Pulkkinen, Petteri; Koivunen, Visa

Model-Based Online Learning for Resource Sharing in Joint Radar-Communication Systems

Published in:
2022 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2022 - Proceedings

DOI:
10.1109/ICASSP43922.2022.9747269

Published: 01/01/2022

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Pulkkinen, P., & Koivunen, V. (2022). Model-Based Online Learning for Resource Sharing in Joint Radar-Communication Systems. In *2022 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2022 - Proceedings* (pp. 4103-4107). (IEEE International Conference on Acoustics, Speech and Signal Processing ; Vol. 2022-May). IEEE. https://doi.org/10.1109/ICASSP43922.2022.9747269

MODEL-BASED ONLINE LEARNING FOR RESOURCE SHARING IN JOINT RADAR-COMMUNICATION SYSTEMS

Petteri Pulkkinen[⋆†], *Visa Koivunen*[⋆]

[⋆]Department of Signal Processing and Acoustics, Aalto University, Espoo, Finland
[†]Saab Finland Oy, Helsinki, Finland

ABSTRACT

The ever-increasing congestion in the radio spectrum has made coexistence and co-design for radar and communication systems an important problem to address. The radio spectrum is a rapidly time-frequency-space varying resource, and learning is required to use the spectrum and mitigate the interference. This paper proposes a model-based online learning (MBOL) framework to enable a structured way to formulate efficient online learning algorithms for resource sharing in joint radar-communication (JRC) systems. As an example, we apply the MBOL framework for allocating frequency resources in non-cooperative shared spectrum scenarios. The proposed MBOL algorithm learns a predictive model using online convex optimization (OCO) and chooses the best frequency channels in uncertain interference environments. The algorithm outperforms the considered baseline algorithms in terms of regret that quantifies the cost of learning.

***Index Terms*—** coexistence, joint radar-communication systems, model-based learning, online convex optimization

1. INTRODUCTION

The increased number of systems operating in radio frequencies is leading to a congested or even contested frequency spectrum. Consequently, future radars should be able to operate in coexistence with communication systems or even be co-designed joint radar-communication (JRC) systems [1, 2]. Radar sensing and communication systems share the same spatial and frequency resources. Hence, it is necessary to develop methods for managing and mitigating the interference. There are two types of coexistence; cooperative and non-cooperative coexistence. In the cooperative scenario, awareness about the radio environment state is exchanged among the agents and sub-systems. In non-cooperative scenarios, no or little information is shared; hence the state of the spectrum has to be estimated in each node.

The spectrum state is a rapidly time-frequency-space varying due to the coexisting systems and mobility. Therefore, online learning methods using local observations and the information from cooperative systems are required to utilize the spectrum resources efficiently. We propose a model-based online learning (MBOL) framework for sharing resources among radar and communication systems under the cooperative and the non-cooperative coexistence. The framework combines ideas from model-based reinforcement learning (MBRL) [3] and learning-based model predictive control (LMPC) [4]. To our best knowledge, MBRL or LMPC approaches have not been employed earlier in radar or JRC systems. As an intuitive example of how to use the MBOL framework, we consider agile use of frequency resources in a non-cooperative spectrum sharing context.

Online learning for channel sensing and access has been extensively researched in the context of cognitive radios [5, 6]. Recently, online learning methods for radar and communication coexistence have been proposed in [7–10]. In [7], a novel reinforcement learning (RL) approach for non-cooperative coexistence was proposed, and [8] further extends the approach using a deep RL algorithm. These RL algorithms select linear frequency modulated (LFM) waveforms with different center-frequencies and bandwidths from a library of waveforms. A sample efficient RL method was proposed in [9] where a contextual multi-armed bandit (MAB) algorithm was used. For cooperative scenarios, an RL algorithm was proposed in [10] to control how the bandwidth is shared between radar and communication sub-systems.

The RL algorithms in [7–10] proposed for spectrum sharing are based on model-free reinforcement learning (MFRL) [11] and do not take advantage of rich structural information on communications and sensing systems and radio spectrum. Consequently, MFRL algorithms are sample inefficient, especially when the cardinality of the decision space is large or when the decision space is continuous, which is typically the case in JRC systems. Moreover, with MFRL it is complicated to satisfy constraints that are typically employed in JRC systems. The constraints are imposed to ensure hardware limits as well as to guarantee desired performance levels between different sub-systems [12].

The MBOL approach proposed in this paper enables learning with large and constrained decision spaces, makes the learned system more explainable, and enables learning policies sample efficiently by replacing real-world experience with the learned model. We propose a model-based online

convex optimization (MOCO) algorithm that uses the MBOL framework and online convex optimization (OCO) [13]. The performance of the MOCO algorithm is studied in simulations using Markovian and adversarial interference environments. The results show superior performance compared to the considered baseline algorithms in terms of regret.

2. MODEL BASED ONLINE LEARNING

Model-based methods utilize certain levels of modeling information to solve a problem. Learning algorithms are employed when the model structure or parameters are unknown in advance. Models allow for faster learning and improved explainability. MBRL can be applied to solve a problem that is modeled as a Markov decision process (MDP) or a partially observable Markov decision process (POMDP) [3]. A related method in the field of optimal control is the LMPC [4]. The main difference between MBRL and LMPC is their application domain, and typically MBRL considers models as a black-box. In model predictive control (MPC), the models are considered to be *sufficiently descriptive* [14]. Therefore, an LMPC algorithm may use less generic parametric models to save computation and learning time.

We use MBOL to refer to MBRL and LMPC algorithms that learn model parameters online. The MBOL problem is generally modeled as an POMDP comprising the *agent*, *environment*, and *sensor* as visualized in Fig. 1. The agent acts in the environment with state $\mathbf{s}\left[k\right]$ using actions $\mathbf{a}\left[k\right]$ where k is the time index. The transitions of the states are described by a transition function $\mathbf{s}\left[k+1\right] \sim f(\mathbf{s}\left[k\right], \mathbf{a}\left[k\right])$ which is a probability distribution. The agent can not observe the state directly, instead it uses a sensor which gives an observation $\mathbf{z}\left[k\right] \sim h(\mathbf{s}\left[k\right], \mathbf{a}\left[k-1\right])$.

The controller optimizes policy π such that loss function $J_H(\pi; I_k)$ for planning horizon H is minimized where I_k is the history of past actions and observations. In addition, constraints for actions $\mathbf{a}\left[k\right] \in \mathcal{K}$ are imposed. Transition and sensor models are required to minimize $J_H(\pi; I_k)$. However, even if the transition model is unknown, it is reasonable to assume that the sensor model is known. In order to solve the problem, the transitions are modeled as a parameterized function $f_{\boldsymbol{\beta}}(\mathbf{s}\left[k\right], \mathbf{a}\left[k\right])$ where $\boldsymbol{\beta}$ are the parameters to be learned.

Probing actions is required in MBOL to balance the trade-off between acquiring more information about $\boldsymbol{\beta}$ while maximizing the controller performance. This is called dual-control in MPC literature [15] and exploration–exploitation trade-off in RL literature [11]. The optimal exploration–exploitation trade-off can be built into the models and cost functions. However, the optimal solution is typically non-practical. Hence, heuristics have to be used instead. For example, it is possible to include the exploration-exploitation trade-off to the parameter $\boldsymbol{\beta}$ by using optimism under uncertainty [16] or Thompson sampling techniques [17].

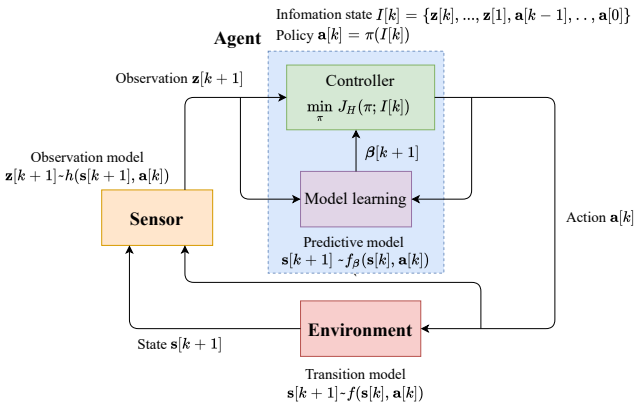


Fig. 1: Model-based online learning (MBOL) framework.

3. MBOL ALGORITHM FOR NON-COOPERATIVE SPECTRUM SHARING

Consider a wide-band JRC system operating in a congested radio spectrum with N sub-channels. The time and frequency are slotted such that the interference is assumed to evolve from slot to slot based on a certain dynamical model. One slot should be at least as short as the coherence time of the interference scenario. The interference is modeled using vector $\mathbf{x}\left[k\right] \in \{0,1\}^N$, indicating whether channels are occupied or idle. We refer to this vector as an internal state because its transitions may depend on histories of $\mathbf{x}$ and other information encoded into the environment state $\mathbf{s}\left[k\right]$. The transitions of the internal state $x_n\left[k\right] \;\; \forall \;\; n \in \{1, ..., N\}$ are modeled using parameterized Bernoulli distribution $x_n\left[k+1\right] \sim f_{\boldsymbol{\beta}}(\mathbf{s}_n\left[k\right])$ where $\mathbf{s}_n\left[k\right]$ is a channel-specific environment state. The action $\mathbf{a}\left[k\right] \in \{0,1\}^N$ decides whether or not to transmit certain sub-channels at the time slot $k+1$. The state transitions are assumed to not depend on the recent $\mathbf{a}\left[k\right]$, which is a reasonable assumption with short time slots.

The JRC system is assumed to have spectrum sensors that detect whether channels are idle or occupied. Thus a noisy observation of the internal state $\mathbf{z}\left[k\right] \sim h(\mathbf{x}\left[k\right])$ is obtained after the action $\mathbf{a}\left[k-1\right]$ has taken place. The observation $\mathbf{z}\left[k\right]$ is assumed independent of $\mathbf{a}\left[k-1\right]$. The element $z_n\left[k\right]$ corresponds to a detector output at a sub-channel n with probabilities of detection P_{d} and false alarm P_{fa}. The noise variance can be estimated from target and interference-free range-Doppler cells to control P_{fa} consequently. However, P_{d} may not be known in advance, but it can be estimated from the received signals. To simplify the notation, we assume that P_{d} and P_{fa} are equal and fixed for each sub-channel n.

Collisions refer to selected sub-bands $a_n\left[k\right] = 1 \;\forall\; n = \{1, \ldots, N\}$ where the spectrum is occupied during the transmission $x_n\left[k+1\right] = 1$. Missed opportunities refer to the sub-bands that were not selected $a_n\left[k\right] = 0$ and non-occupied

$x_n[k+1] = 0$ during the transmission. We define parameter α to weight the severity of collisions in comparison to missed opportunities. The cost function is written as follows

$$J_H(\pi; I_k) = \mathbb{E}\left[\sum_{i=1}^{H} \alpha \mathbf{a}[k+i-1]^{\mathrm{T}} \mathbf{x}[k+i] + (1 - \alpha)(\mathbf{1} - \mathbf{a}[k+i-1])^{\mathrm{T}}(\mathbf{1} - \mathbf{x}[k+i])\right] \quad (1)$$

where π is the policy and H is the planning horizon.

3.1. Controller

Minimizing the cost function in (1) is a non-convex problem when allowing the states to be dependent on the actions $\mathbf{a}$. Solving this dynamic optimization problem would require, for example, deep MBRL [3] techniques which are not considered here. The controller optimization is remarkably simplified when using myopic optimization, i.e., $H = 1$. Thus the cost function can be simplified to

$$J_1(\mathbf{a}; I_k) \triangleq \mathbf{a}^{\mathrm{T}} \mathbf{p}[k+1] - (1-\alpha)\mathbf{a}^{\mathrm{T}}\mathbf{1} \quad (2)$$

where $\mathbf{p}[k+1] = \mathbb{E}[\mathbf{x}[k+1]]$ denotes *a priori* probability of the channels being occupied at next time slot $k+1$. Since the equation (2) is linear, the problem can be solved efficiently. The sub-channels associated with the negative gradient of (2) are selected. Thus the action is

$$\mathbf{a}_k = \mathbf{1}_{\{\mathbf{p}[k+1] < (1-\alpha)\}} \quad (3)$$

which indicates that channels being occupied with a priori probability less than $1-\alpha$ are selected.

3.2. Model learning

Assume that the *a priori* channel occupancy probability can be modeled using generalized linear model. Thus we write

$$p_n[k+1] = \sigma\left(\boldsymbol{\beta}_n^{\mathrm{T}} \mathbf{s}_n[k]\right) \quad (4)$$

where is $\sigma(\cdot)$ is a sigmoid function and vector $\boldsymbol{\beta}_n$ contain model parameters for channel n. The vector $\mathbf{s}_n[k]$ is selected to contain L recent observations $\{z_n[k-i+1]\}_{i=1}^{L}$ and actions $\{a_n[k-i]\}_{i=1}^{L}$ as well as a constant equal to 1. Thus the model parameter vector $\boldsymbol{\beta}_n$ and the state vector $\mathbf{s}_n[k]$ have dimension $2L+1$.

Learning the model parameters $\boldsymbol{\beta}$ can be formulated as an OCO problem. In OCO, decisions are made sequentially to minimize convex loss functions where the loss functions are revealed only after making the decisions [13]. A suitable loss function is a logistic loss that is derived from the maximum likelihood criterion for Bernoulli distribution [18]. The logistic loss can be written as follows

$$l_k(\boldsymbol{\beta}_n) = -x_n[k+1] \log \sigma\left(\boldsymbol{\beta}_n^{\mathrm{T}} \mathbf{s}_n[k]\right) - (1 - x_n[k+1]) \log\left(1 - \sigma\left(\boldsymbol{\beta}_n^{\mathrm{T}} \mathbf{s}_n[k]\right)\right). \quad (5)$$

Furthermore, an unbiased gradient estimator of (5) is

$$\widehat{\nabla} l_k(\boldsymbol{\beta}_n) = \left[\sigma\left(\boldsymbol{\beta}_n^{\mathrm{T}} \mathbf{s}_n[k]\right) - \hat{x}_n[k+1]\right] \mathbf{s}_n[k] \quad (6)$$

where $\hat{x}_n[k]$ is an unbiased estimator for the channel occupancy. The estimator is obtained as follows

$$\hat{x}_n[k] = \frac{z_n[k] - P_{\mathrm{fa}}}{P_{\mathrm{d}} - P_{\mathrm{fa}}} \quad (7)$$

under the assumption that detection P_{d} and false alarm P_{fa} probabilities are known. For each channel n, model parameters $\boldsymbol{\beta}_n$ are updated using Online Gradient Descent (OGD) algorithm [13] as follows

$$\boldsymbol{\beta}_n[k+1] = \boldsymbol{\beta}_n[k] - \frac{D}{G\sqrt{k}} \widehat{\nabla} l_k(\boldsymbol{\beta}_n[k]) \quad (8)$$

where D and G are upper-bounds for $L2$-norms of the action space diameter and the gradient in (6), respectively.

4. NUMERICAL EXAMPLES

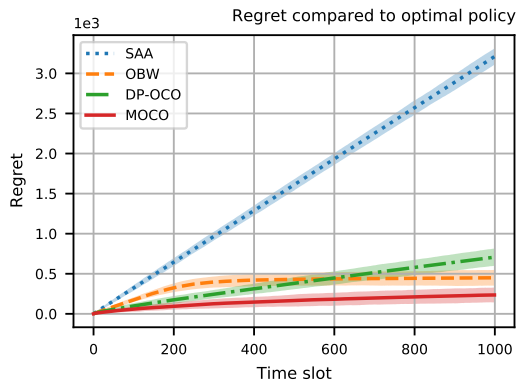
We refer to the algorithm proposed in Section 3 as model-based online convex optimization (MOCO) algorithm. Numerical examples demonstrate the performance of the MOCO algorithm in two types of interference environments. In both environments the number of sub-channels $N = 32$ and the cost trade-off parameter $\alpha = 0.75$. For the detector, probabilities of detection $P_{\mathrm{d}} = 0.9$ and false alarm $P_{\mathrm{fa}} = 0.001$ are fixed in simulations. The MOCO uses memory length $L = 5$ and the OGD parameters are set $D/G = 1$.

The performance of the MOCO algorithm is compared to the following methods: (i) Sense and Avoid (SAA) algorithm which greedily access channels which were sensed unoccupied during the recent time-slot [19], (ii) model-based approach which learns hidden Markov model (HMM) transition probabilities online using an algorithm presented in [20], we refer this algorithm as online Baum-Welch (OBW) algorithm, and (iii) directly learn channel selection probabilities using the OGD algorithm [13], this algorithm is referred to as direct policy online convex optimization (DP-OCO) algorithm. Unfortunately, we can not use the MFRL algorithms proposed in [7–9] as a baseline because they do not scale to the large action space that is typical in JRC problems.

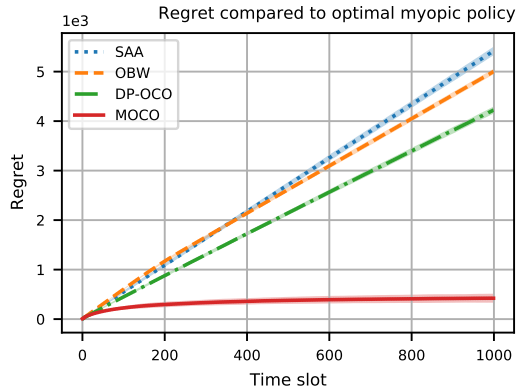
We compare the algorithms using regret

$$R_K = \mathbb{E}\left[\sum_{k=1}^{K} J_1(\mathbf{a}[k]; \mathbf{x}[k+1]) - J_1(\mathbf{a}^*[k]; \mathbf{x}[k+1])\right] \quad (9)$$

which is the expected amount of cumulative immediate costs the algorithm has suffered in comparison to the policy π^*. The policy π^* is either the optimal policy or the best known policy. Sub-linear regret is desired which indicates that the algorithm performance converges close to the policy π^* asymptotically.



(a) Markovian scenario.



(b) Adversarial scenario.

Fig. 2: MOCO obtains a sub-linear regret which is remarkably lower than the other algorithms have in both Markovian and adversarial interference environments.

4.1. Markovian interference environment

In the Markovian interference environment, the internal state vector $\mathbf{x}[k]$ transitions between idle and occupied channels based on N independent two-state Markov chains. The transition probability matrices are randomly generated for each channel by selecting the transition probabilities from a uniform distribution. A myopic controller is optimal in this scenario since the actions do not affect the state transition or observation probabilities of the Markov chains.

Fig. 2a shows regret with 95% confidence intervals for the considered algorithms obtained using 100 Monte Carlo runs. First of all, we see that the SAA algorithm performs poorly in the simulated scenario since it does not use any model to predict the channel occupancy before accessing it. The DP-OCO algorithm converges to a particular channel access policy quickly, but the regret is not sub-linear. It stems from the fact that OCO algorithm is guaranteed to have sub-linear regret only against the best single decision in hindsight [13]. Since the optimal decisions are highly dynamic, DP-OCO experiences a linear regret. On the other hand, OBW can learn the transition probabilities of the HMMs and calculates the belief states using the belief update rule. Even though OBW is not guaranteed to converge to a global minimum [20], the regret is sub-linear in the considered case. However, the regret of the MOCO algorithm is even lower and sub-linear. It is a considerably quicker learner than the other algorithms indicated by the lowest regret.

4.2. Adversarial interference environment

In the adversarial scenario, the interference source wants to deteriorate the JRC system performance intentionally. The intentional interference is modeled using a counter $c_n[k] \in \mathbb{R}^+$ for each channel n where $c_n[0] = 0$. When the JRC system accesses a channel, the counter is increased by $c_n[k+1] = c_n[k] + 1$. On the other hand, when the channel is not accessed, the counter is decreases by $c_n[k+1] = \max\{c_n[k] - 1, 0\}$. Then the channel is occupied with probability

$$\Pr\{x_n[k] = 1\} = \frac{c_n[k]}{1 + c_n[k]}. \tag{10}$$

Note that the optimal policy may be non-myopic, and therefore it is difficult to obtain. Thus we compare the algorithms to the optimal myopic policy.

Fig. 2b shows regret in the adversarial interference environment obtained with the Monte Carlo simulation. As expected, the OBW algorithm performs poorly in this scenario since it uses a model different from the true environment model. Thus the regret grows linearly and even has the largest regret of the compared algorithms. The DP-OCO algorithm performs slightly better than the SAA algorithm, but the difference is not significant. Also, both algorithms experience linear regret. On the other hand, MOCO outperforms all the other algorithms clearly, and the regret is decidedly sub-linear. However, having a sub-linear regret against an optimal non-myopic policy would likely require using a non-myopic controller.

5. CONCLUSIONS

In this paper, a model-based online learning (MBOL) approach was proposed for resource sharing in joint radar-communication (JRC) systems. In comparison to the model-free learning methods, MBOL provides a structured way to include domain knowledge to the learning problem to increase sample efficiency and algorithm explainability, which is essential in JRC applications. The MBOL applies for resource sharing in JRC and radar systems in general. However, it was demonstrated in the non-cooperative spectrum sharing context. We proposed MBOL-based algorithm that learned channel access policies sample efficiently. It outperformed the considered baseline algorithms in terms of regret in both Markovian and adversarial interference environments.

6. REFERENCES

[1] A. R. Chiriyath, B. Paul, and D. W. Bliss, "Radar-communications convergence: Coexistence, cooperation, and co-design," *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, no. 1, pp. 1–12, 2017.

[2] Fan Liu, Christos Masouros, Athina P. Petropulu, Hugh Griffiths, and Lajos Hanzo, "Joint radar and communication design: Applications, state-of-the-art, and the road ahead," *IEEE Transactions on Communications*, vol. 68, no. 6, pp. 3834–3862, 2020.

[3] Thomas M. Moerland, Joost Broekens, and Catholijn M. Jonker, "Model-based reinforcement learning: A survey," arXiv:2006.16712 [cs.LG], 2021.

[4] Lukas Hewing, Kim P. Wabersich, Marcel Menner, and Melanie N. Zeilinger, "Learning-based model predictive control: Toward safe learning in control," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 3, no. 1, pp. 269–296, 2020.

[5] J. Oksanen and V. Koivunen, "An order optimal policy for exploiting idle spectrum in cognitive radio networks," *IEEE Transactions on Signal Processing*, vol. 63, no. 5, pp. 1214–1227, Mar. 2015.

[6] Jarmo Lunden, Visa Koivunen, and H. Vincent Poor, "Spectrum exploration and exploitation for cognitive radio: Recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 123–140, 2015.

[7] E. Selvi, R. M. Buehrer, A. Martone, and K. Sherbondy, "On the use of Markov decision processes in cognitive radar: An application to target tracking," in *2018 IEEE Radar Conference (RadarConf18)*, Apr. 2018, pp. 0537–0542.

[8] C. E. Thornton, M. A. Kozy, R. M. Buehrer, A. F. Martone, and K. D. Sherbondy, "Deep reinforcement learning control for radar detection and tracking in congested spectral environments," *IEEE Trans. Cogn. Commun. Netw.*, pp. 1–16, 2020.

[9] Charles E. Thornton, R. Michael Buehrer, and Anthony F. Martone, "Constrained online learning to mitigate distortion effects in pulse-agile cognitive radar," in *2021 IEEE Radar Conference (RadarConf21)*, 2021, pp. 1–6.

[10] O. Ma, A. R. Chiriyath, A. Herschfelt, and D. W. Bliss, "Cooperative radar and communications coexistence using reinforcement learning," in *52nd Asilomar Conference on Signals, Systems, and Computers*, 2018, pp. 947–951.

[11] Richard S. Sutton and Andrew G. Barto, *Reinforcement Learning: An Introduction*, The MIT Press, Cambridge, MA, USA, 2nd edition, 2018.

[12] Marian Bică and Visa Koivunen, "Multicarrier radar-communications waveform design for RF convergence and coexistence," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7780–7784.

[13] Elad Hazan, "Introduction to Online Convex Optimization," arXiv:1909.05207 [cs.LG], 2019.

[14] Manfred Morari and Jay H. Lee, "Model predictive control: past, present and future," *Computers & Chemical Engineering*, vol. 23, no. 4, pp. 667–682, 1999.

[15] Ali Mesbah, "Stochastic model predictive control with active uncertainty learning: A Survey on dual control," *Annual Reviews in Control*, vol. 45, pp. 107–117, 2018.

[16] Tor Lattimore and Csaba Szepesvári, *Bandit Algorithms*, Cambridge University Press, Cambridge, MA, USA, 2020.

[17] Daniel J. Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen, "A tutorial on Thompson sampling," *Found. Trends Mach. Learn.*, vol. 11, no. 1, pp. 1–96, July 2018.

[18] Christopher M. Bishop, *Pattern recognition and machine learning*, Springer-Verlag, Berlin, Heidelberg, 2006.

[19] Anthony F. Martone, Kelly D. Sherbondy, Jacob A. Kovarskiy, Benjamin H. Kirk, Ram M. Narayanan, Charles E. Thornton, R. Michael Buehrer, Jonathan W. Owen, Brandon Ravenscroft, Shannon Blunt, Austin Egbert, Adam Goad, and Charles Baylis, "Closing the loop on cognitive radar for spectrum sharing," *IEEE Aerospace and Electronic Systems Magazine*, vol. 36, no. 9, pp. 44–55, 2021.

[20] Gianluigi Mongillo and Sophie Deneve, "Online learning with Hidden Markov models," *Neural computation*, vol. 20, pp. 1706–1716, Aug. 2008.