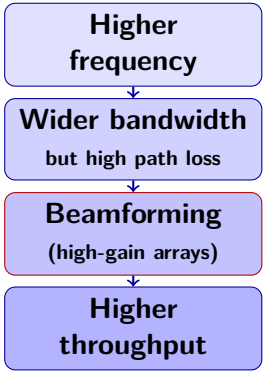
Tan et al., IEEE ComSurv 2024

77 GHz



Drone landing & sensing

mmE-Loc, SenSys '25



Due to high path loss in the mmWave band, we need beamforming.

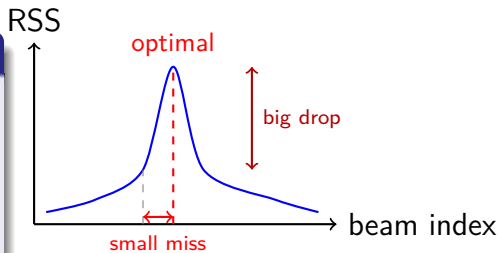
Cost of misalignment

Narrow beams, narrow tolerance

beam only 7° wide $\Longrightarrow$

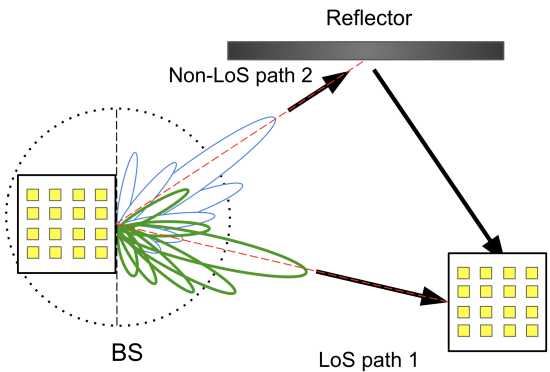
a **small** pointing error = **17 dB drop**

(Nitsche et al. 2015, 7° beamwidth, 18° off)



A tiny horizontal miss is a large vertical loss.

System model



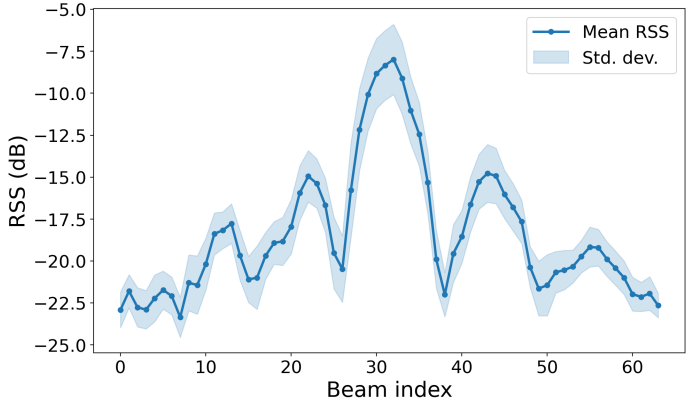
- BS with an antenna array, 1 RF chain
- UE with a (quasi-)omnidirectional antenna; still or mobile
- BS picks a beam
- Signal arrives via a few paths (LoS + reflections)

Goal: pick the beam that maximizes RSS.

Shown with 2 paths for clarity; the model and our method handle any number of paths k.

Challenges

1. RSS across beams is noisy and multimodal.
Sidelobes and multipath make unimodal methods miss the best beam.



Real RSS (DeepSense 6G, scenario 17): noisy, multimodal.

Challenges

1. RSS across beams is noisy and multimodal.
Sidelobes and multipath make unimodal methods miss the best beam.

2. Sample efficiency: converge in few rounds.
The optimal beam stays stable only for a limited time, algorithm must converge to a good beam within beam shifts window (300 ms in practice) in few rounds.

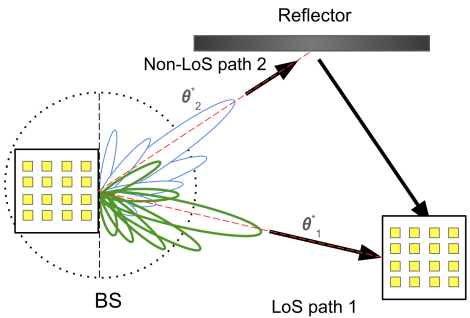
Challenges

1. RSS across beams is noisy and multimodal.
Sidelobes and multipath make unimodal methods miss the best beam.

2. Sample efficiency: converge in few rounds.
The optimal beam stays stable only for a limited time, algorithm must converge to a good beam within beam shifts window (300 ms in practice) in few rounds.

3. Low computational latency.
Per-step computation must finish within one packet/frame interval (10 ms in 5G NR), so it keeps up with the transmission pipeline.

Channel and reward model

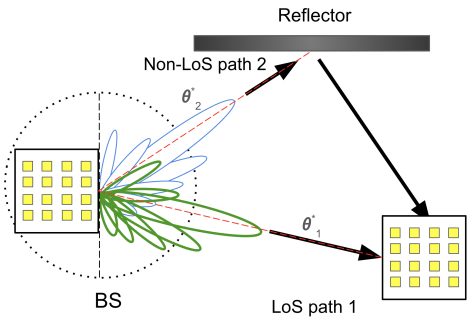


Beam pattern of antenna config $\mathbf{f}_a$:

$$h_a(\theta) = \mathbf{f}_a^\top v(\theta)$$

$v(\theta)$: array response vector at AoD θ

Channel and reward model



Beam pattern of antenna config $\mathbf{f}_a$:

$$h_a(\theta) \;=\; \mathbf{f}_a^\top v(\theta)$$

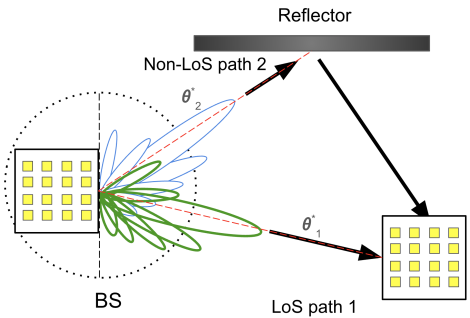
$v(\theta)$: array response vector at AoD θ

Linear channel with k paths:

$$h_a(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*) \;=\; \sum_{i=1}^{k} \beta_i^*\, h_a\left(\theta_i^*\right)$$

β_i^*: complex gain of path i; θ_i^*: AoD of path i

Channel and reward model



Beam pattern of antenna config $\mathbf{f}_a$:

$$h_a(\theta) \;=\; \mathbf{f}_a^\top v(\theta)$$

$v(\theta)$: array response vector at AoD θ

Linear channel with k paths:

$$h_a(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*) \;=\; \sum_{i=1}^{k} \beta_i^*\, h_a\left(\theta_i^*\right)$$

β_i^*: complex gain of path i; θ_i^*: AoD of path i

Observed RSS (in dB) with Gaussian noise η:

$$r \;=\; 30 + 10\log_{10}\left|h_a(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*)\right|^2 + \eta$$

Online learning protocol & regret

Per round $t = 1, \ldots, T$**:**

- BS picks beam $a_t \in [K]$
- Observes noisy RSS r_t (dB)

Cumulative regret:

$$\mathrm{Regret}_T \;=\; \sum_{t=1}^{T} \Big(\max_{a} R(\mathbf{f}_a, \boldsymbol{\theta}^*, \boldsymbol{\beta}^*) \;-\; R(\mathbf{f}_{a_t}, \boldsymbol{\theta}^*, \boldsymbol{\beta}^*) \Big)$$

We want regret bounded by k, the number of paths, not K, the number of beams.

Estimating RSS: maximum likelihood

Given data $S_m = \{(a_t, r_t)\}_{t=1}^m$, fit the channel by least squares:

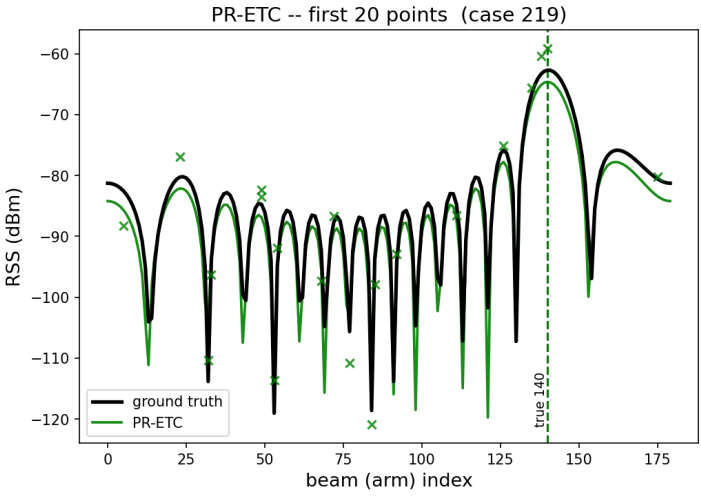
$$(\hat{\boldsymbol{\theta}}, \hat{\beta}) \;=\; \arg\min_{\boldsymbol{\theta},\beta} \sum_{t=1}^{m} \Big(r_t - R(\mathbf{f}_{a_t}, \boldsymbol{\theta}, \beta) \Big)^2$$

Only $2k$ unknowns $(\hat{\boldsymbol{\theta}}, \hat{\beta})$ to recover — independent of the number of beams K.

PR-ETC: $M = 20$

1. **Explore.** Pick random beams for M rounds.
2. **Estimate.** Fit the channel *once* by MLE.
3. **Commit.** Pick the best predicted beam thereafter.

One hyperparameter (M); MLE solved once.



PR-Greedy: $t = 5 \to 10 \to 15 \to 20$

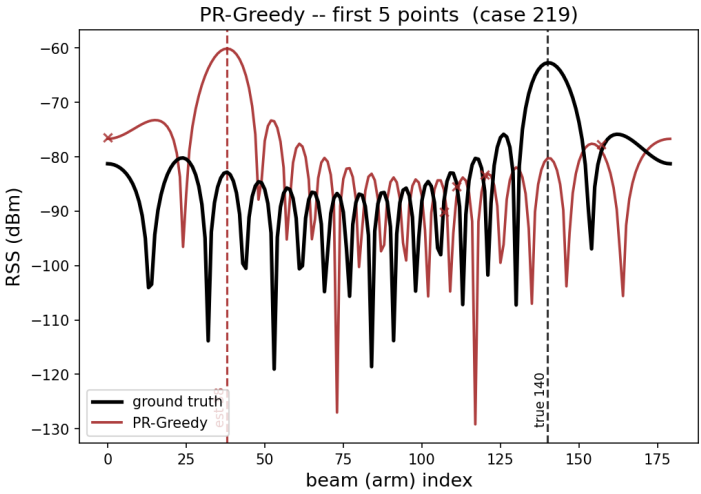
0. **Init.** Start from any channel guess.

1. **Greedy.** Pick the best predicted beam.

2. **Observe.** Get noisy RSS.

3. **Re-fit.** Update MLE on all samples so far.

Hyperparameter-free (only k); MLE rerun every step.



PR-Greedy: $t = 5 \to 10 \to 15 \to 20$

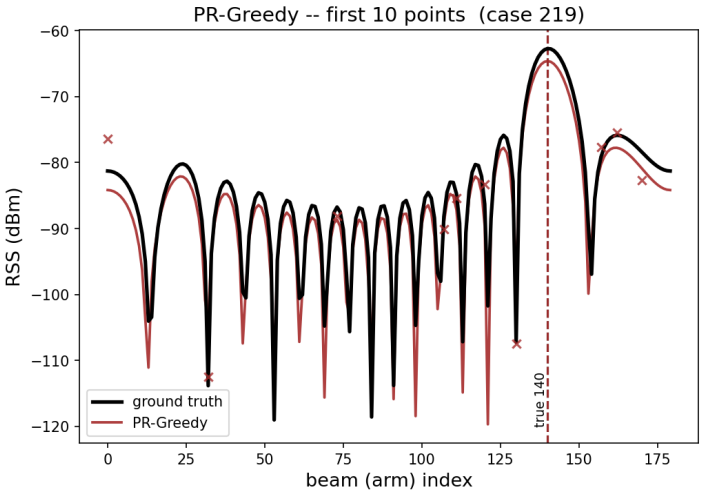
0. **Init.** Start from any channel guess.

1. **Greedy.** Pick the best predicted beam.

2. **Observe.** Get noisy RSS.

3. **Re-fit.** Update MLE on all samples so far.

Hyperparameter-free (only k); MLE rerun every step.



PR-Greedy: $t = 5 \to 10 \to 15 \to 20$

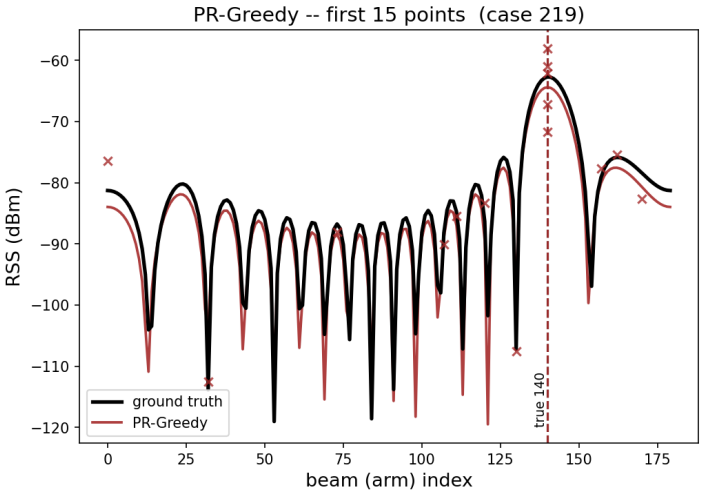
0. **Init.** Start from any channel guess.

1. **Greedy.** Pick the best predicted beam.

2. **Observe.** Get noisy RSS.

3. **Re-fit.** Update MLE on all samples so far.

Hyperparameter-free (only k); MLE rerun every step.



PR-Greedy: $t = 5 \to 10 \to 15 \to 20$

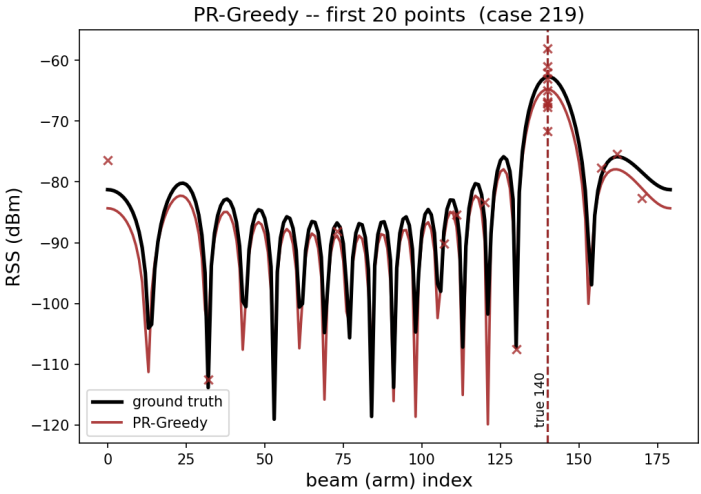
0. **Init.** Start from any channel guess.

1. **Greedy.** Pick the best predicted beam.

2. **Observe.** Get noisy RSS.

3. **Re-fit.** Update MLE on all samples so far.

Hyperparameter-free (only k); MLE rerun every step.



PR-ETC: regret independent of K

Informal guarantee. Under three regularity assumptions on $R(\cdot)$ (bounded, locally Lipschitz, strongly-convex MSE landscape), with M tuned appropriately:

$$\mathrm{Regret}_T \;=\; \tilde{O}\Big(T^{2/3}\,k^{1/3}\Big)\,.$$

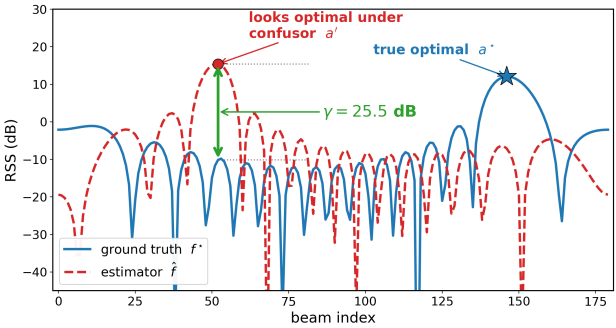
- Standard ETC: $\tilde{O}(T^{2/3}K^{1/3})$; UCB: $\tilde{O}(\sqrt{KT})$
- Both vacuous when $K \approx T$; PR-ETC and PR-Greedy replace K with $k \ll K$

Formal guarantees can be found in the paper.

PR-Greedy: identifiability gives regret bounds

γ-self-identifiability (Slivkins 2025). Any wrong $(\boldsymbol{\theta}, \boldsymbol{\beta})$ makes difference at its (suboptimal) argmax beam by $\geq \gamma$ dB.

Informal guarantee. Each suboptimal beam is selected at most $\tilde{O}\left(\frac{k}{\gamma^2}\right)$ times.



Verified on DeepMIMO: empirically $\gamma \approx 0.07$ dB.

Formal guarantees can be found in the paper.

Slivkins, Xu, Zuo. "Greedy Algorithm for Structured Bandits: A Sharp Characterization of Asymptotic Success/Failure." 2025.

Experimental setup

Datasets

- **DeepMIMO** (simulated, ray tracing)
  4,952 BS-UE pairs, 28 GHz, $K = 180$
- **DeepSense 6G** (real measurements)
  12 outdoor scenarios, 60 GHz, $K = 64$

Horizon

- $T = 200$

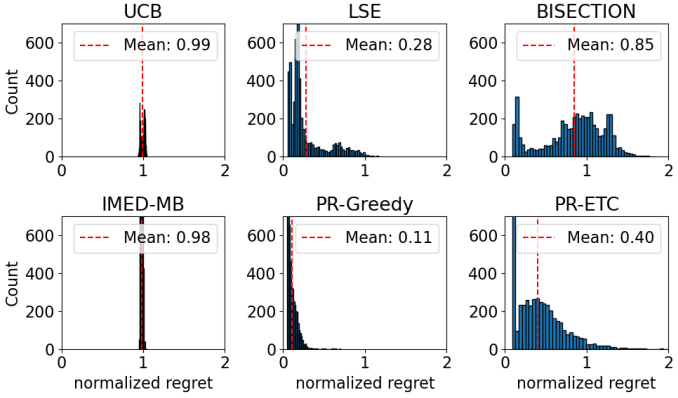
Baselines

- UCB (classical)
- LSE (unimodal)
- BISECTION (unimodal)
- IMED-MB (multimodal, 10 peaks)

Metric: normalized regret $= \mathrm{Regret}_T /$ regret of random policy.
0: optimal, 1: uniform random, > 1: worse than random.

DeepMIMO: Alkhateeb, ITA 2019.
DeepSense 6G: Alkhateeb et al., IEEE Commun. Mag. 2023.

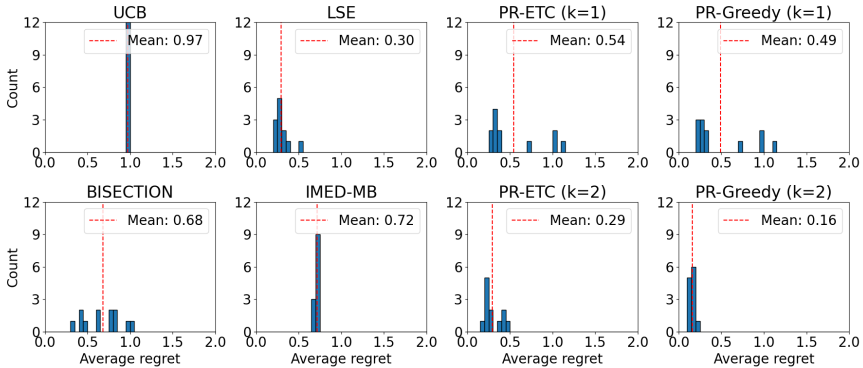
Experiment 1: DeepMIMO simulated environments



Distribution of normalized regret at $T = 200$ across 4,952 BS-UE pairs.

PR-Greedy achieves the lowest regret.

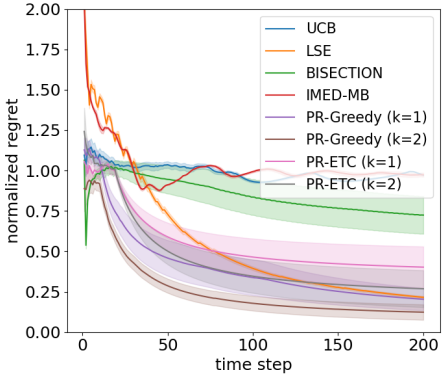
Experiment 2: DeepSense 6G real-world datasets



Distribution of normalized regret at $T = 200$ across 12 outdoor scenarios.

PR-Greedy achieves the lowest regret on real RSS traces.

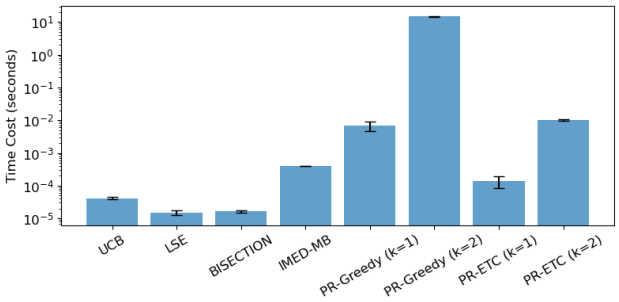
Experiment 2: learning curve (DeepSense 6G scenario 24)



Per-step regret, averaged over 10 trials; shaded bands $= \pm 1$ s.e.

PR-Greedy ($k = 2$) and PR-ETC ($k = 2$) reach low regret early.

Computational cost

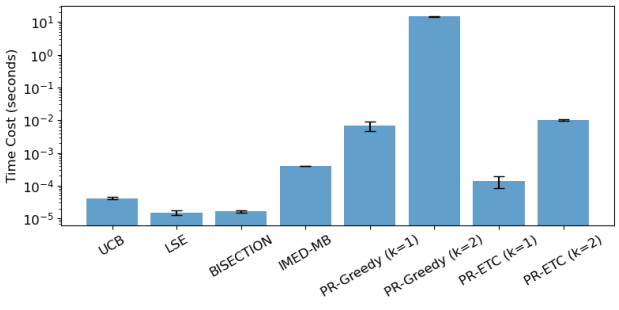


Per-step time, DeepSense 6G ($T = 200$)

Per step on a CPU (mean)

- PR-GREEDY: $\sim$ 377 ms (MLE every step)
- PR-ETC: $\sim$ 8 ms (MLE once)
- UCB / LSE / BISECTION: < 1 ms

Computational cost



Per-step time, DeepSense 6G ($T = 200$)

Per step on a CPU (mean)

- PR-GREEDY: $\sim$ 377 ms (MLE every step)
- PR-ETC: $\sim$ 8 ms (MLE once)
- UCB / LSE / BISECTION: < 1 ms

Beam-update intervals are 160 to 310 ms (Feng et al. 2025); PR-ETC is the safe choice when latency matters.

Takeaway

- PR-ETC: $\tilde{O}(T^{2/3}k^{1/3})$ regret, independent of number of beams
- PR-GREEDY: lowest empirical regret on DeepMIMO and DeepSense 6G
- Robust to misspecified k

Takeaway

- PR-ETC: $\tilde{O}(T^{2/3}k^{1/3})$ regret, independent of number of beams
- PR-GREEDY: lowest empirical regret on DeepMIMO and DeepSense 6G
- Robust to misspecified k

Open questions

- Faster MLE (compressed sensing, warm starts)
- Near-field channels
- Non-stationary extension (mobility, blockage)

Takeaway

- PR-ETC: $\tilde{O}(T^{2/3}k^{1/3})$ regret, independent of number of beams
- PR-GREEDY: lowest empirical regret on DeepMIMO and DeepSense 6G
- Robust to misspecified k

Open questions

- Faster MLE (compressed sensing, warm starts)
- Near-field channels
- Non-stationary extension (mobility, blockage)

Thang Duong

Hao Qin

Thank you!

arXiv

Backup: formal guarantee for PR-ETC

Theorem (formal)

Under Assumptions 1, 2, and 3,

$$\mathrm{Regret}_T \;\le\; \tilde{O}\Big(R_{\max}\, T^{2/3} \left(k\sigma^2 (\log|B| + \log|\Theta|) \right)^{1/3} \Big)$$

with $M = T^{2/3}(k\sigma^2(\log|B| + \log|\Theta|))^{1/3}$.

Assumptions.

1. **Bounded reward.** $R \in [0, R_{\max}]$.
2. **Identifiability.** Distinct (θ, β) induce distinguishable expected rewards (MSE sense), so exploration can recover them.
3. **Local Lipschitzness.** R is locally Lipschitz in $(\theta, \log\beta)$: small estimation error $\Rightarrow$ small reward gap.

Backup: formal guarantee for PR-Greedy

γ-self-identifiability (Slivkins 2025, adapted)

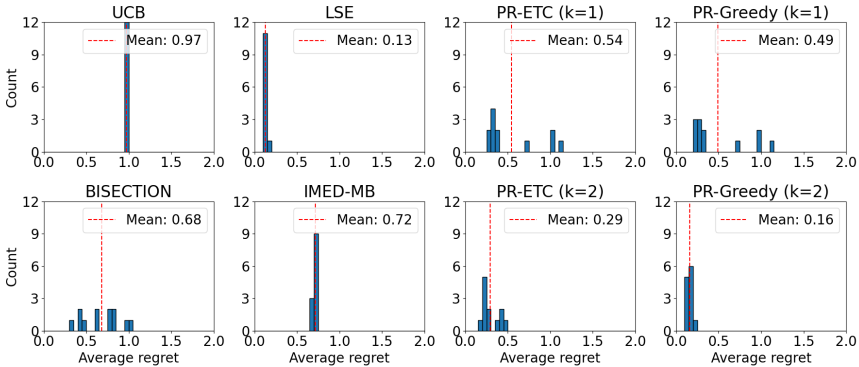
Any *wrong* parameter $(\boldsymbol{\theta}, \boldsymbol{\beta})$ whose greedy beam is suboptimal can be refuted: its predicted RSS differs from the true RSS by $\geq \gamma$.

Theorem (formal)

With probability $1 - \delta$:

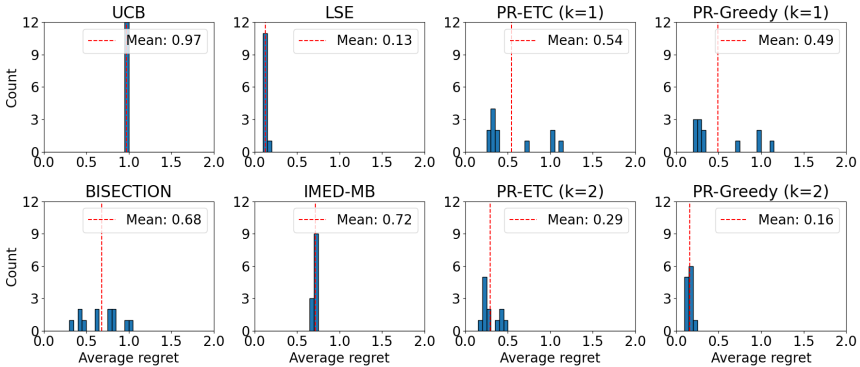
$$\mathrm{Regret}_T \;\leq\; \tilde{O}\Big(K\, R_{\max} \cdot \tfrac{k}{\gamma^2}(\log|B| + \log|\Theta|)\Big)\,.$$

Backup: DeepSense 6G regret (alt aggregation)



Distribution of normalized regret at $T = 200$ across 12 outdoor scenarios.

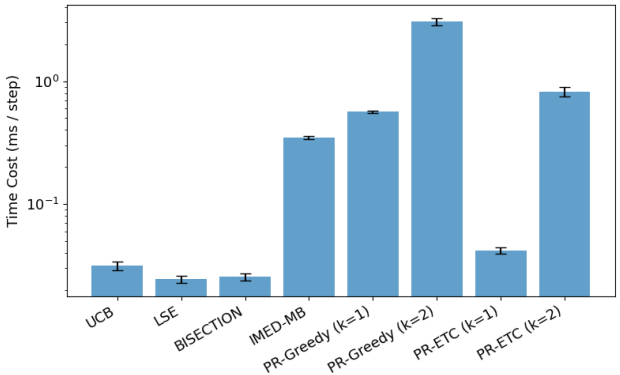
Backup: DeepSense 6G regret (alt aggregation)



Distribution of normalized regret at $T = 200$ across 12 outdoor scenarios.

PR-Greedy achieves the lowest regret on real RSS traces.

Backup: computational cost (optimized MLE)



Per-step time, DeepSense 6G

Per step on a CPU (mean)

- PR-GREEDY ($k = 2$): ~ 3.0 ms
- PR-ETC ($k = 2$): ~ 0.8 ms
- IMED MB: ~ 0.35 ms
- UCB / LSE / BISECTION: < 0.05 ms

Numbers from the optimized MLE implementation; the paper version is on the main deck.