ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.9.1.P145

Distributed Multi-Armed Bandit for Satellite-Assisted Integrated Sensing and Communication in Micro-UAV Swarms

Zuohong Xu [1] *, Changhao Du [2], Zhou Zhang [3], Yizhen Jia [1],
Yingqi Jia [1],Tong Wu [1], Lihong Lv [1], Cuowu Xiruo [1]

[1] the 32008 Troops of PLA, Beijing, China;

[2] Beijing Institute of Technology, China;

[3] Hunan University, Changsha, China.

zuohong.xu@outlook.com

Abstract. This paper considers micro unmanned aerial vehicle (UAV) swarms operating in confined or infrastructure-denied urban environments, where they conduct target sensing, support local communications, and perform remote data backhaul via edge relay nodes. We propose a distributed cooperative multi-armed bandit framework for satellite-assisted integrated sensing and communication (ISAC) in micro-UAV swarms. Each UAV selects a joint configuration including sensing waveform, frequency channel, transmit power, bandwidth allocation, and modulation scheme as a bandit arm. A perturbation-augmented neighbor-fusion UCB algorithm enables UAVs to share reward statistics while avoiding co-channel clustering, and adapt to satellite handover without a central scheduler. Simulation results show that the proposed method reaches the fastest convergence and handover recovery among UCB-based distributed bandit methods, and that cooperative neighbor fusion yields statistically significant gains over independent learning in convergence speed, throughput, ISAC interference reduction, and handover recovery.

Keywords: Integrated sensing and communication; Micro-UAV swarm; LEO satellite; Multi-armed bandit; Distributed resource allocation; Online learning.

1. Introduction

Micro unmanned aerial vehicles (UAVs) are increasingly deployed in confined, GPS-denied, and infrastructure-denied environments such as urban buildings, tunnels, and post-disaster areas [1]. In these scenarios, UAVs must detect weak and slow-moving targets, maintain communication links, and support real-time command delivery without prior map information [2]. Integrated sensing and communication (ISAC) lets a UAV share spectrum and hardware between target sensing and data transmission [3], but the two functions then contend for the same resources and interfere with each other [4]. When terrestrial infrastructure is unavailable, low-Earth-orbit (LEO) satellite links can provide a resilient relay channel to remote command centers [5, 6], yet satellite-assisted UAV networks face time-varying visibility, long propagation delay, and handover events.

Unlike conventional UAV-satellite communication scenarios that primarily optimize backhaul throughput, micro-UAV missions in confined environments require joint consideration of sensing reliability, local communication, and command latency. A configuration that maximizes satellite throughput may be unsuitable if it degrades target detection or increases sensing-communication interference. Therefore, the resource allocation problem should be formulated as a joint ISAC-backhaul decision-making task rather than a pure satellite uplink optimization.

These uncertainties make offline optimization unsuitable. Multi-armed bandit learning offers a lightweight online decision-making framework that balances exploration and exploitation with limited computation [7]. This paper proposes a satellite-assisted ISAC cooperative multi-armed bandit (SA-ISAC-MAB) framework for micro-UAV swarms. The main contributions are:

(1) We formulate satellite-assisted micro-UAV ISAC as a distributed cooperative bandit problem with joint sensing-communication-backhaul configuration selection.

(2) We design a composite reward that weights detection probability, throughput, latency, ISAC interference, and co-channel interference.

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.9.1.P145

(3) We propose a perturbation-augmented neighbor-fusion UCB algorithm with satellite handover reset, which prevents co-channel clustering while accelerating cooperative learning without a central scheduler.

(4) We validate the method with paired-environment Monte Carlo runs and 95% confidence intervals. The results show statistically significant gains over independent learning and the fastest convergence and handover recovery among UCB-based distributed bandit methods.

ISAC enables simultaneous sensing and communication using shared spectrum [3], and has been applied to UAV platforms [4]. However, existing work assumes known channel statistics and centralized control. LEO constellations provide coverage for UAVs in infrastructure-denied areas [6], with resource allocation studied via optimization [8] and deep reinforcement learning [9], but these require significant computation. Cooperative bandit algorithms let several agents share information and speed up learning [10, 11]; representative methods include Thompson Sampling [12] and sliding-window UCB [13]. Our work extends cooperative bandits to satellite-assisted ISAC and adds a channel-level perturbation mechanism that addresses the co-channel clustering problem of deterministic cooperative policies.

2. System Model

2.1 Hybrid Confined-Edge Architecture.

We consider a hybrid micro-UAV swarm consisting of sensing UAVs and relay UAVs in an infrastructure-denied environment. Sensing UAVs operate inside or near confined structures and perform local ISAC-based target detection. Due to blockage and limited antenna aperture, these sensing UAVs may not maintain direct satellite connectivity. Instead, sensing results are forwarded to relay UAVs located near building edges, rooftops, or open-air positions. These relay UAVs provide intermittent LEO satellite backhaul to a remote command center. The system thus operates in three layers: (1) confined-space sensing, (2) local cooperation via short-range side links, and (3) satellite backhaul via LEO satellite at altitude 550 km. At any given time, a satellite s is visible with elevation angle $\theta_s(t)$. The network operates in discrete time steps Δt.

2.2 ISAC-Backhaul Configuration as Bandit Arm.

At time step t, each UAV n selects a joint ISAC-backhaul configuration from $A = \{a_1, ..., a_K\}$, where:

$$a_k = (w_k, f_k, p_k, b_k, m_k) \tag{1}$$

$w_k \in \{0,1\}$ is the sensing-communication waveform type (sensing-focused or ISAC-balanced), $f_k \in \{1, ..., F\}$ the frequency channel, p_k the transmit power level, b_k the bandwidth split among sensing, local communication, and backhaul, and m_k the modulation scheme. The total number of arms is $K = 2 \times F \times P \times B \times M$.

2.3 Environmental State Observation.

Each UAV n observes an environmental state vector:

$$z_n(t) = [\theta_s(t), \mathrm{SNR}_n(t), M_n(t), I_n(t)] \tag{2}$$

where $\theta_s(t)$ is the satellite elevation angle, $\mathrm{SNR}_n(t)$ the local signal-to-noise ratio, $M_n(t)$ the multipath intensity, and $I_n(t)$ the estimated ISAC interference level. These features are locally measurable and used for reward computation and handover detection.

2.4 Composite Reward Function.

After selecting arm $a_n(t)$, UAV n receives a composite reward:

$$r_n(t) = \omega_1 \overline{D}_n(t) + \omega_2 \overline{T}_n(t) - \omega_3 \overline{\Lambda}_n(t) - \omega_4 \overline{I}_n^{\mathrm{ISAC}}(t) - \omega_5 \overline{I}_n^{\mathrm{co}}(t) + \omega_6 \, 1\{\theta_s > \theta_{\min}\} \tag{3}$$

where $\overline{D}_n(t)$, $\overline{T}_n(t)$, $\overline{\Lambda}_n(t)$, $\overline{I}_n^{\mathrm{ISAC}}(t)$, and $\overline{I}_n^{\mathrm{co}}(t)$ are the normalized (to [0,1]) target detection probability, satellite backhaul throughput, command access latency, ISAC mutual interference, and

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.9.1.P145
co-channel interference, respectively. The weights $\omega_i > 0$ reflect mission-level preferences and can be adjusted according to whether sensing accuracy, command latency, or backhaul throughput is prioritized. The detection probability is modeled as:

$$D_n(t) = \sigma(\alpha_0 + \alpha_1 \mathrm{SNR}_n(t) - \alpha_2 M_n(t) - \alpha_3 I_n^{\mathrm{ISAC}}(t)) \quad (4)$$

where $\sigma(\cdot)$ is the sigmoid function that models how detection depends on sensing SNR, multipath, and ISAC interference.

2.5 Interference Model.

The co-channel interference experienced by UAV n is:

$$I_n^{\mathrm{co}}(t) = \sum_{j \neq n,\, f_j = f_n} \frac{g_{js}(t) p_j(t)}{L_{js}(t)} \quad (5)$$

where $g_{js}(t)$ is the channel gain from UAV j to satellite s, $p_j(t)$ the transmit power, and $L_{js}(t)$ the path loss. This coupling means each UAV's optimal action depends on others' choices.

2.6 Global Objective.

$$\max \sum_{t=1}^{T} \sum_{n=1}^{N} r_n(t) \quad (6)$$

with group regret $\mathrm{GR}(T) = T \sum_n r_n^* - \sum_t \sum_n r_n(t)$, where r_n^* is the expected reward of the optimal arm for UAV n.

3. Distributed Cooperative Bandit Algorithm

3.1 Algorithm Design.

Each UAV n maintains empirical mean estimates $\hat{\mu}_{n,a}(t)$ and pull counts $N_{n,a}(t)$ for each arm. A key challenge in cooperative bandits is that deterministic policies such as standard UCB produce correlated arm selections when agents share similar reward estimates, leading to co-channel clustering and increased interference. To address this, we augment the UCB index with a channel-level random perturbation that breaks decision correlation while preserving structured exploration:

$$a_n(t) = \arg\max \left(\hat{\mu}_{n,a}(t) + c\sqrt{\frac{\ln t}{N_{n,a}(t)}} + \varepsilon_{n,f(a)}(t)\right) \quad (7)$$

where $f(a)$ denotes the channel of arm a, and $\varepsilon_{n,k}(t) \sim \mathrm{N}(0, \xi^2)$ is an independent zero-mean Gaussian perturbation drawn anew by each UAV for each channel at every time step. Since the perturbation is independent across UAVs, neighboring agents with identical UCB indices will select different channels with non-zero probability. This promotes channel diversity without explicit coordination. The perturbation scale ξ is chosen so that the perturbation is small relative to the $O\left(\sqrt{\frac{\ln t}{N_{n,a}(t)}}\right)$ exploration bonus in the exploitation regime; we validate this scale by a sensitivity analysis over ξ. After observing reward $r_n(t)$, UAV n broadcasts $(a_n(t), r_n(t))$ to neighbors via side links and fuses collected samples:

$$\hat{\mu}_{n,a}(t) \leftarrow \frac{1}{|S_{n,a}(t)|} \sum_{(j,\, r_j)}^{S_{n,a}(t)} r_j \quad (8)$$

where $S_{n,a}(t)$ contains reward samples for arm a from UAV n and its neighbors. The perturbation and neighbor fusion work complementarily: the perturbation prevents channel clustering while fusion accelerates reward estimation. When a satellite handover occurs (detected via beacon signal loss), all UAVs reset their statistics to re-learn for the new satellite channel.

Algorithm 1 SA-ISAC-MAB for Satellite-Assisted Micro-UAV ISAC
Require: arm set A, exploration parameter c, perturbation scale ξ, horizon T
Ensure: ISAC-backhaul policy π_n for each UAV $n \in N$
1: **Initialization (each UAV n in parallel):**
2: **for** a = 1 **to** K **do**
3: Select arm a; observe reward r_n; set $\hat{\mu}_{n,a} \leftarrow r_n$, $N_{n,a} \leftarrow 1$

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.9.1.P145

Broadcast (a, r_n) to neighbors; fuse samples via Eq. (8)
end for
Main Loop (each UAV n in parallel):
for t = K+1 **to** T **do**
if satellite handover detected **then**
Reset $\hat{\mu}_{n,a} \leftarrow 0$, $N_{n,a} \leftarrow 0$ for all a; go to Init
end if
Observe environmental state $z_n(t)$; compute reward-related metrics
Draw $\varepsilon \sim N(0, \xi^2)$ for each channel k
Compute $UCB_{n,a}(t) = \hat{\mu}_{n,a} + c\sqrt{(\ln t / N_{n,a})} + \varepsilon_{n,f}(a)$ for all a
Select $a_n(t) = \arg\max_a UCB_{n,a}(t)$
Perform ISAC operation; transmit via satellite using configuration $a_n(t)$
Observe composite reward $r_n(t)$
$N_{n,a_n}(t) \leftarrow N_{n,a_n}(t) + 1$
Broadcast $(a_n(t), r_n(t))$ to neighbors; fuse samples via Eq. (8)
end for
return Local policy π_n

3.2 Regret Analysis.

Under the idealized assumptions of homogeneous agents, decoupled rewards, and a bounded perturbation, the cooperative UCB framework of [10] yields sub-linear group regret:

$$\mathrm{GR}(T) = \tilde{O}(\sqrt{NKT}) + O(HK) \quad (9)$$

where N is the UAV count, K the arm count, T the horizon, and H the handover count. The first term reflects cooperative exploration benefit; the second captures the fixed re-initialization cost after H handover resets. We note that this bound is inherited from [10]: our model also adds a Gaussian (unbounded) perturbation, a piecewise-stationary elevation $\theta_s(t)$, and a coupled reward through co-channel interference, so we do not extend the guarantee to the full SA-ISAC-MAB setting here. In practice, the perturbation scale ξ is set so that the perturbation is small relative to the exploration bonus in the exploitation regime, bounding its empirical effect on the regret constant.

4. Simulation Results

4.1 Simulation Setup.

We simulate N=8 micro-UAVs in a hybrid confined-edge scenario with LEO satellite connectivity at 550 km. Each UAV selects from K=12 arms (2 waveform types × 3 channels × 2 power levels). Channel model includes free-space path loss, Rayleigh fading, and co-channel interference (factor 0.14 per same-channel UAV). Detection follows the sigmoid model of Eq. (4) with $(\alpha_0, \alpha_1, \alpha_2, \alpha_3) = (0.5, 0.85, 0.1, 0.15)$. The composite reward of Eq. (3) is computed with all components normalized to [0,1] and equal weights $\omega_i=1$. Each UAV is connected to 3 neighbors via short-range side links in a ring-cross topology. The oracle assigns each UAV its optimal arm under round-robin channel allocation. The time horizon is T=2000 with one handover at t=1000. The perturbation scale is $\xi = 0.05$, validated by a sensitivity analysis over ξ.

We use a paired evaluation protocol. All schemes share an identical environment realization per Monte Carlo run (a common random seed drives the reward noise), while algorithm-internal randomness (exploration draws, perturbation, Thompson sampling) uses a scheme-specific seed. This paired design allows paired Wilcoxon signed-rank significance tests between any two schemes on the same 100 runs, which independent seeds per scheme would not support. Results report the mean together with 95% confidence intervals. Seven schemes are compared: (1) SA-ISAC-MAB (proposed); (2) Independent UCB; (3) Thompson Sampling; (4) ε-Greedy; (5) SW-UCB; (6) Greedy; (7) Fixed Allocation.

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.9.1.P145

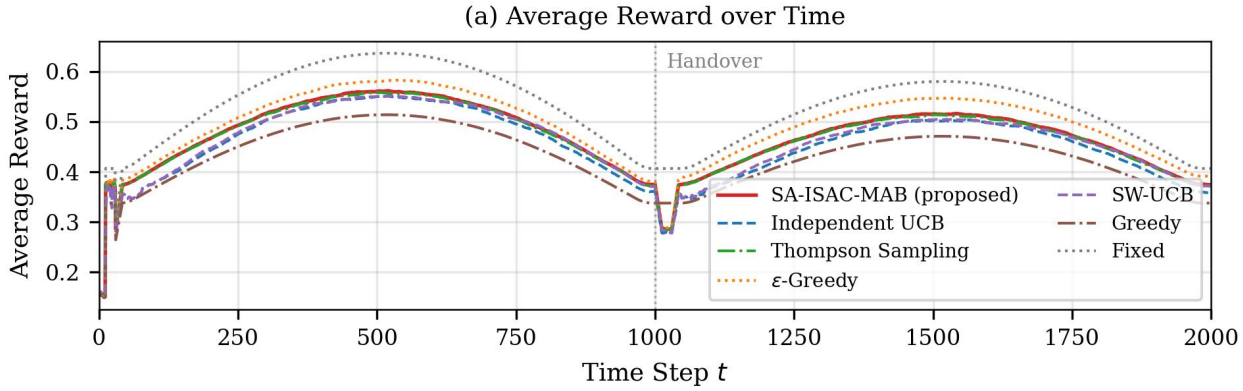


Fig. 1 Average reward over time. The vertical line marks satellite handover at t=1000. SA-ISAC-MAB converges faster than Independent UCB and SW-UCB through cooperative neighbor fusion and channel-level perturbation, and is statistically indistinguishable from Thompson Sampling on convergence speed.

Table 1. Performance Comparison (N=8, T=2000, 100 paired runs, mean)

Scheme	Converge	Detect.(%)	Tput.(%)	ISAC Intf.	Recovery
SA-ISAC-MAB (proposed)	250 rds	63.8	79.0	0.118	319 rds
Independent UCB	281 rds	63.7	74.4	0.131	461 rds
Thompson Sampling	252 rds	63.9	80.5	0.121	324 rds
ε-Greedy	225 rds	64.3	87.8	0.114	221 rds
SW-UCB	277 rds	64.0	80.8	0.122	427 rds
Greedy	504 rds	64.3	70.8	0.160	600 rds
Fixed Allocation	164 rds	64.3	100.0	0.088	146 rds

Convergence/Recovery measured in rounds from the end of the K-arm initialization. Fixed Allocation uses the oracle assignment by construction, hence its near-100% throughput and trivially fast metrics.

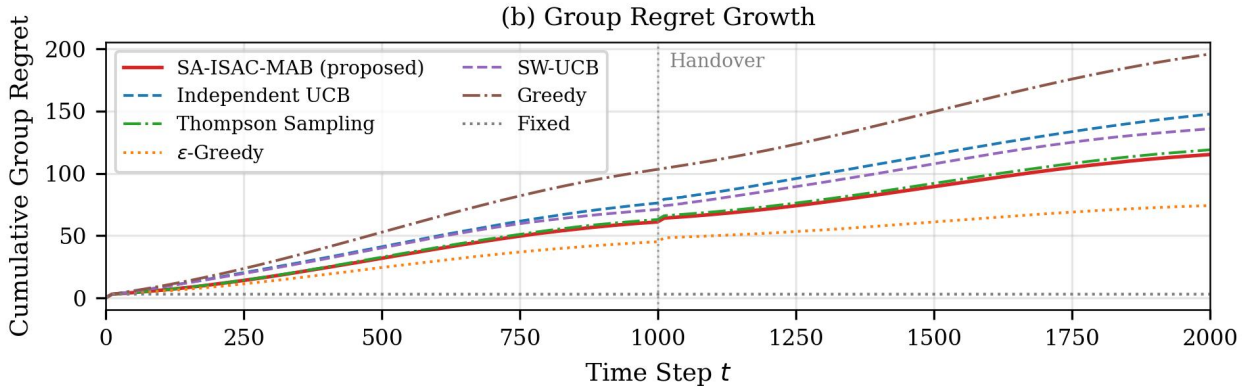


Fig. 2 Cumulative group regret relative to the oracle. SA-ISAC-MAB maintains the lowest regret among UCB-based schemes.

The proposed method does not dominate every metric, and we report the comparison in full rather than selecting only favorable ones. On raw throughput, ε-Greedy (87.8%) clearly leads, followed by SW-UCB (80.8%) and Thompson Sampling (80.5%), all above SA-ISAC-MAB (79.0%); ε-Greedy also recovers fastest (221 rounds). We therefore scope the claims carefully. Relative to Independent UCB, the natural baseline that isolates the cooperative gain, SA-ISAC-MAB improves every metric with high statistical significance (paired Wilcoxon $p < 10^{-4}$ on convergence, recovery, throughput, and ISAC interference). Among the UCB-based methods (Independent UCB, SW-UCB, and the proposed method), SA-ISAC-MAB reaches the fastest convergence and recovery. We do not claim superiority through a "provable-regret" filter: Thompson Sampling itself carries a $\tilde{O}(\sqrt{KT})$ bound [12]

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.9.1.P145
and SW-UCB a piecewise-stationary bound [13], so excluding them on those grounds would be incorrect. The practical case for SA-ISAC-MAB is that it approaches the throughput of TS and SW-UCB while converging and recovering faster than Independent UCB, without any exploration-rate tuning.

Detection probability is the one metric on which the schemes barely differ. All learning-based schemes fall in the range 63.7%–64.3%, because detection is dominated by the sigmoid response to SNR, which varies only modestly across arm configurations. We acknowledge that this dimension is nearly non-discriminative under the present parameterization; sharpening the SNR-to-detection sensitivity would expose the sensing–communication trade-off more clearly, and we leave that to later work.

Where the channel-level perturbation shows up is interference and recovery. SA-ISAC-MAB reaches an ISAC interference of 0.118, a 9.9% reduction versus Independent UCB (0.131) that is significant ($p < 10^{-4}$): by decorrelating neighboring selections, the perturbation spreads UAVs across channels and lowers sensing-communication conflict. ε-Greedy attains a marginally lower interference (0.114), consistent with its forced uniform exploration, but among the UCB-based methods SA-ISAC-MAB has the lowest value. After handover, SA-ISAC-MAB recovers within 319 rounds, 30.8% faster than Independent UCB (461 rounds) and faster than SW-UCB (427 rounds); ε-Greedy recovers fastest (221 rounds) for the same forced-exploration reason, and Greedy, which lacks exploration, recovers slowest (600 rounds).

5. Conclusion

This paper proposes a distributed cooperative multi-armed bandit framework for satellite-assisted ISAC in micro-UAV swarms. Each UAV independently learns its ISAC-backhaul configuration using perturbation-augmented UCB decisions, neighbor reward sharing, and satellite handover reset. Random perturbation mitigates co-channel clustering, while neighbor fusion accelerates reward estimation. Experiments show that SA-ISAC-MAB significantly outperforms independent learning and other UCB-based methods in convergence and recovery, achieving throughput close to Thompson Sampling and SW-UCB.

References

[1] Shakhatreh H, Sawalmeh A H, Al-Fuqaha A, et al. Unmanned aerial vehicles (UAVs): A survey on civil applications and key research challenges. IEEE Access, 2019, 7: 48572–48634. DOI: 10.1109/ACCESS.2019.2909530.

[2] Mozaffari M, Saad W, Bennis M, et al. A tutorial on UAVs for wireless networks: Applications, challenges, and open problems. IEEE Commun. Surveys & Tutorials, 2019, 21(3): 2334–2360. DOI: 10.1109/COMST.2019.2902862.

[3] Liu F, Cui Y, Masouros C, et al. Integrated sensing and communications: Toward dual-functional wireless networks for 6G and beyond. IEEE J. Sel. Areas Commun., 2022, 40(6): 1728–1767. DOI: 10.1109/JSAC.2022.3156632.

[4] Yao Y, Dong D, Cai C, et al. Multi-UAV-assisted Internet of Remote Things communication within satellite–aerial–terrestrial integrated network. EURASIP J. Adv. Signal Process., 2024, 2024(1): 10. DOI: 10.1186/s13634-023-01101-3.

[5] Azari M M, Rosas F, Chen K C, et al. Ultra reliable UAV communication using altitude and cooperation diversity. IEEE Trans. Commun., 2018, 66(1): 330–342. DOI: 10.1109/TCOMM.2017.2757506.

[6] Kodheli O, Lagunas E, Maturo N, et al. Satellite communications in the new space era: A survey and future challenges. IEEE Commun. Surveys & Tutorials, 2021, 23(1): 70–109. DOI: 10.1109/COMST.2020.9210567.

[7] Lattimore T, Szepesvári C. Bandit Algorithms. Cambridge University Press, 2020. DOI: 10.1017/9781108571401.

ISSN: 3007-9640 DOI: https://doi.org/10.70114/acmsr.2026.9.1.P145
[8] Zhang M, Yang X, Bu Z. Beam-hopping-based resource allocation in integrated satellite-terrestrial networks. Sensors, 2024, 24(14): 4699. DOI: 10.3390/s24144699.

[9] Huang C, Chen X, Chen G, et al. Deep reinforcement learning-based resource allocation for hybrid bit and generative semantic communications in space-air-ground integrated networks. arXiv preprint, 2024: arXiv:2412.05647 (accepted by IEEE J. Sel. Areas Commun.). DOI: 10.48550/arXiv.2412.05647.

[10] Landgren P, Srivastava V, Leonard N E. Distributed cooperative decision-making in multiarmed bandits: Frequentist and Bayesian algorithms. Proc. IEEE CDC, 2016: 166–171. DOI: 10.1109/CDC.2016.7798264.

[11] Avner O, Mannor S. Multi-user communication networks: A coordinated multi-armed bandit approach. IEEE/ACM Trans. Networking, 2019, 27(6): 2192–2207. DOI: 10.1109/TNET.2019.2935043.

[12] Agrawal S, Goyal N. Analysis of Thompson Sampling for the multi-armed bandit problem. Proc. COLT, 2012: 39.1–39.26.

[13] Garivier A, Moulines E. On upper-confidence bound policies for switching bandit problems. Proc. ALT, 2011: 174–188.