Multi-Armed Bandit Learning for ISAC Systems in Time-Varying Environments

Multi-Armed Bandit Learning for ISAC Systems in Time-Varying Environments

Giorgio Taricco, *Fellow, IEEE*

***Abstract*—Integrated sensing and communication (ISAC) systems aim to jointly perform data transmission and environmental sensing, leading to coupled and often conflicting design objectives. This paper investigates a lightweight online learning framework for ISAC based on multi-armed bandit (MAB) algorithms. A finite beam codebook is used for sequential beam selection, while the sensing–communication trade-off is captured through a normalized reward combining communication rate and a CRLB-based sensing metric. To address time-varying environments, a sliding-window UCB strategy is adopted for beam selection and the trade-off parameter is adapted online on a slower timescale. Numerical results under abrupt changes, gradual drift, and multiple change points show that the proposed approach tracks environmental variations and improves dynamic-regret performance over fixed-α and stationary baselines.**

***Index Terms*—Integrated sensing and communication (ISAC), multi-armed bandits, online learning, beamforming.**

I. Introduction

The rapid evolution of wireless networks toward sixth-generation (6G) systems is driving the convergence of communication and sensing functionalities into a unified framework known as Integrated Sensing and Communication (ISAC) [1]. By enabling a single platform to perform data transmission and environmental sensing, ISAC improves spectral efficiency, reduces hardware redundancy, and supports applications such as autonomous driving, smart cities, and human activity recognition. Despite its potential, ISAC introduces significant design challenges. Communication and sensing objectives are inherently coupled and often conflicting. For instance, beamforming strategies that maximize communication rate may be suboptimal for sensing accuracy, while radar-oriented waveforms may degrade communication performance. This leads to a fundamental trade-off that must be dynamically managed. The problem is further complicated by uncertainty and time variability in wireless environments. Channels are affected by fading, interference, and mobility, while sensing performance depends on unknown and evolving target parameters. Many conventional optimization approaches rely on accurate models and slowly varying assumptions, which are often unavailable or impractical in real-time ISAC systems. Learning-based approaches provide a promising alternative. In particular, Multi-Armed Bandit (MAB) algorithms offer a principled framework for sequential decision-making under uncertainty [2], [3]. In the ISAC context, transmission configurations such as beamforming vectors, waveforms, or power levels can be treated as bandit arms, while observed communication and sensing metrics define the reward, enabling online adaptation without explicit environmental models.

In this paper, we focus on adaptive beam selection in time-varying ISAC environments with a finite beam codebook and fixed waveform and transmit power. The proposed framework uses a sliding-window bandit algorithm for non-stationary beam selection and a slower outer adaptation of the sensing–communication trade-off parameter. A normalized reward is used so that the communication and sensing terms can be combined consistently despite their different scales.

A. Related Literature

Multi-armed bandit (MAB) methods have been investigated for several ISAC and beam-management problems. They have been applied to ISAC optimization, including user–target pairing and interference management [4], beam selection using contextual bandits with deep learning and sensing side information [5], beam alignment and tracking in mmWave systems [6], and radar-assisted beam selection via sensing-aided search-space reduction [7]. Beyond bandit formulations, model-based and reinforcement learning approaches have been proposed for ISAC waveform and resource optimization [8], while related MAB techniques address beam tracking and rate adaptation in directional wireless systems mainly for communication performance [9].

In contrast, this paper embeds a CRLB-based sensing objective into a non-stationary bandit formulation and studies online adaptation of both the beam index and the trade-off parameter. The emphasis is on a lightweight model-free method suitable for real-time implementation.

II. System Model

We consider a monostatic ISAC system operating over discrete time slots $t = 1, \dots, T$, where a transmitter simultaneously performs communication and sensing.

As illustrated in Fig. 1, the monostatic ISAC transmitter serves both communication and sensing functions within the same hardware platform. At each slot, the system selects an action $a_t \in \mathcal{A}$ corresponding to a transmission configuration defined by a beamforming vector $\boldsymbol{w}_t \in \mathbb{C}^{N_t}$, a waveform $s_t(n)$, and transmit power P_t. The transmitted signal is given by

$$\boldsymbol{x}_t(n) = \sqrt{P_t}\, \boldsymbol{w}_t\, s_t(n). \tag{1}$$

The received signal at the communication user is

$$y_t(n) = \boldsymbol{h}_t^{\mathsf{H}} \boldsymbol{x}_t(n) + z_t(n), \tag{2}$$

G. Taricco is with the Department of Electrical and Telecommunications Engineering, Politecnico di Torino, 10129 Torino, Italy.
E-mail: giorgio.taricco@polito.it.

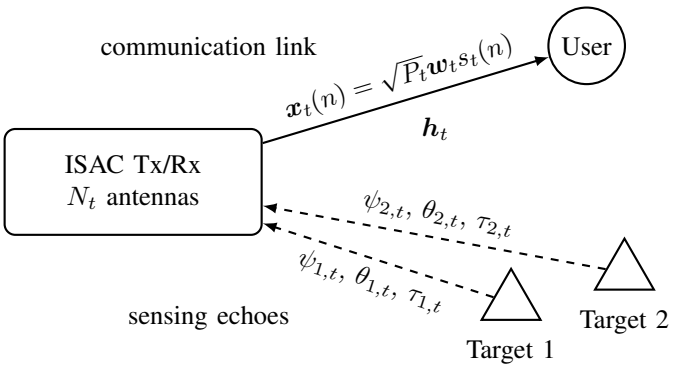

Fig. 1. Illustration of the monostatic ISAC system model. The transmitter sends a communication waveform toward the user while simultaneously probing the environment and receiving echoes from surrounding targets.

where $\boldsymbol{h}_t$ denotes the channel vector and $z_t(n)$ is Gaussian additive noise with distribution $\mathcal{CN}(0, \sigma^2)$. The corresponding achievable rate is

$$R_t^{\mathsf{comm}} = \log_2\left(1 + \frac{P_t|\boldsymbol{h}_t^{\mathsf{H}}\boldsymbol{w}_t|^2}{\sigma^2}\right). \tag{3}$$

For sensing, the transmitted signal is reflected by the targets back to the transmitter. Let $\boldsymbol{a}(\theta) \in \mathbb{C}^{N_t}$ denote the transmit array steering vector associated with spatial angle θ. Considering a linear array with element positions x_k and wavenumber $\kappa \triangleq 2\pi/\lambda$, the steering vector is

$$\boldsymbol{a}(\theta) = \frac{1}{\sqrt{N_t}}\left(e^{j\kappa x_1 \sin\theta}, e^{j\kappa x_2 \sin\theta}, \ldots, e^{j\kappa x_{N_t} \sin\theta}\right)^{\mathsf{T}}. \tag{4}$$

The sensing observation is modeled as

$$\boldsymbol{y}_t^{\mathsf{sense}}(n) = \sum_{k=1}^{K_t} \psi_{k,t}\boldsymbol{a}(\theta_{k,t})\boldsymbol{a}^{\mathsf{H}}(\theta_{k,t})\boldsymbol{x}_t(n-\tau_{k,t}) + \boldsymbol{n}_t(n), \tag{5}$$

where $\psi_{k,t}$, $\theta_{k,t}$, and $\tau_{k,t}$ represent the reflection coefficient, angle, and delay of the kth target, respectively. The sensing performance is quantified using a CRLB-based metric derived from the Fisher information matrix (FIM). Let $\boldsymbol{\theta}_t = (\theta_{1,t}, \ldots, \theta_{K_t,t})^{\mathsf{T}}$ denote the vector of target angles. The FIM is approximated by a diagonal matrix

$$\boldsymbol{J}_t = \mathrm{diag}(J_{1,t}, \ldots, J_{K_t,t}), \tag{6}$$

where each diagonal element

$$J_{k,t} = \frac{2}{\sigma_s^2}|\boldsymbol{w}_t^{\mathsf{H}}\boldsymbol{a}'(\theta_{k,t})|^2 \tag{7}$$

captures the sensitivity of the received signal with respect to the angle $\theta_{k,t}$ (here, $\boldsymbol{a}'(\theta)$ denotes the derivative of the steering vector with respect to the angle and σ_s^2 is the sensing noise variance). This approximation neglects cross-terms between different targets and provides a tractable surrogate for the estimation accuracy. Therefore, the sensing term should be interpreted as a tractable beam-dependent surrogate rather than as an exact multi-target CRLB in the presence of target coupling. A scalar sensing metric is then defined as

$$R_t^{\mathsf{sense}} = -\mathrm{tr}(\boldsymbol{J}_t^{-1}), \tag{8}$$

so that lower estimation variance corresponds to higher reward.

Since R_t^{comm} and R_t^{sense} have different scales and physical units, they are normalized before aggregation using fixed bounds associated with the considered beam codebook and system setup. For a generic metric $x_t(a)$ defined on $\mathcal{A}$, let x_{min} and x_{max} denote fixed lower and upper normalization bounds, and define

$$x_t(a) \mapsto \bar{x}_t(a) \triangleq \frac{x_t(a) - x_{\mathsf{min}}}{x_{\mathsf{max}} - x_{\mathsf{min}} + \varepsilon}, \tag{9}$$

where $\varepsilon > 0$ avoids division by zero. This normalization is applied to both $R_t^{\mathsf{comm}}(a)$ and $R_t^{\mathsf{sense}}(a)$, with separate bounds for the two metrics. The overall reward is then written as

$$r_t = \alpha_t \bar{R}_t^{\mathsf{comm}}(a_t) + (1-\alpha_t)\bar{R}_t^{\mathsf{sense}}(a_t), \tag{10}$$

where $\alpha_t \in [0,1]$ controls the sensing–communication trade-off. The system operates under uncertainty due to time-varying channels, unknown target parameters, and noise. Consequently, the reward associated with each action is stochastic and may evolve over time. The action space $\mathcal{A}$ is assumed finite, enabling a direct bandit formulation with the objective of maximizing the expected cumulative reward while adapting to environmental changes.

III. Bandit Formulation and Learning Algorithm

The beam-selection problem is modeled as a non-stationary stochastic multi-armed bandit over the finite beam set $\mathcal{A}$. Each arm corresponds to a transmission configuration. In the general formulation, a transmission configuration is defined by the tuple $u \equiv (\boldsymbol{w}, s, P)$, where $\boldsymbol{w} \in \mathbb{C}^{N_t}$ is the beamforming vector, s is the transmit waveform, and P is the transmit power. In the numerical study, s and P are fixed, so each bandit arm $a \in \mathcal{A}$ corresponds only to one beamforming vector from the finite codebook. When beam a_t is selected at time t, the system observes a stochastic normalized reward of the form

$$r_t = \mu_{a_t,\alpha_t,t} + \nu_t, \tag{11}$$

where $\mu_{a,\alpha,t}$ is the time-varying expected reward associated with beam a and trade-off value α, and ν_t accounts for channel fading, sensing uncertainty, and noise. The corresponding expected reward is

$$\mu_{a,\alpha,t} = \mathbb{E}\left[\alpha\bar{R}_t^{\mathsf{comm}}(a) + (1-\alpha)\bar{R}_t^{\mathsf{sense}}(a)\right]. \tag{12}$$

Since the environment is time varying, the learning performance is evaluated through the dynamic regret

$$R_T^{\mathsf{dyn}} = \sum_{t=1}^{T}\left(\mu_t^* - \mathbb{E}[r_t]\right), \qquad \mu_t^* = \max_{a\in\mathcal{A},\,\alpha\in\Gamma}\mu_{a,\alpha,t}, \tag{13}$$

where $\Gamma \subset [0,1]$ denotes the admissible set of trade-off values. Since α_t is selected online by the controller, the dynamic oracle is defined over the product action space $\mathcal{A} \times \Gamma$.

To adapt to non-stationarity, a sliding-window upper-confidence-bound (SW-UCB) algorithm is used for beam selection. For each beam a, let $N_a^W(t)$ and $\hat{\mu}_a^W(t)$ denote, respectively, the number of selections and the empirical average reward over the most recent W slots. The beam selected at slot t maximizes the sliding-window UCB index among beams with positive window count, while beams with zero window

Algorithm 1 Two-timescale SW-UCB for adaptive beam selection

Initialize the beam codebook $\mathcal{A}$, the trade-off set Γ, the window length W, and the outer update period B.
For initialization, select each beam once using a fixed trade-off parameter α_{init} and record the corresponding normalized rewards.
for $t = 1, 2, \ldots, T$ **do**
  if $t \in \{1, B+1, 2B+1, \ldots\}$ **then**
    Compute $N_\alpha^B(t)$ and $\hat{\mu}_\alpha^B(t)$ for all $\alpha \in \Gamma$.
    if there exists $\alpha \in \Gamma$ such that $N_\alpha^B(t) = 0$ **then**
      Select one such trade-off value α_t.
    else
      Set $\alpha_t = \arg\max_{\alpha\in\Gamma}\left(\hat{\mu}_\alpha^B(t) + \sqrt{\frac{2\log b(t)}{N_\alpha^B(t)}}\right)$.
    end if
  else
    Set $\alpha_t = \alpha_{t-1}$.
  end if
  Compute $N_a^W(t)$ and $\hat{\mu}_a^W(t)$ for all $a \in \mathcal{A}$.
  if there exists $a \in \mathcal{A}$ such that $N_a^W(t) = 0$ **then**
    Select one such beam a_t.
  else
    Select $a_t = \arg\max_{a\in\mathcal{A}}\left(\hat{\mu}_a^W(t) + \sqrt{\frac{2\log t}{N_a^W(t)}}\right)$.
  end if
  Observe $R_t^{\mathsf{comm}}(a_t)$ and $R_t^{\mathsf{sense}}(a_t)$, normalize them, and form $r_t = \alpha_t \bar{R}_t^{\mathsf{comm}}(a_t) + (1-\alpha_t)\bar{R}_t^{\mathsf{sense}}(a_t)$.
  Update the sliding-window statistics of the selected beam and the current α_t.
end for

count are explored first. The trade-off parameter is adapted online over the finite admissible set Γ. An outer SW-UCB learner updates α_t every B slots using the previous block-average reward, while the inner learner updates the beam every slot. For beam selection, let $N_a^W(t)$ denote the number of times beam a was selected in the most recent W slots before time t, and let $\hat{\mu}_a^W(t)$ denote the corresponding empirical mean normalized reward. At the outer timescale, let $N_\alpha^B(t)$ denote the number of previous blocks in which trade-off value α was selected, let $\hat{\mu}_\alpha^B(t)$ denote the corresponding empirical mean block reward, and let $b(t)$ denote the current block index. Arms with zero count in the relevant statistics are explored before applying the corresponding UCB index. This structure reflects that beam selection reacts to faster environmental variations, whereas α_t is updated on a slower system-level timescale.

Standard alternatives such as ε-greedy, stationary UCB, Thompson sampling, contextual bandits, and discounted or sliding-window variants are well known in the MAB literature [10]–[14]. Here we focus on SW-UCB because of its low complexity and its ability to discount outdated observations in piecewise-stationary environments.

IV. Numerical Results

We consider a monostatic ISAC system with $N_t = 8$ transmit antennas and a finite codebook of $K = 25$ beams uniformly located from -60° to $+60^\circ$ in 5° steps. The waveform and transmit power are fixed, so each arm corresponds to one beamforming direction. The horizon is $T = 2000$, the baseline SNR is 10 dB, the sliding-window length is $W = 120$, and the outer update period is $B = 20$ slots. The trade-off parameter is adapted online from the finite admissible set $\Gamma = \{0.1, 0.2, \ldots, 0.9\}$. The communication channel and target angles follow von Mises models with concentration parameters $\xi_c = 205$ and $\xi_s = 821$, corresponding approximately to angular spreads of 4° and 2°. We consider three non-stationary scenarios: (i) an abrupt change at $t = 1000$, where the channel mean changes from -10° to 22°, and the target means change from $(-18^\circ, 12^\circ)$ to $(8^\circ, 28^\circ)$; (ii) a gradual drift over $[800, 1200]$ between the same values; and (iii) changes at $t = 700$ and $t = 1400$, where the channel and target means follow the same transitions forward and backward. The baseline comparisons are averaged over 16 Monte Carlo trials using common random realizations across different algorithms, while the parameter sweeps use 10 trials.

Figure 2 shows the learning behavior in the abrupt-change scenario. The left panel reports the average cumulative dynamic regret for the proposed SW-UCB with adaptive α_t, fixed-α SW-UCB baselines, and stationary UCB. The adaptive scheme achieves the lowest cumulative dynamic regret and outperforms stationary UCB after the change point. The right panel shows that α_t evolves over time instead of collapsing to a fixed value, supporting the outer adaptation loop.

Figure 3 summarizes the performance corresponding to the three non-stationary scenarios for $W = 120$. The left panel reports the cumulative dynamic regret at $t = T$, while the right panel reports the mean normalized reward over the last 200 slots. The adaptive-α_t method yields the best or nearly best regret in all scenarios and the highest late-time reward, showing that the adaptive trade-off remains beneficial under abrupt changes, gradual drift, and repeated transitions.

Figure 4 reports sensitivity to the main simulation parameters in the abrupt-change scenario. The left and middle panels show the window and SNR sweeps with raw values and smoothed trends. The right panel reports the mean normalized reward over the last 200 slots versus the number of beams. The window sweep indicates that $W = 120$ gives a favorable reactivity–noise trade-off, while the SNR and beam-count sweeps show higher late-time reward as communication conditions and angular resolution improve.

V. Conclusion

This paper investigated a lightweight bandit-based approach for online beam selection in time-varying ISAC systems. A normalized communication–sensing reward was combined with a two-timescale sliding-window UCB algorithm, where the beam is updated at every slot and the trade-off parameter is adapted on a slower timescale. Numerical results under abrupt changes, gradual drift, and multiple change points showed that the proposed adaptive-α_t strategy improves dynamic-regret performance and steady-state reward relative to fixed-α and stationary baselines. Future work includes extending the framework to richer sensing models and to joint adaptation of beamforming, waveform, and power.

References

[1] F. Liu, Y. Cui, C. Masouros, J. Xu, T. X. Han, Y. C. Eldar, and S. Buzzi, "Integrated sensing and communications: Toward dual-functional wireless networks for 6G and beyond," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 6, pp. 1728–1767, 2022.

This article has been accepted for publication in IEEE Wireless Communications Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LWC.2026.3696048

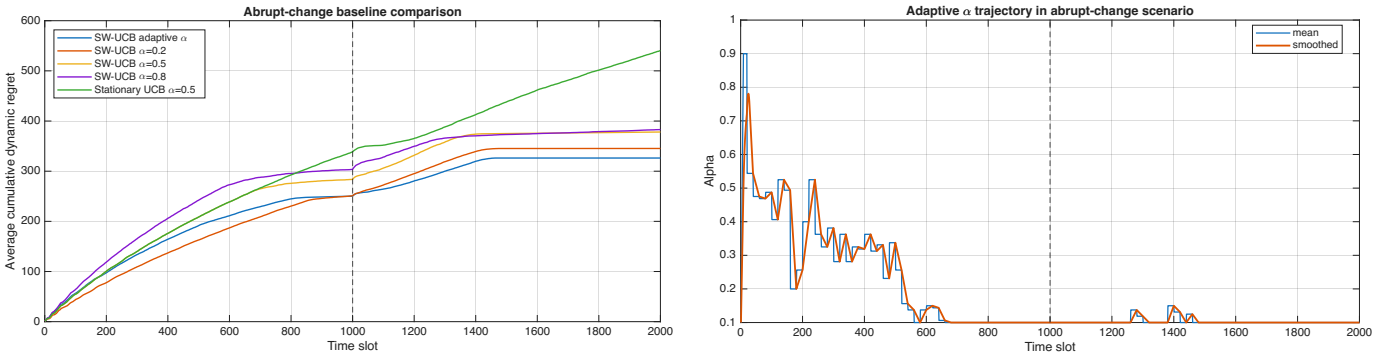

Fig. 2. Abrupt-change learning behavior. Left: average cumulative dynamic regret for the proposed SW-UCB with adaptive α_t, fixed-α SW-UCB baselines, and stationary UCB. Right: average and smoothed trajectories of the adaptive trade-off parameter.

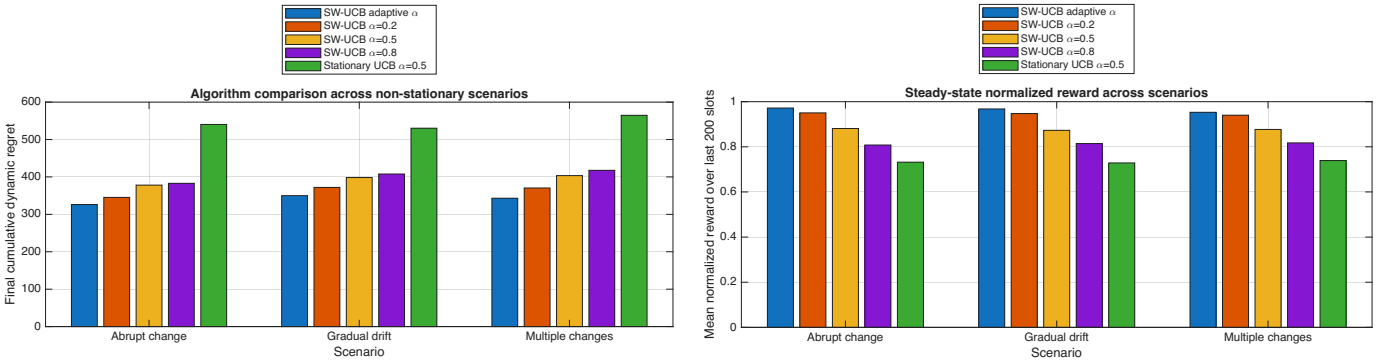

Fig. 3. Cross-scenario summary for $W = 120$. Left: cumulative dynamic regret at $t = T$ corresponding to abrupt-change, gradual-drift, and multiple-change scenarios. Right: mean normalized reward over the last 200 slots.

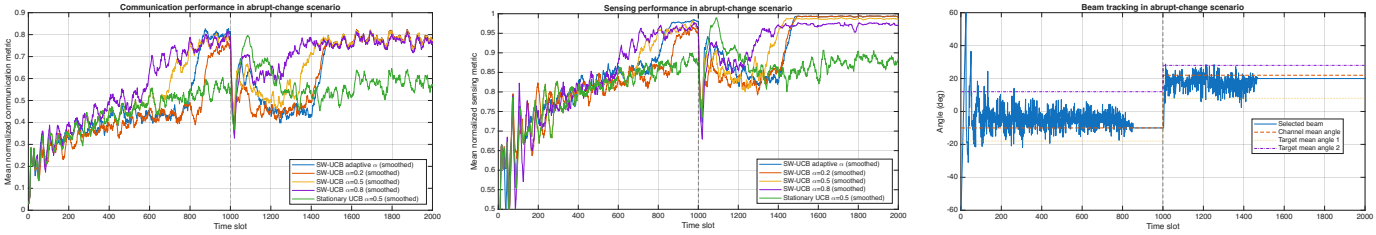

Fig. 4. Sensitivity studies for the proposed adaptive-α_t SW-UCB method in the abrupt-change scenario, with baseline values $W = 120$, SNR $= 10$ dB, and $K = 25$ unless varied. Left: final cumulative dynamic regret versus W, with raw values and a smoothed trend. Middle: mean normalized reward over the last 200 slots versus SNR, with raw and smoothed curves. Right: mean normalized reward over the last 200 slots versus the number of beams.

[2] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[3] A. Slivkins, "Introduction to multi-armed bandits," *Foundations and Trends in Machine Learning*, vol. 12, no. 1–2, pp. 1–286, 2019.

[4] A. Nasser, A. Celik, and A. M. Eltawil, "A multi-armed bandit approach for user-target pairing in NOMA-aided ISAC," in *2024 IEEE 35th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2024, pp. 1–6.

[5] M. Farzanullah, H. Zhang, A. Bin Sediq, A. Afana, and M. Erol-Kantarci, "Beam selection in ISAC using contextual bandit with multi-modal transformer and transfer learning," 2025. [Online]. Available: https://arxiv.org/abs/2503.08937

[6] N. Blinn and M. R. Bloch, "Multi-armed bandit dynamic beam zooming for mmWave alignment and tracking," *IEEE Transactions on Wireless Communications*, vol. 24, no. 9, pp. 7908–7922, 2025.

[7] A. Sneh, S. S. Ram, S. J. Darak, and A. Tewari, "Beam alignment in multipath environments for integrated sensing and communication using bandit learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 18, no. 5, pp. 871–885, 2024.

[8] P. Pulkkinen and V. Koivunen, "Model-based online learning for active ISAC waveform optimization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 18, no. 5, pp. 737–751, 2024.

[9] S. Sarkar, M. Krunz, S. Badran, J. M. Jornet, D. Manzi, and R. Kulkarni, "GAMBIT: A generalized adaptive multi-armed bandit for intelligent beam tracking and rate adaptation," *IEEE Transactions on Vehicular Technology*, vol. 74, no. 11, pp. 17 683–17 696, 2025.

[10] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2–3, pp. 235–256, 2002.

[11] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3–4, pp. 285–294, 1933.

[12] S. Agrawal and N. Goyal, "Analysis of Thompson sampling for the multi-armed bandit problem," in *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, vol. 23, 2012, pp. 39.1–39.26.

[13] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proceedings of the 19th International Conference on World Wide Web (WWW)*, 2010, pp. 661–670.

[14] A. Garivier and E. Moulines, "On upper-confidence bound policies for switching bandit problems," in *Proceedings of the 22nd International Conference on Algorithmic Learning Theory (ALT)*, 2011, pp. 174–188.