Satisficing with Binary Feedback: Multi-User mmWave Beam and Rate Adaptation via Combinatorial Bandits

Emre Özyıldırım
Dept. of Electrical and Electronics Eng.
Bilkent University
Ankara, Türkiye
emre.ozyildirim@ug.bilkent.edu.tr

Barış Yaycı
Dept. of Computer Eng.
Bilkent University
Ankara, Türkiye
b.yayci@ug.bilkent.edu.tr

Umut Eren Akturk
Dept. of Electrical and Electronics Eng.
Bilkent University
Ankara, Türkiye
eren.akturk@bilkent.edu.tr

Cem Tekin
Dept. of Electrical and Electronics Eng.
Bilkent University
Ankara, Türkiye
cemtekin@ee.bilkent.edu.tr

Abstract

We study downlink beam and rate adaptation in a multi-user mmWave MISO system where multiple base stations (BSs), each using analog beamforming from finite codebooks, serve multiple single-antenna user equipments (UEs) with a unique beam per UE and discrete data transmission rates. BSs learn about transmissions success based on ACK/NACK feedback. To encode service goals, we introduce a satisficing throughput threshold τ_r and cast joint beam and rate adaptation as a combinatorial semi-bandit over beam-rate tuples. Within this framework we propose SAT-CTS, a lightweight, threshold-aware policy that blends conservative confidence estimates with posterior sampling, steering learning toward meeting τ_r rather than merely maximizing. We evaluate the performance via cumulative satisficing regret to τ_r alongside standard regret and fairness. Experiments under time varying sparse multipath channels show that SAT-CTS consistently reduces satisficing regret and maintains competitive standard regret, while achieving favorable average throughput and fairness across users, indicating that modest, feedback-efficient learning can equitably allocate beams and rates to meet QoS targets without channel state knowledge.

1 Introduction

We study downlink data transmission in a multi-user, multi-base station Multiple Input Single Output (MISO) system, where multiple base stations (BSs) serve single-antenna user equipments (UEs) [12]. In millimeter-wave (mmWave) communications, transmitters employ highly directional narrow beams to mitigate severe path loss and blockage. Reliable links require these beams to be accurately steered toward the UEs [25, 18]. Each UE can be served by one of several candidate beams across different BSs. A BS transmits data to a UE using the selected beam at the highest feasible rate, determined by the chosen modulation and coding scheme (MCS). We model this as a joint beam and rate adaptation problem without channel state information (CSI), where after each transmission the BS receives binary ACK/NACK feedback for the selected beam–rate pair.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI and ML for Next-Generation Wireless Communications and Networking (AI4NextG).

Table 1: Comparison with related works.

Work	Combinatorial Setup	MAB & Comm. System Together	Satisficing Threshold	Beam + Data (Align & Rate)
MAMBA[2]	✗	✓	✗	✓
PE[18]	✗	✓	✗	✗
CCBM[10]	✓	✓	✗	✗
CCVB for BS[13]	✓	✓	✗	✗
FBA[6]	✗	✗	✗	✗
CCV-MAB[11]	✓	✗	✗	✗
Our work	✓	✓	✓	✓

We consider a centralized architecture in which a learner coordinates beam and rate assignments for the BSs and UEs at the beginning of each time slot over an ultra-low latency control channel. At the end of each slot, BSs return feedback to the learner, which updates assignments to improve system performance.

We cast this problem as a satisficing combinatorial multi armed bandit (CMAB) with semi-bandit feedback and develop a learning algorithm that ensures acceptable per-UE average throughput. Our approach follows Herbert Simon's bounded-rationality perspective, where agents aim to reach an aspiration level under limited time and information [15]. Instead of focusing on convergence to a unique maximizer, we measure how quickly the assignments achieve a satisficing threshold through the notion of satisficing regret.

We evaluate this metric in a simulated multi-BS, multi-UE MISO system with realistic channel vectors generated by the DeepMIMO simulator [1]. UEs are modeled as stationary within each time slot, consistent with the quasi-static assumption used in prior beam-training studies [18, 16]. Our proposed algorithm SAT-CTS, which takes the satisficing threshold as an input and tests the selected super arm's performance under this threshold, outperforms standard Combinatorial Thompson Sampling (CTS)[21] and Combinatorial Upper Confidence Bound (CUCB) [3] baselines in terms of satisficing regret.

2 Related Work

In mmWave networks, the central challenge is assigning beams together with appropriate rates so that data transmissions succeed under directional links and SNR thresholds. Early non-bandit approaches tackle only the alignment phase: e.g., Hassanieh et al. [6] design fast probing to identify a good beam but do not address multi-user assignment or data rate selection and do not consider the problem in a multi armed bandit (MAB) formulation. Bandit formulations then appear for alignment: Wei et.al.[18] and Wu et.al.[22] treat fast beam alignment as pure exploration, identifying the best beam from pilot measurements (received signal strength) without sending real data or choosing rates. The recent Contextual Combinatorial Beam Management [10] work line extends alignment to a contextual, combinatorial setting with multiple UEs and BSs and a probing budget, still operating on pilot signals and maximizing a contextual reward. By contrast, Contextual Combinatorial Volatile Bandits for multi user small BS association [13] emphasizes rate/association decisions under context and volatility, but does not perform directional beam selection. Relatedly, CCV-MAB [11] addresses contextual combinatorial bandits with time-varying availability, offering regret guarantees via adaptive discretization, yet it remains communication agnostic (no beam/rate modeling or ACK/NACK signals). MAMBA [2] jointly adapts beam and MCS from ACK/NACK, yet remains a single-user, single-BS formulation without combinatorial matching.

Our contribution We propose SAT-CTS, which performs joint beam–rate adaptation together with BS-UE association in a combinatorial setting with semi-bandit (ACK/NACK) feedback, and introduces a satisficing threshold which indicates the target average throughput per user. This fills the gap between alignment-only pilot methods and single-link bandits by providing a target-aware, multi-user assignment mechanism that directly operates on the data plane.

3 Problem Formulation

As shown in Figure 1 we consider a multi-user mmWave MISO system where B BSs, each with N transmit antennas, serve M single-antenna UEs.

Each BS $b \in [B]$ selects its beams from its predefined analog beamforming codebook of size K [7], defined as $\mathcal{C}_b := \{ \mathbf{f}_{b,k} \in \mathbb{C}^N : k \in [K]\}$, with each beamforming vector $\mathbf{f}_{b,k}$ normalized, $\|\mathbf{f}_{b,k}\|_2 = 1$. The mmWave channel, as observed in measurement studies [24, 23], follows a multipath model with a small number of propagation paths, resulting in a sparse structure. While our learning algorithms are not restricted to work under a specific channel model, a common model for the channel vector from base station b to user m is the $L_{m,b}$-path Saleh–Valenzuela model [14], given as

$$\mathbf{h}_{m,b} = \sqrt{\frac{N}{L_{m,b}}} \sum_{\ell=1}^{L_{m,b}} \beta_{m,b,\ell}\, \mathbf{a}(\cos\theta_{m,b,\ell}), \tag{1}$$

where $\beta_{m,b,\ell}$ and $\theta_{m,b,\ell}$ are the complex gain and angle-of-departure of path ℓ for that specific UE-BS link, and the array steering vector is $\mathbf{a}(\cos\theta_{m,b,\ell}) = [1, e^{j\frac{2\pi}{\lambda} d \cos\theta_{m,b,\ell}}, \ldots, e^{j\frac{2\pi}{\lambda} d(N-1)\cos\theta_{m,b,\ell}}]^T$ defined as in [4]. Here, λ represents the carrier wavelength and d represents the antenna spacing. We assume a quasi-static channel model, where the channel vector $\mathbf{h}_{m,b}$ remains constant during a time slot, but can vary between time slots.

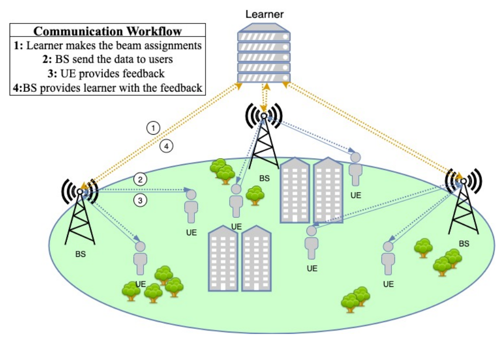


Figure 1: System model.

Consider BS b transmitting symbol s to UE m with transmit power p_b. Without loss of generality, assume $s = 1$. When BS b transmits with with beam $\mathbf{f}_{b,k}$, the received signal at UE m is given by $y_m = \sqrt{p_b}\mathbf{h}_{m,b}^H\mathbf{f}_{b,k} + n_m$ where $n_m \sim \mathcal{CN}(0, \sigma_m^2)$ is the complex additive white Gaussian noise (AWGN) [18]. The instantaneous received-signal-strength (RSS) is equal to the magnitude squared of the received signal, which is given as $\mathrm{RSS}_m(\mathbf{f}_{b,k}) = |y_m|^2 = |\sqrt{p_b}\,\mathbf{h}_{m,b}^H\mathbf{f}_{b,k} + n_m|^2$. Letting $a_{m,b,k} = \mathbf{h}_{m,b}^H\mathbf{f}_{b,k}$, we have

$$\mathrm{RSS}_m(\mathbf{f}_{b,k}) = \left|\sqrt{p_b}\, a_{m,b,k} + n_m\right|^2 = p_b\, |a_{m,b,k}|^2 + 2\sqrt{p_b}\, \Re\{a_{m,b,k}^*\, n_m\} + |n_m|^2.$$

Note that $2\sqrt{p_b}\,\Re\{a_{m,b,k}^*\, n_m\} \sim \mathcal{N}\big(0,\ 2p_b\, |a_{m,b,k}|^2\, \sigma_m^2\big)$ and the noise-power term $|n_m|^2$ is an exponential random variable with mean σ_m^2 and variance σ_m^4. Assuming a high SNR regime, i.e., $p_b\, |a_{m,b,k}|^2 \gg \sigma_m^2$, we obtain the following approximation.

$$\mathrm{RSS}_m(\mathbf{f}_{b,k}) = |\sqrt{p_b} a_{m,b,k} + n_m|^2 \approx p_b |a_{m,b,k}|^2 + \mathcal{N}(0, 2p_b|a_{m,b,k}|^2\sigma_m^2). \tag{2}$$

The expected RSS within a time slot is then given as $\mathbb{E}[\mathrm{RSS}_m(\mathbf{f}_{b,k})] = p_b\left|\mathbf{h}_{m,b}^H\mathbf{f}_{b,k}\right|^2$, which yields a signal to noise ratio (SNR) equal to $p_b\left|\mathbf{h}_{m,b}^H\mathbf{f}_{b,k}\right|^2/\sigma_m^2$, which is also used in [4].

Combinatorial Multi-armed Bandit Formulation Beam k of BS b is denoted by tuple (b, k). We denote the set of all beams by $\mathcal{K} := [B] \times [K]$. Additionally, we define a discrete set of feasible transmission rates $\mathcal{R} = \{r_1, r_2, \ldots, r_R\}$ where $r_1 < r_2 < \ldots < r_R$ represent the available data rates in bits per channel use. Different data rates can be achieved by choosing a different MCS. We assume $|\mathcal{K}| \geq M$. Let T represent the time horizon. We consider a centralized system where the learner coordinates beam and rate selection for downlink transmission.

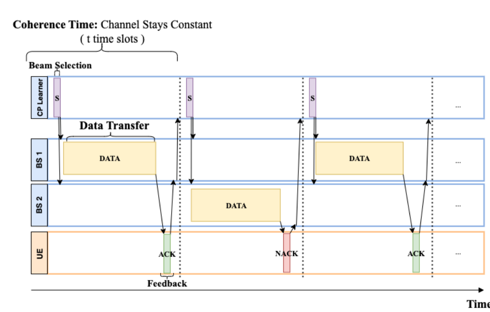


Figure 2: Process of beam and rate assignment.

As presented in Figure 2, the following events take place sequentially at each time slot t. At the beginning of each time slot $t \in [T]$, the learner assigns exactly one beam to each UE via a mapping $\pi_t : \{1, \ldots, M\} \to \mathcal{K}$, and selects a transmission rate for each UE via a mapping $\rho_t : \{1, \ldots, M\} \to \mathcal{R}$. Here, $\pi_{m,t} = (b_{m,t}, k_{m,t})$ represents the beam assigned to UE m with $b_{m,t}$ representing the assigned BS and $k_{m,t}$ representing the beam of the assigned BS. For each UE $m \in [M]$, the learner communicates this information with the associated BS through an ultra-low latency control channel. After this communication, BS $b_{m,t}$ uses its beam $\pi_{m,t} = (b_{m,t}, k_{m,t})$ to transmit data to UE m at rate $\rho_{m,t}$.

We assume that each beam $k \in \mathcal{K}$ can serve at most one UE as in [19, 20]. Let $\mathcal{S} := \{\{(\pi_1, \rho_1), \ldots, (\pi_M, \rho_M)\} : \pi_m \in \mathcal{K}, \rho_m \in \mathcal{R}, \pi_m \neq \pi_n \text{ for } n \neq m\}$ represent the set of super arms. For a super arm $s \in \mathcal{S}$, let $\pi_m(s)$ be the beam assigned to UE m and $\rho_m(s)$ be the rate selected for UE m.

For a given rate $\rho_{m,t}$ and beam allocation $\pi_{m,t} = (b, k)$, the transmission is successful if the instantaneous SNR exceeds the threshold required for the selected rate. Define the SNR threshold for rate $r \in \mathcal{R}$ as $\gamma_{\text{th}}(r) = 2^r - 1$ based on the Shannon-Hartley theorem. Note that we assume that inter-cell interference on SNR is ignored. At the end of time slot t, BS b receives the transmission success indicator (ACK/NACK feedback) from UE m, which is given by:

$$x_{m,t} = \begin{cases} 1 & \text{if } \frac{p_b |\mathbf{h}_{m,b}^H \mathbf{f}_{b,k}|^2}{\sigma_m^2} \geq \gamma_{\text{th}}(\rho_{m,t}) \quad \text{(ACK)} \\ 0 & \text{otherwise} \quad \text{(NACK - outage)} \end{cases}$$

which is the similar to the model that is used in previous works [5, 17]. The instantaneous reward for UE m at time t is $r_{m,t} = \rho_{m,t} \times x_{m,t}$, where the UE receives the selected rate $\rho_{m,t}$ if transmission is successful and zero otherwise. The transmission success probability for UE m with beam (b, k) and rate r is denoted as $\psi_{m,(b,k),r} := \mathbb{P}\left(\frac{p_b |\mathbf{h}_{m,b}^H \mathbf{f}_{b,k}|^2}{\sigma_m^2} \geq \gamma_{\text{th}}(r)\right)$, where the randomness is over the channel vector $\mathbf{h}_{m,b}$ which is sampled as i.i.d random variable across rounds. An optimal super arm is a super arm that maximizes the expected total throughput, i.e., $(\pi^*, \rho^*) \in \arg\max_{(\pi,\rho) \in \mathcal{S}} \mathbb{E}\left[\sum_{m=1}^{M} \rho_m \times x_{m,t}\right] = \arg\max_{(\pi,\rho) \in \mathcal{S}} \sum_{m=1}^{M} \rho_m \times \psi_{m,\pi_m,\rho_m}$ where ρ_m is the transmission rate selected for UE m, $x_{m,t} \in \{0, 1\}$ is the ACK/NACK feedback, and $\psi_{m,\pi_m,\rho_m} = \mathbb{P}(x_{m,t} = 1)$ is the transmission success probability for UE m with beam π_m and rate ρ_m.

Given assignment $S_t = \{(\pi_{1,t}, \rho_{1,t}), \ldots, (\pi_{M,t}, \rho_{M,t})\}$, the learner observes at the end of slot t the per-UE ACK/NACK feedback $x_{m,t} \in \{0, 1\}$ for $m = 1, \ldots, M$. This information is communicated by the BSs to the learner via the ultra-low latency control link.

Define the *satisficing threshold* as $\tau_r \in \mathbb{R}_+$ representing the target average throughput per UE. This threshold can be interpreted as the desired minimum average data rate that should be achieved across all UEs. For instance, in 6G scenarios one can aim for average throughput $\tau_r = 2.5$ Gbits/sec per UE per time slot. The *per-round satisficing regret* of S_t is $\Delta(S_t) := [\tau_r - \frac{1}{M}\sum_{m=1}^{M} \rho_{m,t} \cdot \psi_{m,\pi_{m,t},\rho_{m,t}}]_+$, and the *cumulative satisficing regret* over T time slots is $\mathcal{R}_S(T) := \sum_{t=1}^{T} \Delta(S_t)$.

4 Fast Beam and Rate Adaptation via SAT-CTS

Satisficing Combinatorial Thompson Sampling (SAT-CTS) aims to select an appropriate BS-beam-rate combination for each UE to meet the target throughput requirement efficiently. Each arm (b, k, r) keeps a Beta prior on its success probability ψ and its plays and success counts. At each round t, the algorithm samples $\tilde{\psi} \sim \text{Beta}(A, B)$, forms $\theta = r\tilde{\psi}$, and builds a TS super arm according to total throughput. It also computes empirical means $\hat{\psi}$ and the half-width $c(t, n) = \sqrt{0.5 \log(\max\{2, t\}) / \max(1, n)}$, giving indices $\text{LCB} = r(\hat{\psi} - c)$, $\text{MEAN} = r\hat{\psi}$, and $\text{UCB} = r(\hat{\psi} + c)$; maximization of each gives three candidate super-arms whose sums are compared to the target $M\tau_r$. The gate plays the first that satisfies the target in the order LCB, MEAN, UCB respectively; otherwise it falls back to the TS proposal. After observing per user ACK/NACK feedback, it updates counts and Beta posteriors. The Algorithm 1 is the short version of the pseudo code, the full version is available as Algorithm 4 in the Appendix.

Algorithm 1 SAT-CTS (Short Version)

Require: Users M; beam set $\mathcal{K}$; rate set $\mathcal{R}$; horizon T; target τ_r

Init: Set Beta priors, play counters, and the feasible assignment set $\mathcal{S}$.
for $t = 1$ to T **do**
  TS step: Sample all BS–beam–rate triplets for all UEs and build a TS score table.
  Index step: From observed data, form three score tables: LCB, MEAN, and UCB.
  Best sets: For each table (LCB/MEAN/UCB) select its best assignment; also select the TS best assignment.
  Gate:
    If the LCB candidate meets the target, play it; else if the MEAN candidate meets the target, play it;
    else if the UCB candidate meets the target, play it; otherwise play the TS candidate.
  Play & observe: Execute the chosen assignment; collect per-user ACK/NACK (semi-bandit feedback).
  Update: Update counters and Beta priors only for the arms that were played.
end for

5 Experiments

5.1 Experimental Setup

We consider a multi-cell mmWave communication system with 3 BSs, where each BS is equipped with a uniform linear array (ULA) of 64 antenna elements with full wavelength spacing. Each BS can form 120 directional beams, resulting in a total of 360 beams across the network that serve 12 users distributed across the coverage area. Our implementation is available at https://github.com/Bilkent-CYBORG/Satisficing-with-Binary-Feedback-for-Combinatorial-Beam-Alignment. The channel characteristics are obtained using the DeepMIMO dataset [1], specifically the `city_3_houston_28` scenario, which provides realistic channel realizations based on ray-tracing simulations in an urban environment. The system operates with a bandwidth of 50 MHz [1], enabling high-throughput communication in the millimeter-wave band. The system employs adaptive modulation with three discrete rate levels: $\{6, 8, 12\}$ bits/symbol, which correspond to achievable data rates of 300 Mbps, 400 Mbps, and 600 Mbps respectively with the 50 MHz bandwidth. The target performance which is equal to satisficing threshold is set at 8 bits/symbol for realizable case, corresponding to 400 Mbps per user, and in non-realizable case it is set to 25 bits/symbol which corresponds to 1.25 Gbps. For all measurements, simulations were repeated for 100 iterations and the average of 100 iterations with a standard deviation is plotted. For the optimization oracle, Hungarian Algorithm is used [9]. Figure 4 shows the deployment of users and BSs, the channel gain randomness is achieved by modeling the array steering vector in (1) as a random variable. As benchmarks, CUCB [3] and CTS [21] algorithms were used. Their pseudo codes are given in Algorithms 2 and 3 in the Appendix. Additionally, we measured the cumulative satisficing regret for varying thresholds and assessed user-level fairness on the same performance metric as our objective, throughput. Let $r_{m,t} = \rho_{m,t} x_{m,t}$ be UE m's throughput (bits/symbol) at slot t; define cumulative throughput $G_m(T) = \sum_{t=1}^{T} r_{m,t}$. We report Jain's Fairness Index [8]:

$$J(T) = \frac{\left(\sum_{m=1}^{M} G_m(T)\right)^2}{M \sum_{m=1}^{M} G_m^2(T)} \in \left[\frac{1}{M}, 1\right],$$

where $J = 1$ indicates perfectly even throughput across UEs and smaller values indicate disparity. The cumulative fairness over time is also measured in simulation.

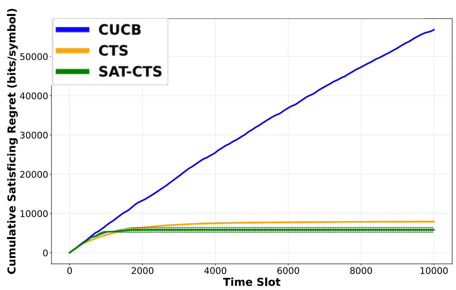
5.2 Results and Discussion

Table 2: Fairness on cumulative throughput with a realizable target (mean $\pm$ std).

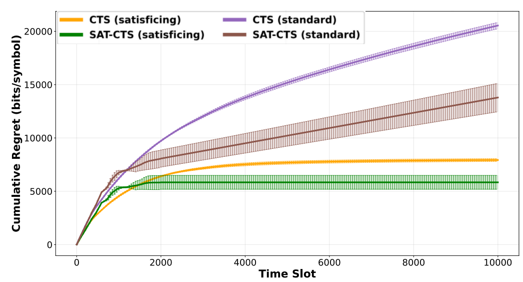
Algorithm	t=1000	t=1500	t=2500	t=5000	t=7500	t=10000
CUCB	0.640 ± 0.030	0.700 ± 0.020	0.732 ± 0.028	0.774 ± 0.018	0.787 ± 0.015	0.804 ± 0.013
CTS	0.660 ± 0.030	0.700 ± 0.025	0.792 ± 0.021	0.838 ± 0.017	0.866 ± 0.014	0.888 ± 0.012
$\mathrm{SAT-CTS}$	0.740 ± 0.020	0.775 ± 0.015	0.787 ± 0.012	0.793 ± 0.011	0.796 ± 0.010	0.798 ± 0.010

Table 3: Cumulative satisficing regret with changing satisficing thresholds at $t = 10000$ (mean in bits/symbol).

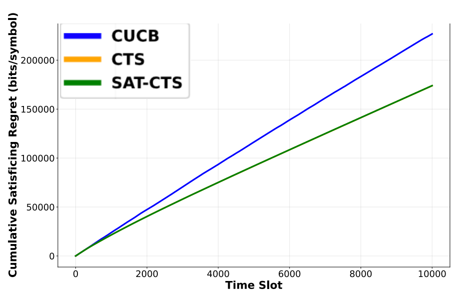
	Threshold τ			
Algorithm	**6**	**8**	**10**	**12**
CTS	5113	7568	19382	34561
SAT-CTS	2547	5561	17623	33789



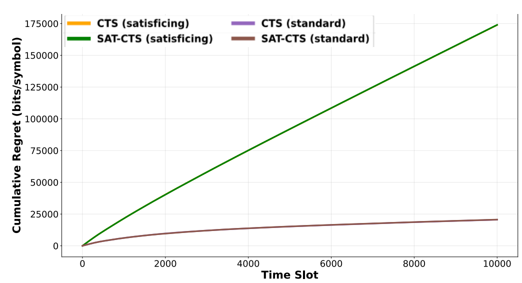
(a) Satisficing regret comparison of three algorithms on same plot with realizable target.



(b) Satisficing and standard regret comparison of CTS and SAT-CTS on same plot with realizable target.



(c) Satisficing regret comparison of three algorithms on same plot with non realizable target.



(d) Satisficing and standard regret comparison of CTS and SAT-CTS on same plot with non realizable target.

Figure 3: Cumulative regret comparison graphs.

Figure 3 indicates the cumulative regret. Since the standard deviation is very small compared to the cumulative regret values, some error bars are not visible. Under a *realizable* threshold (Fig. 3a), SAT- CTS achieves the threshold fastest and then stabilizes, resulting in the lowest cumulative *satisficing* regret throughout the horizon. In the *realizable* case, SAT-CTS achieves approximately 25% lower *satisficing* regret than CTS at $T = 10000$ in the steady region. CUCB's satisficing regret looks nearly linear over our reported horizon, but this reflects an insufficient time horizon rather than true linear asymptotics; once CUCB reliably finds a feasible super-arm, its curve is expected to flatten (converge) beyond the measured T. On *standard* regret with *realizable* threshold (Fig. 3b), SAT- CTS also maintains a slight but persistent advantage over CTS in the considered time horizon, indicating that the satisficing gate not only achieves the target early but also steers learning toward high-throughput super-arms without sacrificing exploitation efficiency.

When the target is non realizable (Fig. 3c), all methods obtain *satisficing* regret roughly linearly simply because no super arm can meet the threshold. In SAT- CTS, the decision gate (LCB $\rightarrow$ mean $\rightarrow$ UCB) frequently finds that no candidate achieves the target; it therefore falls back to the Thompson–sampling candidate, which is effectively the CTS choice on the super–arm. As a result, SAT- CTS behaves like CTS for most rounds and their curves are nearly indistinguishable, with small differences only in the early phase when SAT- CTS briefly attempts to satisfy the impossible constraint. On the *standard* regret (Fig. 3d) they performed better, since the satisficing threshold is higher than the maximum achievable average throughput. CUCB remains dominated in this regime as well and its regret is excluded from the plot since it incurs much higher regret than CTS and SAT-CTS. Fairness results in Table 2 reveal a trade off between throughput and equity, but they also highlight an early phase fairness advantage for SAT- CTS. By $t = 1000$, SAT- CTS achieves a Jain index of ≈ 0.74—about 0.07–0.09 above CTS and ≈ 0.12 above CUCB and it maintains the smallest variability, reflecting stable, high-confidence assignments that quickly balance cumulative service across users. Although CTS reaches and slightly passes on long horizons thanks to continued exploration of Thompson sampling, SAT- CTS provides the best short and middle horizon fairness and throughput balance in our experiments, while CUCB starts lowest and improves slowly across the horizon.

6 Conclusion and Future Research

We formulated a multi-user beam–rate selection problem with ACK/NACK feedback as a satisficing combinatorial bandit with a per-user throughput threshold and a no–beam-sharing assignment, for which we proposed SAT-CTS, an if gated policy that blends conservative and exploratory indices. In experiments on time-varying channels using DeepMIMO as an accurate simulator of real life environments, SAT-CTS reached realizable targets in shorter periods of time and with lower cumulative satisficing regret than baselines found in the literature, while behaving comparably to CTS when the target was infeasible. As future work, it is possible to extend our formulations and approaches to contextual combinatorial bandits that exploit side information (e.g., geometry, mobility) to accelerate learning and improve robustness. One could also explicitly incorporate fairness objectives as constraints or via multi-objective optimization.

Acknowledgments

This work was supported in part by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) BİLGEM Grant; TÜBİTAK Grant 124E065; by the Turkish Academy of Sciences Distinguished Young Scientist Award Program (TÜBA-GEBİP-2023); by TÜBİTAK 2024 Incentive Award.

References

[1] Ahmed Alkhateeb. DeepMIMO: A generic deep learning dataset for millimeter wave and massive MIMO applications. In *Proc. Information Theory and Applications Workshop*, 2019.

[2] Irmak Aykin, Berk Akgun, Mingjie Feng, and Marwan Krunz. MAMBA: A multi-armed bandit framework for beam tracking in millimeter-wave systems. In *Proc. 2020 IEEE Conference on Computer Communications*, pages 1469–1478, 2020.

[3] Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *Proc. 30th International Conference on Machine Learning*, pages 151–159, 2013.

[4] Debamita Ghosh, Manjesh K. Hanawal, and Nikola Zlatanov. UB3: Fixed budget best beam identification in mmwave massive MISO via pure exploration unimodal bandits. *IEEE Transactions on Wireless Communications*, 23(10):12658–12669, 2024.

[5] Ruchir Gupta, K. Lakshmanan, and Abhay Kumar Sah. Beam alignment for mmwave using non-stationary bandits. *IEEE Communications Letters*, 24(11):2619–2622, 2020.

[6] Haitham Hassanieh, Omid Abari, Michael Rodriguez, Mohammed Abdelghany, Dina Katabi, and Piotr Indyk. Fast millimeter wave beam alignment. In *Proc. 2018 Conference of the ACM Special Interest Group on Data Communication*, SIGCOMM '18, page 432–445, 2018.

[7] Shiwen He, Jiaheng Wang, Yongming Huang, Björn Ottersten, and Wei Hong. Codebook-based hybrid precoding for millimeter wave multiuser systems. *IEEE Transactions on Signal Processing*, 65(20):5289–5304, 2017.

[8] Raj Jain, Dah-Ming W. Chiu, and William R. Hawe. A quantitative measure of fairness and discrimination for resource allocation in shared computer systems. Technical Report DEC-TR-301, Digital Equipment Corporation, September 1984.

[9] Harold W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

[10] Zhizhen Li, Xuanhao Luo, Mingzhe Chen, Chenhan Xu, Shiwen Mao, and Yuchen Liu. Contextual combinatorial beam management via online probing for multiple access mmWave wireless networks. *IEEE Journal on Selected Areas in Communications*, 43(3):959–972, 2025.

[11] Andi Nika, Sepehr Elahi, and Cem Tekin. Contextual combinatorial volatile multi-armed bandit with adaptive discretization. In *Proc. 23rd International Conference on Artificial Intelligence and Statistics*, pages 1486–1496, 2020.

[12] A.J. Paulraj, D.A. Gore, R.U. Nabar, and H. Bolcskei. An overview of MIMO communications - a key to gigabit wireless. *Proceedings of the IEEE*, 92(2):198–218, 2004.

[13] Muhammad Anjum Qureshi, Andi Nika, and Cem Tekin. Multi-user small base station association via contextual combinatorial volatile bandits. *IEEE Transactions on Communications*, 69(6):3726–3740, 2021.

[14] A. A. M. Saleh and R. A. Valenzuela. A statistical model for indoor multipath propagation. *IEEE Journal on Selected Areas in Communications*, 5(2):128–137, 1987.

[15] Herbert A. Simon. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1):99–118, 1955.

[16] Zihan Tang, Jun Wang, Jintao Wang, and Jian Song. A high-accuracy adaptive beam training algorithm for MmWave communication. In *2018 IEEE Globecom Workshops*, pages 1–6, 2018.

[17] David Tse and Pramod Viswanath. *Fundamentals of Wireless Communication*, page 187. Cambridge University Press, Cambridge, UK, 2005.

[18] Thomas Vincent, Ehsan Ghadimi, Vincenzo Lottici, and Ahmed Alkhateeb. Fast mmwave beam alignment via correlated bandit learning. In *IEEE International Conference on Communications*, pages 1–6, 2022.

[19] Junyuan Wang, Huiling Zhu, Lin Dai, Nathan J Gomes, and Jiangzhou Wang. Low-complexity beam allocation for switched-beam based multiuser massive MIMO systems. *IEEE Transactions on Wireless Communications*, 15(12):8236–8248, 2016.

[20] Junyuan Wang, Huiling Zhu, Nathan J Gomes, and Jiangzhou Wang. Frequency reuse of beam allocation for multiuser massive MIMO systems. *IEEE Transactions on Wireless Communications*, 17(4):2346–2359, 2018.

[21] Siwei Wang and Wei Chen. Thompson sampling for combinatorial semi-bandits. In *Proc. 35th International Conference on Machine Learning*, pages 5114–5122, 2018.

[22] Wen Wu, Nan Cheng, Ning Zhang, Peng Yang, Weihua Zhuang, and Xuemin Shen. Fast mmwave beam alignment via correlated bandit learning. *IEEE Transactions on Wireless Communications*, 18(12):5894–5908, 2019.

[23] Xianyue Wu, Cheng-Xiang Wang, Jian Sun, Jie Huang, Rui Feng, Yang Yang, and Xiaohu Ge. 60-GHz millimeter-wave channel measurements and modeling for indoor office environments. *IEEE Transactions on Antennas and Propagation*, 65(4):1912–1924, 2017.

[24] Hequn Zhang, Yue Zhang, John Cosmas, Nawar Jawad, Wei Li, Robert Muller, and Tao Jiang. mmwave indoor channel measurement campaign for 5G new radio indoor broadcasting. *IEEE Transactions on Broadcasting*, 68(2):331–344, 2022.

[25] Pei Zhou, Xuming Fang, Yuguang Fang, Yan Long, Rong He, and Xiao Han. Enhanced random access and beam training for millimeter wave wireless local networks with high user density. *IEEE Transactions on Wireless Communications*, 16:7760–7773, 2017.

A Appendix

Table 4: Summary of notation for multi-UE multi-BS mmWave MISO system.

Symbol	Description
System Parameters	
B	Number of BSs
M	Number of users
K	Number of beams in each BS codebook
W	System bandwidth
N	Number of antennas at BS b
$\mathcal{C}_b = \{\mathbf{f}_{b,k}\}_{k=1}^{K}$	Beamforming codebook for BS b
λ	Carrier wavelength for BS b
d	Antenna element spacing at BS b
p	Transmit power of BS b
Channel and Signal Parameters	
$\mathbf{h}_{m,b}$	Channel vector from BS b to user m
$L_{m,b}$	Number of propagation paths between BS b and user m
$\beta_{m,b,\ell}$	Complex gain of the ℓ-th path for the (m,b) link
$\theta_{m,b,\ell}$	Angle-of-departure of the ℓ-th path for the (m,b) link
$\mathbf{a}(\cos\theta_{m,b,\ell})$	Array steering response vector at spatial angle θ
$\gamma_{m,b}$	LoS path gain for the link between BS b and user m
$\theta_{m,b}$	LoS angle-of-departure from BS b to user m
$\mathbf{f}_{b,k}$	k-th beamforming vector from BS b
y_m	Received baseband signal at user m
$\text{RSS}_m(\mathbf{f}_{b,k})$	Measured received power at user m for beam $\mathbf{f}_{b,k}$
$a_{m,b,k}$	Instantaneous projection: $a_{m,b,k} = \mathbf{h}_{m,b}^{H}\mathbf{f}_{b,k}$
Noise Parameters	
n_m	Additive complex Gaussian noise at user m, $n_m \sim \mathcal{CN}(0, \sigma_m^2)$
σ_m^2	Noise variance at user m

Figure 4: User and BS locations for the experimental setup.

Algorithm 2 CUCB: Combinatorial Upper Confidence Bound

Require: M users, beam set $\mathcal{K} = [B] \times [K]$, rate set $\mathcal{R} = \{r_1 < \cdots < r_R\}$, time horizon T

Initialize:
  $n_{m,(b,k),r} \leftarrow 0, \quad \hat{\psi}_{m,(b,k),r} \leftarrow 0$ for all $m \in [M],\ (b,k) \in \mathcal{K},\ r \in \mathcal{R}$ {Counts and success-rate estimates}
for $t = 1$ to T **do**
  Step 1: UCB index computation on success probability
  for each $m \in [M],\ (b,k) \in \mathcal{K},\ r \in \mathcal{R}$ **do**
    if $n_{m,(b,k),r} = 0$ **then** $\mathrm{UCB}_{m,(b,k),r} \leftarrow +\infty$ **else**
    $\mathrm{UCB}_{m,(b,k),r} \leftarrow \hat{\psi}_{m,(b,k),r} + \sqrt{\dfrac{2\log t}{n_{m,(b,k),r}}}$
    $\theta_{m,(b,k),r} \leftarrow r \cdot \mathrm{UCB}_{m,(b,k),r}$
  end for

  Step 2: Construct super-arm via optimal assignment (no sharing)
    $\mathcal{S} = \{\{(\pi_1, \rho_1), \ldots, (\pi_M, \rho_M)\} : \pi_m \in \mathcal{K},\ \rho_m \in \mathcal{R},\ \pi_m \neq \pi_n\ \forall m \neq n\}$
    $\mathcal{A}_t \leftarrow \arg\max_{s \in \mathcal{S}} \sum_{m=1}^{M} \theta_{m,\pi_m(s),\rho_m(s)}$

  Step 3: Assign super-arm and update estimates (ACK/NACK)
    Assign $\mathcal{A}_t = \{(\pi_{1,t}, \rho_{1,t}), \ldots, (\pi_{M,t}, \rho_{M,t})\}$ with $\pi_{m,t} = (b_{m,t}, k_{m,t})$
    Observe $x_{m,t} \in \{0,1\}$ for $m = 1, \ldots, M$
  for each user $m \in [M]$ **do**
    $n_{m,\pi_{m,t},\rho_{m,t}} \leftarrow n_{m,\pi_{m,t},\rho_{m,t}} + 1$
    $\hat{\psi}_{m,\pi_{m,t},\rho_{m,t}} \leftarrow \hat{\psi}_{m,\pi_{m,t},\rho_{m,t}} + \dfrac{x_{m,t} - \hat{\psi}_{m,\pi_{m,t},\rho_{m,t}}}{n_{m,\pi_{m,t},\rho_{m,t}}}$
  end for
end for

Algorithm 3 CTS: Combinatorial Thompson Sampling

Require: M users, beam set $\mathcal{K} = [B] \times [K]$, rate set $\mathcal{R} = \{r_1 < \cdots < r_R\}$, time horizon T

Initialize:
  $A_{m,(b,k),r} \leftarrow 1, \quad B_{m,(b,k),r} \leftarrow 1$ for all $m \in [M],\ (b,k) \in \mathcal{K},\ r \in \mathcal{R}$ {Beta priors on ψ}
for $t = 1$ to T **do**
  Step 1: Thompson sampling from posterior
  for each $m \in [M],\ (b,k) \in \mathcal{K},\ r \in \mathcal{R}$ **do**
    $\tilde{\psi}_{m,(b,k),r} \sim \mathrm{Beta}(A_{m,(b,k),r}, B_{m,(b,k),r})$
    $\theta_{m,(b,k),r} \leftarrow r \cdot \tilde{\psi}_{m,(b,k),r}$ {expected throughput sample}
  end for

  Step 2: Construct super-arm via optimal assignment (no sharing)
    $\mathcal{S} = \{\{(\pi_1, \rho_1), \ldots, (\pi_M, \rho_M)\} :\ \pi_m \in \mathcal{K},\ \rho_m \in \mathcal{R},\ \pi_m \neq \pi_n\ \forall m \neq n\}$
    $\mathcal{A}_t \leftarrow \arg\max_{s \in \mathcal{S}} \sum_{m=1}^{M} \theta_{m,\pi_m(s),\rho_m(s)}$

  Step 3: Assign super-arm and update posteriors (ACK/NACK)
    Assign $\mathcal{A}_t = \{(\pi_{1,t}, \rho_{1,t}), \ldots, (\pi_{M,t}, \rho_{M,t})\}$ with $\pi_{m,t} = (b_{m,t}, k_{m,t})$
    Observe $x_{m,t} \in \{0,1\}$ for $m = 1, \ldots, M$
  for each user $m \in [M]$ **do**
    $A_{m,\pi_{m,t},\rho_{m,t}} \leftarrow A_{m,\pi_{m,t},\rho_{m,t}} + x_{m,t}$
    $B_{m,\pi_{m,t},\rho_{m,t}} \leftarrow B_{m,\pi_{m,t},\rho_{m,t}} + (1 - x_{m,t})$
  end for
end for

Algorithm 4 SAT-CTS (Full pseudo code)

Require: M users; beam set $\mathcal{K} = [B] \times [K]$; rate set $\mathcal{R} = \{r_1 < \cdots < r_R\}$; horizon T; target τ_r

Initialize:
    For all $m \in [M],\ (b,k) \in \mathcal{K},\ r \in \mathcal{R}$:
        $A_{m,(b,k),r} \leftarrow 1,\ B_{m,(b,k),r} \leftarrow 1,\ n_{m,(b,k),r} \leftarrow 0,\ s_{m,(b,k),r} \leftarrow 0$
    $\mathcal{S} = \{\{(\pi_1,\rho_1),\ldots,(\pi_M,\rho_M)\} : \pi_m \in \mathcal{K},\ \rho_m \in \mathcal{R},\ \pi_m \neq \pi_n\ \forall m \neq n\}$
    Primitives: $\mathrm{BestAssign}(\mathrm{Score}) := \arg\max_{s \in \mathcal{S}} \sum_{m=1}^{M} \mathrm{Score}_{m,\pi_m(s),\rho_m(s)}$, $\mathrm{Avg}(\mathrm{Score}, s) := \frac{1}{M} \sum_{m=1}^{M} \mathrm{Score}_{m,\pi_m(s),\rho_m(s)}$
for $t = 1$ to T **do**
    Step 1: Thompson Sampling
    for each $m, (b,k), r$ **do**
        $\tilde{\psi}_{m,(b,k),r} \sim \mathrm{Beta}(A_{m,(b,k),r}, B_{m,(b,k),r});\quad \theta_{m,(b,k),r} \leftarrow r\, \tilde{\psi}_{m,(b,k),r}$
    end for
    $\widehat{\mathcal{S}}_{\mathrm{TS}} \leftarrow \mathrm{BestAssign}(\theta)$
    Step 3: Indices (LCB/MEAN/UCB)
    for each $m, (b,k), r$ **do**
        $\hat{\psi} \leftarrow s_{m,(b,k),r} / \max(1, n_{m,(b,k),r})$; $c \leftarrow \sqrt{0.5\ \log(\max\{2,t\}) / \max(1, n_{m,(b,k),r})}$
        $\mathrm{LCB}_{m,(b,k),r} \leftarrow r \cdot \max\{0, \hat{\psi} - c\}$, $\mathrm{MEAN}_{m,(b,k),r} \leftarrow r \cdot \hat{\psi}$, $\mathrm{UCB}_{m,(b,k),r} \leftarrow r \cdot (\hat{\psi} + c)$
    end for
    $S_{\mathrm{LCB}} \leftarrow \mathrm{BestAssign}(\mathrm{LCB})$, $S_{\mathrm{MEAN}} \leftarrow \mathrm{BestAssign}(\mathrm{MEAN})$, $S_{\mathrm{UCB}} \leftarrow \mathrm{BestAssign}(\mathrm{UCB})$
    $z_L \leftarrow \mathrm{Avg}(\mathrm{LCB}, S_{\mathrm{LCB}})$, $z_M \leftarrow \mathrm{Avg}(\mathrm{MEAN}, S_{\mathrm{MEAN}})$, $z_U \leftarrow \mathrm{Avg}(\mathrm{UCB}, S_{\mathrm{UCB}})$
    Step 4: Gate

$$\mathcal{A}_t \leftarrow \begin{cases} S_{\mathrm{LCB}}, & z_L \geq \tau_r \\ S_{\mathrm{MEAN}}, & z_M \geq \tau_r \\ S_{\mathrm{UCB}}, & z_U \geq \tau_r \\ \widehat{\mathcal{S}}_{\mathrm{TS}}, & \text{otherwise} \end{cases}$$

    Step 5: Play & update (ACK/NACK semi-bandit feedback)
    Assign $\mathcal{A}_t = \{(\pi_{1,t}, \rho_{1,t}), \ldots, (\pi_{M,t}, \rho_{M,t})\}$; observe $x_{m,t} \in \{0,1\}$
    for each m **do**
        $n_{m,\pi_{m,t},\rho_{m,t}} \leftarrow n_{m,\pi_{m,t},\rho_{m,t}} + 1$; $s_{m,\pi_{m,t},\rho_{m,t}} \leftarrow s_{m,\pi_{m,t},\rho_{m,t}} + x_{m,t}$
        $A_{m,\pi_{m,t},\rho_{m,t}} \leftarrow 1 + s_{m,\pi_{m,t},\rho_{m,t}}$; $B_{m,\pi_{m,t},\rho_{m,t}} \leftarrow 1 + n_{m,\pi_{m,t},\rho_{m,t}} - s_{m,\pi_{m,t},\rho_{m,t}}$
    end for
end for