Applied Artificial Intelligence Research

INNOVIAIR

Unimodal Bandit-Based Beam Selection for Ka-Band Mobile Direct-to-LEO Links

Tianchi Zhao[a,*], Hanwen Liu[a]

[a]School of Intelligence Science and Technology, Beijing University of Civil Engineering and Architecture, Beijing 102616, China

ARTICLEINFO

Keywords:
Beam Selection
Graph-Unimodal Bandits
Thompson Sampling
LEO Downlink
Nonstationary Learning

ABSTRACT

Directional mobile links to low-Earth-orbit satellites require beam selection from limited feedback while the preferred direction changes. This paper studies terminal receive-beam selection on a 20-GHz downlink using a graph of 19 beams arranged on a two-dimensional hexagonal lattice. Transmission success is determined by a fixed SNR threshold, giving all algorithms the same binary ACK/NACK reward and an exact Beta–Bernoulli update for Thompson sampling. unimodal UCB and unimodal Thompson sampling restrict exploration to the empirical leader and its spatial neighbors. Their stationary evaluation is complemented by periodic-restart variants for continuously drifting beam quality. Across 50 independent runs of 4,000 decisions, stationary unimodal Thompson sampling reduces expected-reward loss from 67.5 to 33.7 relative to unstructured Thompson sampling. For a direction drifting at 0.36 degrees per second, restarting every five seconds reduces dynamic loss from 679.7 to 249.7 relative to unrestarted unimodal Thompson sampling. The corresponding delivered goodput increases from 103.1 to 120.4 Mbit/s under the specified transmission duty factor. Sensitivity experiments examine three drift rates and three restart intervals. The results separate the benefit of spatially restricted exploration from the benefit of discarding outdated observations.

1.Introduction

Selecting a directional beam for a mobile satellite terminal requires balancing current data delivery against measurements of alternative beams. A low-Earth-orbit (LEO) satellite changes its position during the learning process, so the value of an old measurement depends on how quickly the geometry evolves. An effective beam-selection model must therefore specify both the feedback used for learning and the time scale over which that feedback remains relevant.

We consider a 20-GHz satellite-to-terminal downlink in a Ka-band satellite system. Here, Ka-band follows satellite-system usage, which includes 17.7–20.2 GHz downlinks and 27.5–30 GHz uplinks[1]. The carrier is specified explicitly to distinguish this usage from the 26.5–40 GHz band-letter definition. The satellite maintains its transmit pointing toward the served terminal; the terminal selects a receive beam. Carrier-frequency compensation is supplied by the communication system.

The terminal uses a single-RF-chain analog receiver and selects one of 19 receive beams per scheduled burst. It observes only the selected beam's local decoding outcome. Initial pointing from ephemeris and platform attitude centers the angular search region; online selection then follows direction changes within that region. Neighboring receive beams have related pointing errors, motivating graph-local exploration when their mean success probabilities are unimodal. The terminal represents a directional mobile platform, such as a vehicle-mounted receiver.

This paper evaluates that spatial structure separately from temporal adaptation. UCB, Thompson sampling (TS), unimodal UCB (UUCB), and unimodal Thompson sampling (UTS) are first compared at fixed geometry. Periodic-restart versions are then evaluated on a continuously moving direction. All methods use the same binary reward, power budget, beam graph, and transmission timing. This construction makes the objective, posterior update, regret metric, and delivered goodput mutually consistent.

* Corresponding author.
E-mail address: zhaotianchi@bucea.edu.cn.
https://doi.org/10.65455/nmf2rj29
Received 8 September 2026; Received in revised form 24 September 2026; Accepted 24 September 2026; Available 5 October 2026
© The Author(s) 2026. This article is published with open access in Applied Artificial Intelligence Research under a Creative Commons Attribution 4.0 International License (CC BY 4.0). https://creativecommons.org/licenses/by/4.0/

2.Related work

2.1.Learning and resource allocation in wireless systems

Satellite beam hopping and beam selection address related decisions at different scales. Tang et al. study resource allocation for LEO beam-hopping satellites in a spectrum-sharing setting[2]. Lin et al. jointly consider beam patterns and bandwidth allocation through multi-agent deep reinforcement learning[3]. Such formulations account for network-level allocation decisions. Here, the action is a terminal receive beam on one scheduled downlink, and the observation is one transmission outcome. Holding power and bandwidth fixed isolates the statistical cost of identifying a useful beam.

Recent work in this journal also examines how structure supports wireless decision making. Lin et al. construct multimodal wireless knowledge graphs for language-model-based network reasoning[4]. Their graph represents domain entities and relationships for retrieval and reasoning. The graph in this study represents physical beam adjacency and determines which actions are explored together. The connection is the explicit use of wireless-domain structure; the feedback and decision mechanisms are different.

2.2.Stochastic and structured bandits

UCB uses upper confidence indices to balance exploration and exploitation[5]. TS samples candidate success probabilities from posterior distributions and selects the largest sample[6]. Their classical stationary analyses assume that each arm retains its reward distribution over time. Sequential updating alone does not establish a guarantee for drifting reward means.

Combes and Proutiere formulate unimodal bandits on a graph and introduce optimal sampling for unimodal bandits (OSUB)[7]. Hashemi et al. exploit directional structure in bandit-based millimeter-wave beam alignment[8]. Trinh et al. establish a Bernoulli UTS analysis that applies to graph-unimodal problems[9]. These results motivate local exploration on a two-dimensional beam graph. The UUCB rule evaluated here uses a Hoeffding-style UCB index; OSUB's KL-index guarantee is therefore kept distinct from the empirical evaluation of UUCB.

2.3.Learning with changing rewards

Nonstationary bandit models measure performance against rewards that change over time. Besbes et al. relate achievable regret to the extent of reward variation[10]. Combes and Proutiere also examine time-varying unimodal rewards[7]. Our temporal mechanism is a fixed periodic restart: previous counts and posterior parameters are discarded at a specified interval. Its value is evaluated using a time-varying oracle and a drift-rate sensitivity experiment. The stationary theory in Section 6 applies to the fixed-geometry model, while the continuous-drift results establish the measured performance of the restart implementation.

3.Main contributions

The study makes three application-level contributions. First, it specifies a common ACK/NACK objective for beam selection and links this objective exactly to fixed-rate delivered goodput. Second, it implements spatial exploration on the true neighbor graph of a 19-beam hexagonal layout, including its boundary beams. Third, it evaluates periodic restarting under continuous angular drift and reports the trade-off between relearning overhead and outdated observations. The experimental package contains the simulation assumptions, random seeds, per-run results, and plotting code.

4.Problem formulation

4.1.Link and feedback model

Let $\mathrm{K}=\{1,\ldots,K\}$ denote the available terminal receive beams. At decision t, the terminal controller selects $A_t \in \mathrm{K}$. The selected transmission experiences instantaneous SNR

$$\rho_{\mathrm{i,t}}=\overline{\rho}_{\mathrm{i}}(\mathrm{t})\mathrm{Z_t} \tag{1}$$

Here $\overline{\rho}_i(t)$ is the mean SNR in linear units and Z_t is unit-mean Rician power fading. We use

$$\mathrm{Z_t}=\left|\sqrt{\frac{\kappa}{\kappa+1}}+\frac{\mathrm{U_t}+\mathrm{jV_t}}{\sqrt{2(\kappa+1)}}\right|^2 \tag{2}$$

where U_t, V_t are independent standard normal variables and κ is the Rician factor in linear units. Fading is independent between decision bursts in this model. Carrier Doppler is compensated before decoding. Inter-beam interference is excluded by the single scheduled-link abstraction.

With satellite transmit power P, receive-beam G/T $H_i(t)$ in dB/K, fixed satellite transmit gain G in dBi, bandwidth B in Hz, and additional loss L_{add} in dB, the link budget is

$$\begin{aligned}\overline{\rho}_{\mathrm{i,dB}}(\mathrm{t}) &= 10\log_{10}\mathrm{P}+\mathrm{H_i(t)}+\mathrm{G}\\ &\quad -\mathrm{L_{FS}(t)}-\mathrm{L_{add}}\\ &\quad +228.6-10\log_{10}\mathrm{B}.\end{aligned} \tag{3}$$

The constant 228.6 is the negative of Boltzmann's constant expressed in dBW/K/Hz. For range d_t in kilometers and frequency f in GHz,

$$L_{\mathrm{FS}}(t)=92.45+20\log_{10}d_t+20\log_{10}f \tag{4}$$

Let η be the transmitted spectral efficiency. We use an ideal outage-based decoding model with threshold $\rho_{\mathrm{th}}=2^{\eta}-1$. The observed reward is

$$Y_t=\mathbf{1}\{\rho_{A_t,t}\geq\rho_{\mathrm{th}}\},\qquad \mu_i(t)=\Pr(Y_t=1 \mid A_t=i) \tag{5}$$

The terminal records its local decoding result as ACK (success) or NACK (failure) before the next decision. The learning update requires no satellite feedback round trip. Continuous SNR determines this outcome inside the simulator; the policy receives only the selected receive beam's binary outcome. The threshold is a fixed-rate abstraction rather than a calibrated modulation-and-coding block-error curve.

Each decision lasts δ seconds, with fraction α used for the payload burst. The payload size is $\alpha\delta B\eta$ bits. Average delivered goodput is

$$\overline{g}_T=\frac{\alpha B\eta}{T}\sum_{t=1}^{T}Y_t \tag{6}$$

Because α, B, and η are shared by all beams, maximizing the expected ACK count is exactly equivalent to maximizing

expected delivered goodput in this model. This equivalence uses the fixed-rate objective throughout the paper.

4.2.Two-dimensional beam graph

The terminal receive-beam centers lie on a two-ring hexagonal lattice in angular coordinates. These vertices represent codebook directions, not satellite ground-coverage cells. For axial coordinates (q,r) satisfying $\max(|q|,|r|,|q+r|)\le 2$, the center is $\mathbf{b}_i=s(q+r/2,\sqrt{3}r/2)$ in angular degrees. The spacing s determines the distance between adjacent centers. Two vertices are joined when their center distance equals s. Define $N(i)$ as the neighbors of beam i and $N^+(i)=\{i\}\cup N(i)$. The 19-vertex graph has 42 edges, maximum degree six, and boundary degrees three or four. Beam labels are identifiers rather than a one-dimensional ordering.

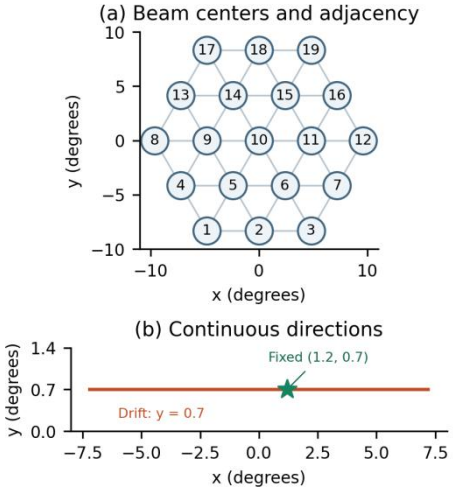


Fig 1. (a) Receive-beam centers and adjacency; (b) Fixed direction (star) and drifting direction (orange line)

For direction $\mathbf{u}_t$, the receive G/T pattern, with boresight value H_0 = 13 dB/K and beam-independent system noise temperature, is

$$H_i(t) = H_0 - 12\left(\frac{\|\mathbf{u}_t-\mathbf{b}_i\|}{\theta_{3dB}}\right)^2 \quad (7)$$

The full half-power beamwidth is θ_{3dB}: an offset of half that width produces 3 dB attenuation. The numerical study uses this smooth pattern, common fading statistics, and common additional loss. Measured sidelobes or beam-dependent blockage require a separate check of the resulting mean-reward graph.

Graph unimodality means that a unique maximizer i^* can be reached from every other beam by a path along which the expected reward strictly increases. We check this property at every evaluated geometry. With a finite graph and unique maximum, it is sufficient to verify that every nonoptimal vertex has a neighbor with a larger mean. The simulation performs this test on the gain ordering, which also orders success probabilities under the common fading model and avoids numerical underflow in very small probabilities. Passing this test establishes the condition for the evaluated layouts and trajectories.

4.3.Stationary and drifting geometries

At fixed geometry, each beam has a constant success probability μ_i. With $\Delta_i = \mu_{i^*} - \mu_i$, the expected regret is

$$R_T = \mathbb{E}[\sum_{t=1}^{T}(\mu_{i^*} - \mu_{A_t})] \quad (8)$$

This stationary model describes a fixed residual direction and range, for example a reference geometry maintained by an external pointing controller. It separates random fading from drift in the mean. Applying it over a physical acquisition window requires that changes in mean beam quality be small over that window.

For continuously changing geometry, let i_t^* =argmax$_i\mu_i(t)$. We instead evaluate.

$$R_T^{dyn} = \mathbb{E}[\sum_{t=1}^{T}(\mu_{i_t^*}(t) - \mu_{A_t}(t))] \quad (9)$$

This oracle follows the best mean-reward beam at each decision. Multiplying the realized sum of mean gaps by $\alpha\delta B\eta$ gives the corresponding conditional expected loss in delivered bits. The oracle uses mean rewards and does not observe future fading.

5.Baseline algorithms

5.1.UCB

Let $N_i(t)$ and $S_i(t)$ be the number of selections and ACKs for beam i before decision t. The empirical mean is $\hat{\mu}_i(t) = S_i(t)/N_i(t)$. After selecting every beam once, UCB chooses the largest index

$$Q_i(t) = \hat{\mu}_i(t) + \sqrt{\frac{2\ln t}{N_i(t)}} \quad (10)$$

Algorithm 1. UCB beam selection
Input: beams K, horizon T.
1. **Initialize**: Ni=Si=0 for all i∈K.
2. **for** t=1,…,K **do**
3. Select At=t; observe ACK/NACK Yt.
4. Set NAt=1, SAt=Yt.
5. **end for**
6. **for** t=K+1,…,T **do**
7. At←argmaxi∈KQi(t).
8. Use Eq. (10); break ties uniformly.
9. Receive with beam At; observe Yt.
10. NAt←NAt+1; SAt←SAt+Yt.
11. **end for**

For independent stationary rewards in [0,1], the standard UCB1 bound[5] is

$$R_T \le \sum_{i:\Delta_i>0}\frac{8\ln T}{\Delta_i} + (1+\frac{\pi^2}{3})\sum_{i:\Delta_i>0}\Delta_i \quad (11)$$

All beams use the same binary observation, so the confidence radius has the required reward scale. Initialization is included in the horizon and in every performance metric.

5.2.Thompson sampling

For each stationary success probability, TS uses an independent Beta(1,1) prior. After the observed selections and ACKs, the posterior is

$$\mu_i \mid \mathcal{H}_t \sim \text{Beta}(S_i+1, N_i-S_i+1) \quad (12)$$

where $\mathscr{H}_t$ is the observed action–reward history. The update is conjugate because Eq. (5) produces Bernoulli observations. No continuous SNR or throughput sample is inserted into a Bernoulli likelihood.

Algorithm 2. Thompson sampling (TS) beam selection
Input: beams K, horizon T.
1. **Initialize**: Ni=Si=0 for all i∈K.
2. **for** t=1,…,K do
3. Select At=t; observe ACK/NACK Yt.
4. Set NAt=1, SAt=Yt.
5. **end for**
6. **for** t=K+1,…,T **do**
7. For i∈K, independently draw
8. ϑi~ Beta(Si+1,Ni−Si+1).
9. At←argmaxiϑi over the sampled beams.
10. Receive with beam At; observe Yt.
11. NAt←NAt+1; SAt←SAt+Yt.
12. **end for**

The stationary Bernoulli analysis of TS establishes logarithmic expected regret[6]. This result applies to the fixed-geometry model; the drifting experiments use the dynamic criterion in Eq. (9).

6.Graph-local algorithms

6.1.Spatial exploration

The empirical leader is L_t=argmax$_i\hat{\mu}_i(t)$, with uniform random tie-breaking. Let ℓ_i count how often beam i has been the leader after initialization. The integer $\gamma\geq 2$ controls forced leader selections. It is an exploration parameter, independent of the number of spatial neighbors. We set γ=2 for both graph-local algorithms.

Algorithm 3. Unimodal UCB (UUCB) beam selection
Input: beams K, horizon T, graph E, γ=2.
1. **Initialize**: Ni=Si=0, ℓi=0 for all i∈K.
2. **for** t=1,…,K **do**
3. Select At=t; observe ACK/NACK Yt.
4. Set NAt=1, SAt=Yt.
5. **end for**
6. **for** t=K+1,…,T **do**
7. Lt←argmaxiSi/Ni; break ties uniformly.
8. ℓLt←ℓLt+1.
9. **if** ℓLt mod γ=0 **then**
10. At←Lt.
11. **else**
12. At←argmaxi∈N+(Lt)Qi(t).
13. Use Eq. (10); break ties uniformly.
14. **end if**
15. Receive with beam At; observe Yt.
16. NAt←NAt+1; SAt←SAt+Yt.
17. **end for**

Algorithm 4. Unimodal Thompson sampling (UTS) beam selection
Input: beams K, horizon T, graph E, γ=2.
1. **Initialize**: Ni=Si=0, ℓi=0 for all i∈K.
2. **for** t=1,…,K **do**
3. Select At=t; observe ACK/NACK Yt.
4. Set NAt=1, SAt=Yt.
5. **end for**
6. **for** t=K+1,…,T **do**
7. Lt←argmaxiSi/Ni; break ties uniformly.
8. ℓLt←ℓLt+1.
9. **if** ℓLt mod γ=0 **then**
10. At←Lt.
11. **else**
12. For i∈N+(Lt), independently draw
13. ϑi~ Beta(Si+1,Ni−Si+1).
14. At←argmaxiϑi over the sampled beams.
15. **end if**
16. Receive with beam At; observe Yt.
17. NAt←NAt+1; SAt←SAt+Yt.
18. **end for**

The augmented neighborhood contains at most seven beams. The direct implementation still scans all K empirical means to find a leader, so its total per-decision cost is $O(K)$ with $O(K+|E|)$ stored graph and count data. Locality restricts the candidate set used for selection. The vectorized reference code evaluates scores for all beams before masking the nonneighbors, so it also has an $O(K)$ score-computation cost.

6.2.Stationary theoretical scope

For a fixed, graph-unimodal Bernoulli instance with a unique best arm, the UTS result of Trinh et al.[9] gives, for each $\varepsilon > 0$,

$$R_T^{\mathrm{UTS}} \leq (1+\varepsilon)\sum_{i\in\mathcal{N}(i^*)}\frac{\Delta_i \ln T}{\mathrm{kl}(\mu_i,\mu_{i^*})} + C(\boldsymbol{\mu},\gamma,\varepsilon) \qquad (13)$$

Here kl(p,q) is the Kullback–Leibler divergence between Bernoulli distributions with means p and q. The constant is horizon-independent and depends on the instance and exploration parameters. Equation (13) describes the graph-neighbor dependence of the logarithmic term. Initialization and the instance-dependent constant remain part of finite-horizon performance. UUCB is evaluated as the specified UCB-based implementation; Eq. (13) is attributed only to UTS under its stationary assumptions.

6.3.Restarting under drift

For a restart interval H, divide time into blocks of H decisions. At each block boundary, reset all selection, success, and leadership counts to zero, restore the Beta priors, and initialize by selecting each beam once. Within a block, run the chosen base rule with the local decision age replacing t in Eq. (10). We denote these methods R-UCB, R-TS, R-UUCB, and R-UTS.

The restart schedule is fixed before an experiment and uses no oracle beam changes. Reward means continue to drift inside each block. The Beta counts then serve as a finite-memory selection rule, and Eq. (9) measures its performance. A shorter block removes old information sooner but spends a larger fraction of decisions on relearning. At H=500, the initial 19 selections occupy 3.8% of each block. These transmissions remain in the goodput and regret accounting.

7.Simulation results

7.1.Parameters and reproducibility

Table 1 gives the fully specified simulation configuration. The gain pattern, Rician factor, decision timing, duty factor, and trajectories define the evaluated single-link model. The experiments are implemented in Python. Fifty independent fading and policy runs are used for each configuration. Across algorithms within a run, common channel samples provide paired environmental conditions.

Table 1. Link and learning parameters

Parameter	Value
Altitude; carrier	1,100 km; 20 GHz downlink
Beam layout	19 vertices; 42 edges
Angular spacing; full half-power width	4.8 ; 6
Satellite transmit gain; link power	30 dBi; 70 W
Terminal boresight G/T	13 dB/K
Bandwidth; additional fixed loss	100 MHz; 2 dB
Rician factor	5 dB
Spectral efficiency; SNR threshold	8 bit/s/Hz; 255 (24.07 dB)
Decision duration; payload duration	10 ms; 2 ms
Horizon; independent runs	4,000 decisions (40 s); 50
UCB exploration coefficient	2 in Eq. (10)
Beta prior; leader period	Beta(1,1); $\gamma = 2$
Nominal restart interval	500 decisions (5 s)
Fixed angular direction	(1.2 , 0.7)
Nominal horizontal angular drift	0.36°/s; vertical coordinate 0.7
Sensitivity drift rates	0.18, 0.36, 0.54°/s
Sensitivity restart intervals	250, 500, 1,000 decisions

Each 10-ms decision contains a 2-ms scheduled downlink payload burst. The other 8 ms is reserved for receive-beam switching, local decoding, and control in this timing model. The terminal updates its policy locally before the next burst. The 20% payload duty factor is imposed equally on every method and is not a satellite-feedback round-trip requirement.

For elapsed time $\tau_t = (t-1)\delta$ and angular drift v , the moving direction is

$$\mathbf{u}_t = (v[\tau_t - (T-1)\delta/2],\ 0.7°) \tag{14}$$

The slant range follows the local tangent-plane approximation $d_t = h\sqrt{1+\tan^2 u_{t,x}+\tan^2 u_{t,y}}$, with angular arguments converted to radians. The direction is measured in a fixed local tangent frame; the receive codebook remains fixed over the 40-s segment after initial centering. This controlled geometry models direction drift within the search region, without continuous recentering by an external tracker. It specifies residual geometry directly instead of inferring an entire orbit from an unspecified time step. The nominal trajectory moves horizontally by about 14.4° over 40 s, and its slant range varies from 1,100.08 to 1,108.82 km. The preferred beam changes from 9 to 10 to 11 at approximately 13.33 and 26.67 s.

The simulator evaluates success probabilities through the noncentral chi-square distribution implied by Eq. (2), with two degrees of freedom and noncentrality 2κ. For every policy run, cumulative sums of the true mean gaps estimate Eqs. (8)–(9); the policies themselves never receive these probabilities. All tested geometries pass the graph-ascent check. Reported intervals are two-sided 95% Student-t confidence intervals across the 50 independent runs, conditional on the chosen geometry and physical model.

7.2.Fixed-geometry comparison

Figure 2 and Table 2 compare the four stationary methods. The best beam has success probability 0.8281. UTS incurs mean expected-reward loss 33.7, compared with 67.5 for TS, a reduction of 50.1%. UUCB reduces the corresponding UCB loss from 344.5 to 147.6. These improvements arise from restricting exploration to a small spatial neighborhood once the leader is useful.

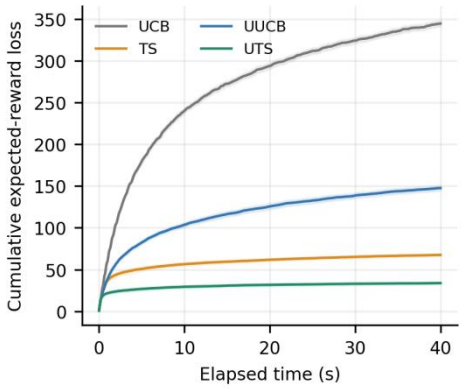


Fig 2. Stationary cumulative expected-reward loss

Note: Lines show means over 50 runs; shaded bands show pointwise 95% confidence intervals.

Table 2. Stationary results, mean ± 95% confidence interval

Method	Reward loss	Goodput (Mbit/s)
UCB	344.5 ± 3.5	118.8 ± 0.3
TS	67.5 ± 1.4	129.8 ± 0.3
UUCB	147.6 ± 3.1	126.6 ± 0.3
UTS	33.7 ± 1.5	131.2 ± 0.3

UTS selects the best beam in 98.62% of decisions over the complete horizon, including initialization. Its delivered goodput is 131.2 Mbit/s under the 20% payload duty factor, compared with 129.8 Mbit/s for TS. Reporting the observed selection fraction avoids assigning an unspecified convergence threshold to a finite simulation trace.

7.3.Continuous drift and tracking

Figure 3 and Table 3 show all eight methods on the nominal moving trajectory. R-UTS has dynamic loss 249.7,

compared with 679.7 for UTS and 458.5 for R-TS. Its reduction relative to UTS is 63.3%, and its delivered goodput is 120.4 Mbit/s. R-UTS also improves on unrestarted UCB, whose dynamic loss is 374.8 in this setting.

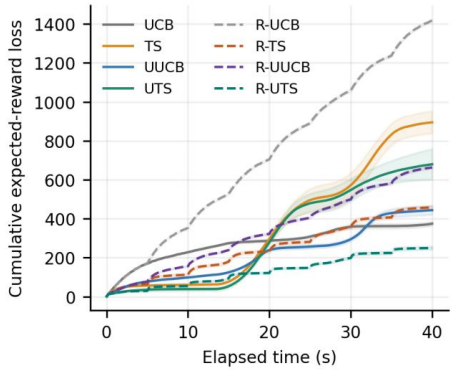


Fig 3. Dynamic expected-reward loss at 0.36°/s

Note: Dashed lines denote fixed five-second restarting. All algorithms include their initialization transmissions.

Table 3. Dynamic results, mean ± 95% confidence interval

Method	Reward loss	Goodput (Mbit/s)
UCB	374.8 ± 10.6	115.3 ± 0.5
TS	894.8 ± 57.1	94.5 ± 2.4
UUCB	443.8 ± 23.4	112.5 ± 1.0
UTS	679.7 ± 78.5	103.1 ± 3.2
R-UCB	1417.6 ± 6.1	73.7 ± 0.4
R-TS	458.5 ± 8.6	112.0 ± 0.4
R-UUCB	663.2 ± 5.9	103.7 ± 0.4
R-UTS	249.7 ± 7.7	120.4 ± 0.5

The ranking differs from the stationary experiment. Full-history TS retains evidence about formerly good beams, while periodic restarting helps it track the changing means. Restarting UCB at the same interval increases its loss because repeated broad exploration is expensive. Thus the measurements support a specific combination of graph-local sampling and finite memory, rather than a uniform benefit from applying the same restart interval to every base algorithm.

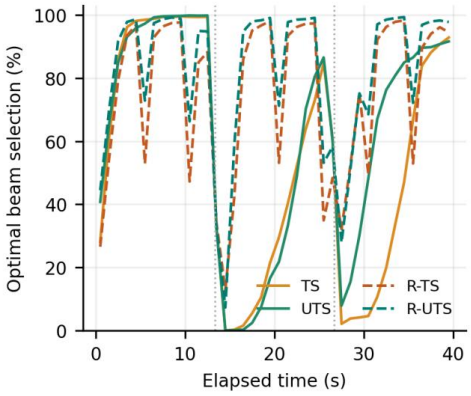


Fig 4. Best-beam selection in nonoverlapping one-second bins

Note: Dashed lines denote restarting. Vertical dotted lines mark changes of the mean-reward oracle; these changes are not provided to any policy.

To measure tracking delay, we find the first completed 100-decision window after an oracle switch in which at least 80% of selections use the new best beam. Over two switches and 50 runs, R-UTS reaches this criterion in all 100 cases, with a median delay of 2.80 s. Full-history UTS reaches it in 90 cases, with a conditional median of 6.59 s; the other ten cases are unrecovered before the next switch or the horizon. The comparison therefore reports both recovery time and recovery count.

7.4.Drift-rate and restart sensitivity

Figure 5 varies angular drift and restart interval on a fixed grid. Each grid point uses 50 independent runs, with a separate random-seed block from the main comparison. All nine settings are retained. R-UTS has lower mean dynamic loss than R-TS in each tested setting. At 0.18°/s, extending the interval from 2.5 to 10 s reduces R-UTS loss from 415.8 to 200.4, as fewer restarts reduce relearning cost. At 0.54°/s, the 10-s interval gives loss 564.1, compared with 283.9 at 5 s, showing the cost of keeping outdated observations too long.

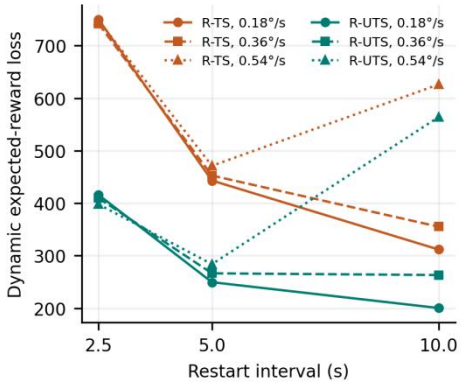


Fig 5. Restart sensitivity for TS and UTS

Note: Markers distinguish the three drift rates. The five-second setting is fixed for the main experiment.

The fixed five-second interval provides effective tracking in the nominal study; the sensitivity grid shows that a longer interval can further reduce loss at slower drift rates. A practical controller can relate the update interval to predicted angular motion, then verify it against packet-level feedback and the available exploration budget. Extending the model to measured antenna patterns, correlated fading, and time-varying blockage requires rechecking both graph structure and the relevant memory length.

8.Conclusion

Beam selection on a two-dimensional spatial graph benefits from exploiting local reward structure. A fixed-rate ACK/NACK formulation aligns the objective with the Beta – Bernoulli update and delivered goodput. In the stationary experiment, UTS halves the expected-reward loss of TS. Under continuous angular drift, periodic restarting makes old observations expire and yields the strongest performance for R-UTS among the tested methods at the specified five-second interval. The sensitivity results show how drift rate and relearning cost jointly determine a useful restart interval.

These findings support graph-local selection for the modeled directional mobile downlink and provide a reproducible basis for evaluating more detailed channel and feedback models.

Author contributions Tianchi Zhao: Conceptualization, Methodology, Software, Simulation experiments, Writing – original draft. Hanwen Liu: Formal analysis, Visualization, Writing – review & editing. All authors have read and approved the final version of the manuscript.

Funding This work was supported by the Cultivation Project Funds for Beijing University of Civil Engineering and Architecture (Grant No. X25020) and the Beijing Funding Program for Overseas High-Level Returned Talents (Grant No. 21997126001).

Data availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

[1] European Space Agency. Dual polarised high power Ka-Band feed[Z/OL]. https://resilience.esa.int/archives/projects/dual-polarised-high-power-ka-band-feed. Accessed 23 September 2026.

[2] TANG J, BIAN D, LI G, et al. Resource Allocation for LEO Beam-Hopping Satellites in a Spectrum Sharing Scenario. IEEE Access, 2021, 9: 56468-56478.

[3] LIN Z, NI Z, KUANG L, et al. Dynamic Beam Pattern and Bandwidth Allocation Based on Multi-Agent Deep Reinforcement Learning for Beam Hopping Satellite Systems. IEEE Transactions on Vehicular Technology, 2022, 71(4): 3917-3930.

[4] LIN W, SONG Y, LIU Y, et al. Constructing Multimodal Wireless Knowledge Graphs for Large Language Model–Based Network Reasoning. Applied Artificial Intelligence Research, 2026, 2(2).

[5] AUER P, CESA-BIANCHI N, FISCHER P. Finite-time Analysis of the Multiarmed Bandit Problem. Machine Learning, 2002, 47: 235-256.

[6] AGRAWAL S, GOYAL N. Analysis of Thompson Sampling for the Multi-armed Bandit Problem//Proceedings of the 25th Annual Conference on Learning Theory. PMLR, 2012: 39.1-39.26.

[7] COMBES R, PROUTIERE A. Unimodal Bandits: Regret Lower Bounds and Optimal Algorithms//Proceedings of the 31st International Conference on Machine Learning. PMLR, 2014: 521-529.

[8] HASHEMI M, SABHARWAL A, KOKSAL C E, et al. Efficient Beam Alignment in Millimeter Wave Systems Using Contextual Bandits//IEEE INFOCOM 2018. Piscataway, NJ: IEEE, 2018: 2393-2401.

[9] TRINH C, KAUFMANN E, VERNADE C, et al. Solving Bernoulli Rank-One Bandits with Unimodal Thompson Sampling[Z/OL]. arXiv:1912.03074, 2019.

[10] BESBES O, GUR Y, ZEEVI A. Optimal Exploration-Exploitation in a Multi-Armed-Bandit Problem with Non-stationary Rewards[Z/OL]. arXiv:1405.3316, 2019.