# Model-Based Online Learning for Joint Radar-Communication Systems Operating in Dynamic Interference

Petteri Pulkkinen*, Visa Koivunen†
*†Signal Processing and Acoustics, Aalto University, Espoo, Finland
*Saab Finland Oy, Helsinki, Finland
Email: *petteri.pulkkinen@aalto.fi, †visa.koivunen@aalto.fi

***Abstract*—This paper addresses the problems of co-design and cooperation among radar and communication systems operating in a shared spectrum scenario. Online learning facilitates using the spectrum flexibly while managing and mitigating rapidly time-frequency-space varying interference. We extend the previously proposed Model-Based Online Learning (MBOL) algorithm [1] to allocate frequency and power resources among co-designed and collaborating sensing and communication systems in dynamic interference scenarios. The proposed MBOL algorithm learns a predictive spectrum model using online convex optimization (OCO), assigns sub-bands between sensing and communications tasks, and optimizes their power for the tasks at hand. The performance of the proposed MBOL method is evaluated in simulations using the proposed constrained regret criterion and shown to improve the sensing and communications performance compared to the baseline method in terms of lower and sub-linear constrained regret.**

***Index Terms*—model-based online learning, joint radar-communications systems, reinforcement learning, online convex optimization, model predictive control**

## I. Introduction

Congested or even contested radio spectrum motives the development of future sensing and communications systems that can cooperate or be co-designed joint radar-communication (JRC) systems for mutual benefit [2], [3]. Cooperating systems share awareness about the radio environment state and other information among the agents and subsystems. In contrast, no or little information is shared in non-cooperative scenarios. It is necessary to develop methods for managing and mitigating the interference in the time, space, and frequency domains to achieve desired performance levels and reduce mutual interference.

The coexisting systems and mobility make the shared spectrum state change rapidly in time, frequency, and spatial dimensions. Therefore, online learning methods using local observations and the information shared among cooperative systems facilitates using the spectrum resources efficiently. In the context of opportunistic spectrum access, online learning has been extensively studied for cognitive radios [4], [5]. In addition, online learning-based spectrum sharing methods for sensing and communications have been proposed in [6]–[8]. The methods in [6]–[8] are based on model-free reinforcement learning (MFRL) [9] and do not take advantage of rich structural information available on sensing and communication systems. Consequently, MFRL algorithms are sample inefficient, especially when the decision space is large or continuous, as in JRC systems. Moreover, with MFRL, it is complicated to satisfy constraints typically employed in JRC systems. For example, constraints are imposed on transmit power and to guarantee desired performance levels in both sensing and communications subsystems [10].

The Model-Based Online Learning (MBOL) approach was proposed in our previous work [1] for resource allocation in JRC systems to address the limitations of MFRL algorithms. Model-based methods utilize certain levels of structural modeling information, and unknown model parameters are obtained by learning. Models make the learning more sample efficient and explainable. We use MBOL to refer to model-based reinforcement learning (MBRL) [11] and learning-based model predictive control (LMPC) [12] algorithms that learn online in real operational environments. The difference between MBRL and LMPC is that the models in MBRL are typically generic to address various problems. However, model predictive control (MPC) considers the models sufficiently descriptive [12]. Therefore, an LMPC algorithm may use more specific parametric models to save computation and learning time. In addition, MBRL algorithms may learn a global decision policy instead of optimizing decisions online, as in LMPC.

The contribution of this paper is to extend the MBOL algorithm proposed in [1] to co-designed JRC systems and cooperative settings. In [1], the non-cooperative and adversarial settings were considered. The proposed MBOL algorithm allocates frequency and power resources in an agile manner for co-designed dual-function radar communication (DFRC) system. It utilizes the problem formulation and the optimization approach developed in [10]. However, in contrast to [10], the proposed MBOL method does not require the interference statistics to be stationary. We evaluate the performance of the proposed MBOL algorithm in simulations using the proposed *constrained regret* criterion. The results show the effectiveness of the MBOL method compared to [10] in dynamic interference scenarios in terms of lower and sub-linear constrained regret.

## II. Problem description

Consider a colocated and co-designed DFRC system operating in a dynamic interference scenario. The scenario

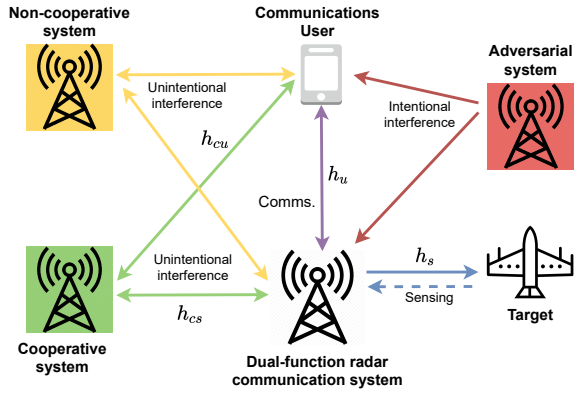



Fig. 1: Co-designed sensing and communications system operating in diverse interference scenarios.

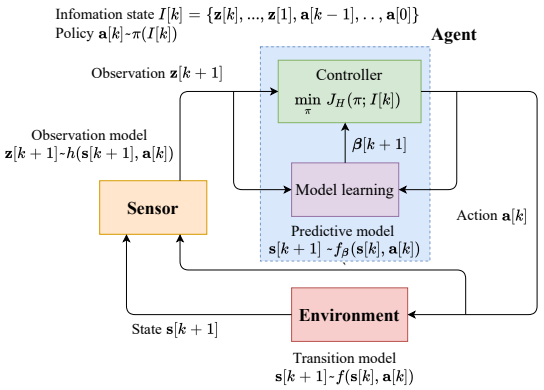



Fig. 2: Model-Based Online Learning (MBOL) framework.

is composed of the DFRC system, a communications user transmitting and receiving data from the DFRC system, and cooperative, non-cooperative, and adversarial communications or sensing systems shown in Fig. 1. The cooperative system represents a system that can share information about the channel state and interference statistics with the DFRC system. In contrast, no information is shared among the non-cooperative system. Lastly, the adversarial system represents intentional interference sources.

In this paper, the goal is to allocate the subcarriers and power resources for the DFRC system such that the communications and sensing subsystems avoid mutual interference and causing interference to the cooperative systems. However, we do not explicitly try to avoid causing interference to the non-cooperative systems as done in [1]. The time is divided into slots similarly as in [6], [7], and the interference is assumed to be quasi-stationary over a single time slot.

## III. Signal Model

For the sensing waveform $x_s(t)$, communications signal $x_u(t)$ and cooperative signal $x_c(t)$, the signal at a receiver (RX) $i$ is written as follows

$$r_i(t) = \sum_{j=1}^{L_i} h_{i,j}(t)\left[x_s(t-\tau_{i,j}) + x_u(t-\tau_{i,j})\right] + \sum_{j=1}^{L_{ci}} h_{ci,j}(t) x_c(t-\tau_{ci,j}) + v_i(t) \quad (1)$$

where the variables $L_i$, $h_{i,j}(t)$, and $\tau_{i,j}$ denote the number of resolvable paths, channel coefficient, and the propagation delay for the path $j$, respectively. In addition, $v_i(t)$ represents white Gaussian noise plus the interference from the non-cooperative and adversarial systems. From now on, $v_i(t)$ is referred to as the *disturbance*. The number of resolvable paths, channel, and propagation delay for the cooperative signal are denoted as $L_{ci}$, $h_{ci,j}(t)$, and $\tau_{ci,j}$, respectively. To indicate the signal at the DFRC RX, we use notation $i = s$ and at the communications user RX $i = u$, as illustrated in Fig. 1.

The RX signal model in (1) can be written in discrete time by first compensating for the line-of-sight path delay, and the Doppler shift is removed in the case of the sensing system. The signal at the RX $i$ is

$$\mathbf{r}_i = \mathbf{H}_i(\mathbf{x}_s + \mathbf{x}_u) + \mathbf{H}_{ci}\mathbf{x}_c + \mathbf{v}_i. \quad (2)$$

The matrices $\mathbf{H}_i$ and $\mathbf{H}_{ci}$ are Toeplitz matrices with dimension $N \times N$. They can be assumed circulant matrices if a cyclic prefix is employed, or if the Toeplitz matrix is large. Thus the channel matrices in the frequency domain are written as diagonal matrices $\mathbf{D}_i$ and $\mathbf{D}_{ci}$, where the diagonal elements are discrete Fourier transforms of the first columns of $\mathbf{H}_i$ and $\mathbf{H}_{ci}$, respectively.

The covariance matrix of the disturbance $\mathbf{v}_i$ is approximated as a circulant matrix. It means that $\mathbf{v}_i$ is modeled as colored noise. Although the shared information from the cooperative system is utilized, the launched signal $\mathbf{x}_c$ is considered to be an interfering signal. Therefore, the interference plus noise covariance matrix in the frequency domain is written as follows

$$\boldsymbol{\Omega}_i = \mathbf{D}_{ci}\mathbf{R}_c\mathbf{D}_{ci}^{\mathrm{H}} + \boldsymbol{\Psi}_i, \quad (3)$$

where $\mathbf{R}_c$ and $\boldsymbol{\Psi}_i$ are the frequency domain covariance matrices of the cooperative signal and disturbances, respectively. We assume that $\mathbf{R}_c$ is diagonal such that $\boldsymbol{\Omega}_i$ for $i \in \{s, u\}$ is also a diagonal matrix. However, note that even if the cooperative signal is orthogonal when launched from the transmitter (TX), it may not be orthogonal at the RX.

## IV. Model-Based Online Learning Formulation

The MBOL problem [1] is generally written as a partially observable Markov decision process (POMDP) [13]. This model comprises three main components: the *agent*, the *environment*, and the *sensor*, as visualized in Fig. 2. The agent desires to control the environment state $\mathbf{s}[k]$ using actions $\mathbf{a}[k]$ where $k$ is the time index. The state transitions are described by a transition function $\mathbf{s}[k+1] \sim f_{\boldsymbol{\beta}}(\mathbf{s}[k], \mathbf{a}[k])$ which is a probability distribution and $\boldsymbol{\beta}$ are the model parameters to learn. The agent observes the state using a sensor, giving the observation $\mathbf{z}[k+1] \sim h(\mathbf{s}[k+1], \mathbf{a}[k])$ where the observation model $h$ is assumed known.

We consider an MPC problem where the policy $\pi(I_k)$ means optimizing set of actions $\{\mathbf{a}[k], \ldots, \mathbf{a}[k+H-1]\}$ with respect to objective

$$\arg\max_{\pi} \ \mathbb{E}_{\pi}\left[\sum_{i=k}^{k+H-1} r(\mathbf{s}[i], \mathbf{s}[i+1], \mathbf{a}[i]) | I_k\right] \tag{4a}$$

$$\text{subject to } \mathbf{s}[i+1] \sim f_{\boldsymbol{\beta}}(\mathbf{s}[i], \mathbf{a}[i]), \tag{4b}$$

$$\mathbb{E}[c_j(\mathbf{a}[i], \mathbf{s}[i])] \in \mathcal{C}_j, \ \forall\, j = 1, \ldots, N_{\mathcal{C}}, \tag{4c}$$

where $r(\mathbf{s}[k], \mathbf{s}[k+1], \mathbf{a}[k])$ is an immediate reward, $H$ is the planning horizon and $I_k$ is the history of actions and observations. The equation (4c) represents all the imposed equality and inequality constraints. The objective (4) is optimized at each time slot $k$, and only the first action $\mathbf{a}[k]$ is taken in the real environment. Finding an optimal solution to (4) may not be feasible even if the parameters $\boldsymbol{\beta}$ are known. Thus, in general, approximate or heuristic algorithms are employed to solve this problem, such as reinforcement learning (RL) [9], sampling-based optimization methods [14], or some relaxation.

### A. States and observations

The state transitions of a POMDP should be Markovian. However, finding a Markovian state presentation may not be straightforward. Therefore, we introduce *internal states* to represent states that may have memory. For the spectrum sharing problem, the internal state $\mathbf{x}[k]$ comprises the channel and disturbance matrices $(\boldsymbol{\Psi}_s, \boldsymbol{\Psi}_u, \mathbf{R}_c, \mathbf{D}_s, \mathbf{D}_u, \mathbf{D}_{cs}, \mathbf{D}_{cu})$ effective at the time slot $k$. The Markovian state $\mathbf{s}[k]$ is approximated taking $L$ most recent internal states $\mathbf{x}$ and actions $\mathbf{a}$. If the internal states are not fully observable, the observations $\mathbf{z}$ are used instead. The observation is a noisy estimate of the internal state where $\boldsymbol{\Psi}_i$ is estimated from secondary data, $\mathbf{R}_c$ is shared via cooperative information sharing, and $\mathbf{D}_i$ and $\mathbf{D}_{ci}$ are estimated from known pilot sequences. However, this paper assumes that the internal state is fully observable, meaning that the observations are noiseless.

### B. Actions

The agent decides the DFRC operational parameters for the upcoming time slot $k+1$ by taking the action $\mathbf{a}[k]$. Here the action is a vector $\mathbf{a} = [p_1, w_1, p_2, w_2, \ldots, p_N, w_N]$ that contain the allocated sub-band powers $p_n$ and selection variables $w_n \in \{s, u\}$ that describe whether channel $n$ is used for sensing purposes ($w_n = s$) or communications ($w_n = u$). The sum of the powers in vector $\mathbf{p}$ is constrained to be less or equal to the maximum total power $P_{\text{tot}}$. We denote the action space satisfying the constraints mentioned above as $\mathcal{A}$.

### C. Rewards and constraints

The communications and sensing system performance is evaluated in terms of mutual information (MI) under the assumption of Gaussian disturbances. For communications, MI corresponds to the communications rate. For sensing, MI is obtained between the received signals and the target scattering coefficient as in [10]. Both MIs depend on the received signal's signal-to-interference-plus-noise ratio (SINR). In the quasi-stationary setting, the MI is defined as follows

$$M_i[k+1] = \sum_{n=1}^{N} \mathbb{I}_{\{w_n[k]=i\}} \log\left(1 + q_{i,n}[k+1] p_n[k]\right) \tag{5}$$

where $q_{i,n}[k+1]$ is the channel quality (channel gain divided by the interference plus noise power) at the slot $k+1$, RX $i$ and sub-carrier $n$. Note that, $w_n[k]$ and $p_n[k]$ for all $n = \{1, ..., N\}$ is the action described in Section IV-B.

Based on (5), the sensing and communications functions are co-designed by imposing a minimum rate constraint

$$\mathbb{E}[M_u[k+1] | I_k] \geq C[k] \tag{6}$$

and using the mutual information $M_s[k+1]$ as the reward function $r$ in (4a). In addition, the performance of the cooperative system is maintained at a desirable level by imposing a constraint for the maximum allowed interference power

$$\mathbf{D}_{cs}[k+1]^{\mathrm{H}} \mathbf{D}_{cs}[k+1] \mathbf{p}[k] \leq \mathbf{c}[k] \tag{7}$$

where $\mathbf{c}[k]$ is the upper-bound defined by the cooperative system.

## V. Proposed MBOL method

The model $f_{\boldsymbol{\beta}}$ in (4b) describes the state transitions for the channel matrices $\mathbf{D}_i$ and $\mathbf{D}_{ci}$ as well as for the interference covariance matrix $\boldsymbol{\Omega}_i$ for all $i \in \{s, u\}$. However, this paper concentrates on learning in the face of dynamic interference. Thus we simplify the model such that $\mathbf{D}_u$, $\mathbf{D}_{cu}$, and $\mathbf{D}_{cs}$ remain unchanged during the learning period due to the non-moving DFRC system, cooperative system, and communications user. The diagonal of the matrix $\mathbf{D}_s$ is modeled as a multivariate complex zero-mean Gaussian random variable according to the Swerling II model. The variances of the diagonal elements are assumed to be known.

The interference can be divided into the interference from the cooperative system and the disturbance, as shown in (3). We assume that the matrix $\mathbf{D}_{ci}\mathbf{R}_c\mathbf{D}_{ci}^{\mathrm{H}}$ is known via cooperative information sharing. However, the disturbance statistics $\boldsymbol{\Psi}_s$ and $\boldsymbol{\Psi}_u$ for the upcoming time slot $k+1$ are predicted from the information in the state $\mathbf{s}[k]$. This prediction is carried out using linear multiclass logistic regression (MLR) algorithm [15]. The classes are different disturbance levels $\sigma^2_{n,l}$ where $n = 1, ..., N$ is the sub-band index and $l = 1, \ldots, N_c$ is the class index. This model is selected because of its simplicity and guaranteed global converge properties. It also generalizes the approach in [1], where we used a binary classification algorithm to predict whether sub-channels are occupied or idle.

The predicted disturbance is quantized linearly on a logarithmic scale to represent the classes. The level $\sigma^2_{n,l}$ is an average of the lower and the upper edges of the quantization bins in the decibel scale. We normalize the power such that 0 dB indicates noise-only disturbance with variance $\sigma^2$. The probability distribution over the levels is expressed using the softmax function. Thus, the distribution is written as follows

$$p(\sigma^2_{n,l}[k+1] | \mathbf{s}_n[k]) = \frac{\exp\left(\boldsymbol{\beta}^T_{n,l} \mathbf{s}_n[k]\right)}{\sum_{j=1}^{N_c} \exp\left(\boldsymbol{\beta}^T_{n,j} \mathbf{s}_n[k]\right)}, \tag{8}$$

where vector $\boldsymbol{\beta}_{n,l}$ for all channels $n$ and levels $l$ are the parameters to be learned, and $\mathbf{s}_n[k]$ is a channel-specific state composed of $L$ previous actions $a_n$ and matrix elements $[\boldsymbol{\Psi}_s]_{n,n}$ and$[\boldsymbol{\Psi}_u]_{n,n}$. Thus, $\mathbf{s}_n$ is decoupled among channels to reduce the number of parameters to learn. Following the MLR methodology, the loss function used for learning the parameters $\boldsymbol{\beta}_{n,l}$ is the logistic loss [15] where the target classes $\sigma^2_{n,l}[k+1]$ are obtained by quantizing the observations. The updates are carried out online using the AdaGrad algorithm [16], which is a specific first-order online convex optimization (OCO) algorithm.

### A. Controller

The objective in (4) is relaxed to a myopic MPC where the horizon $H$ is 1. We also assume that the interference is independent of the most recent action but can depend on the history of actions. This assumption is reasonable when the length of a time slot is short in comparison to the adaptation interval of the interference sources. Thus the controller optimizes

$$\underset{\mathbf{p}[k],\mathbf{w}[k]\in\mathcal{A}}{\arg\max} \quad \mathbb{E}\left[M_s[k+1]|I_k\right] \tag{9a}$$

$$\text{subject to } \mathbb{E}\left[M_u[k+1]|I_k\right] \geq C[k], \tag{9b}$$

$$\mathbf{D}_{cs}[k+1]\mathbf{D}^{\mathrm{H}}_{cs}[k+1]\mathbf{p}[k] \leq \mathbf{c}[k]. \tag{9c}$$

This problem is a non-convex mixed-integer problem, but a heuristic approach proposed in [10] is used to solve it. The algorithm first computes the optimal power allocation for communications by setting $w_n = u \;\forall\; n$ and maximizes $\mathbb{E}[M_u[k+1]|I_k]$ subject to (9c). Secondly, the minimal number of sub-bands meeting the communications constraint in (9b) is found. Lastly, the remaining sub-bands are allocated for the sensing function, and the power is optimized according to (9).

Another subject for matter is the calculation of the expectations in (9a) and (9b). Since we approximate the problem using quantized interference levels, (9a) and (9b) can be approximated as follows (exclude $k$ for clarity)

$$\mathbb{E}\left[M_i\right] \approx \sum_{n=1}^{N}\sum_{l=1}^{N_c} \mathbb{I}_{\{w_n=i\}} p(\sigma^2_{n,l}) \log\left(1+q^{(l)}_{i,n}p_n\right) \tag{10}$$

where $q^{(l)}_{i,n}$ is the sensing channel quality of sub-channel $n$ and class $l$, and the probabilities $p(\sigma^2_{n,l})$ are defined in (8).

## VI. Numerical Examples

We evaluate the proposed MBOL algorithm in four different interference scenarios using Monte Carlo simulations. The scenarios are enumerated and referred to as (i) Markovian, (ii) Deterministic, (iii) Poisson, and (iv) Adversarial. The simulation parameters are shown in Table I. We propose the *constrained regret* as an evaluation criterion which is defined as follows

$$R_K = \mathbb{E}\left[\sum_{k=1}^{K} r^*_k \mathbb{I}_{\{\mathbb{E}[M^*_u[k+1]]\geq C[k]\}} - r_k \mathbb{I}_{\{\mathbb{E}[M_u[k+1]]\geq C[k]\}}\right]. \tag{11}$$

TABLE I: Simulation parameters.

| Desciption | Symbol | Value |
|---|---|---|
| # of channels | $N$ | 8 |
| # of interf. levels | $N_c$ | 5 |
| Memory size | $L$ | 64 |
| Power budget | $P_{\text{tot}}$ | 100 |
| Monte Carlo iterations | - | 48 |
| Model update interval | - | 8 slots |
| AdaGrad: Learning rate | - | 0.01 |
| Noise power (per ch.) | $\sigma^2$ | 0.01 |
| Ch. gain | $\mathbb{E}\left[\mathbf{D}^{\mathrm{H}}_i\mathbf{D}_i\right]$ | Diagonal $\sim \mathcal{U}$(-45dB, 0dB) |
| Coop. interf. | $\mathbf{D}_{ci}\mathbf{R}_c\mathbf{D}^{\mathrm{H}}_{ci}$ | Diagonal $\sim \mathcal{U}$(0, 0.1) |
| Power constraint | $\mathbf{c}$ | $\sim \mathcal{U}(0,\ P_{\text{tot}})$ |
| Minimum rate | $C$ | (i)-(iii): 10, (iv): 5 |
| Disturbance levels | $\sigma^2_{n,l}$ | [0, 35] (normalized dB) |

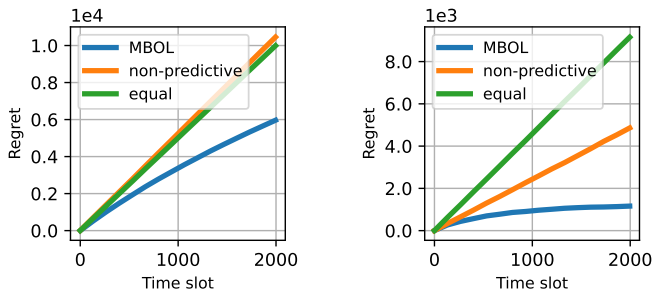
This criterion measures the amount of cumulative reward the agent obtains less compared to the policy $\pi^*$. The constraint violations are accounted by discarding the immediate rewards when the rate in (9b) is not achieved. The policy $\pi^*$ is the controller introduced in Section V-A, but using full knowledge about the environment model, thus no need for learning.

The MBOL algorithm is compared to a controller in V-A that does not use any predictive model; instead, it only uses the most recent internal state. Also, we compare to an algorithm that randomly selects sub-bands for communication or sensing with equal probability and transmits continuously with equal power allocation. Unfortunately, the model-free methods proposed in [6]–[8] can not be used as a baseline because of the prohibitive large and continuous action space, and the employed constraints would require further extensions to the methods.

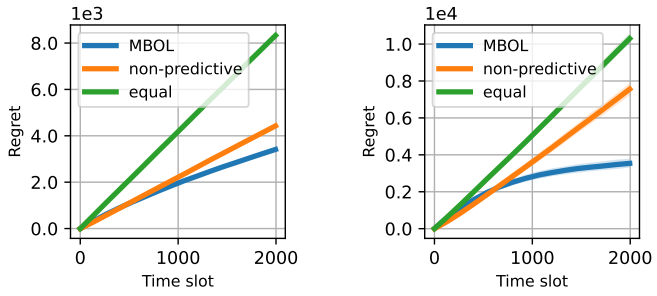
### A. Simulation results

In the considered Markovian environment, each channel contains a single non-cooperative emitter. Each emitter transitions between idle and occupied states based on a randomly generated two-state Markov chain. The interference power at RX is one of the levels $\sigma^2_{n,l}$ sampled from a distribution that favors large levels. The regret in this setting can be seen in Fig. 3a. It shows that the MBOL algorithm obtains sub-linear regret, which indicates that $\pi$ would asymptotically achieve similar performance compared to $\pi^*$. The regret is also clearly smaller compared to the baseline algorithms. Due to lack of coherency between time slots, the non-predictive method performs even worse than the equally allocating baseline.

A deterministic environment contains three frequency agile interference sources. The source dynamics are generated by selecting sub-band indices using the function $\lfloor(0.99\sin^2(2\pi f_m k)N+1)\rfloor$. The variable $f_m$ for each source $m$ is selected randomly between 0.1 and 0.7. The interference power at RX from each source is one of the quantized interference levels. In this case, the interference dynamics are entirely deterministic. Therefore, in principle, the agent could learn to predict the spectrum states perfectly. The sub-linear regret in Fig. 3b reveals that the agent can learn the dynamics to obtain smaller regret than the baseline methods.




(a) Markovian (b) Deterministic




(c) Poisson (d) Adversarial

Fig. 3: Regret in four different interference environments. Proposed MBOL algorithm performs better than the optimization method not using predictive model.

The interference scenario based on Poisson distributions is simulated as follows. Two distributions are used for each channel to sample the number of time-slots being occupied and idle. The expected time for a channel being occupied is two slots, and idle is five. The RX interference power is sampled similarly as in Markovian case (i). The regret performance is shown in Fig. 3c. It shows that the MBOL approach obtains lower regret than the baseline methods. In addition, the regret looks sub-linear but not as distinctly as in the previous cases.

The state transitions are independent of actions in the cases (i)-(iii). However, in case (iv), the interference power depends on the previous actions. The more the agent allocates power to a channel, the easier it is for the adversarial system to detect the agent and interfere with the channels in the next time slot. When the adversarial system does not detect activity, the channel is left interference-free. The clearly sub-linear regret performance of the MBOL approach can be seen in Fig. 3d. However, the performance gain of the proposed approach may decrease when the complexity of the adversarial system increases. Non-myopic planning may be vital in such complicated scenarios.

Fig. 3 shows that the proposed MBOL algorithm outperforms the baseline algorithms in terms of constrained regret in all the considered environments and scenarios. The MLR model can quickly capture the simulated dynamics to achieve a small regret. However, in cases (i) and (iii), the regret is not distinctly sub-linear, indicating that the linear MLR may not capture the dynamics perfectly. The approximation error due to the linear MLR model and the coarse quantization may lead to a performance gap that is asymptotically seen as regret growing linearly. The former problem could be reduced by using a non-linear MLR model.

## VII. Conclusions

In this paper, we proposed a Model-Based Online Learning (MBOL) algorithm for co-designed and cooperative JRC systems. The proposed algorithm learns a probabilistic interference model using OCO and employs a controller to plan over the learned model. The simulation results demonstrate superior performance for the proposed method in terms of low and sub-linear constrained regret. However, the proposed approach could be further improved by developing a non-myopic controller for action-dependent interference scenarios. In addition, it may be possible to improve the performance by using a non-linear MLR transition model.

## References

[1] P. Pulkkinen and V. Koivunen, "Model-based online learning for resource allocation in joint radar-communication systems," in *IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP22)*, pp. 4103–4107, 2022.

[2] A. R. Chiriyath, B. Paul, and D. W. Bliss, "Radar-communications convergence: Coexistence, cooperation, and co-design," *IEEE Trans. on Cogn. Commun. and Netw.*, vol. 3, no. 1, pp. 1–12, 2017.

[3] F. Liu, C. Masouros, A. P. Petropulu, H. Griffiths, and L. Hanzo, "Joint radar and communication design: Applications, state-of-the-art, and the road ahead," *IEEE Trans. on Commun.*, vol. 68, no. 6, pp. 3834–3862, 2020.

[4] J. Oksanen and V. Koivunen, "An order optimal policy for exploiting idle spectrum in cognitive radio networks," *IEEE Trans. on Signal Process.*, vol. 63, pp. 1214–1227, Mar. 2015.

[5] J. Lunden, V. Koivunen, and H. V. Poor, "Spectrum exploration and exploitation for cognitive radio: Recent advances," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 123–140, 2015.

[6] 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.

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

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

[9] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: The MIT Press, 2nd ed., 2018.

[10] M. Bică and V. Koivunen, "Multicarrier radar-communications waveform design for RF convergence and coexistence," in *IEEE Int. Conf. on Acoust., Speech and Signal Process. (ICASSP19)*, pp. 7780–7784, 2019.

[11] T. M. Moerland, J. Broekens, and C. M. Jonker, "Model-based reinforcement learning: A survey." arXiv:2006.16712 [cs.LG], 2021.

[12] L. Hewing, K. P. Wabersich, M. Menner, and M. N. Zeilinger, "Learning-based model predictive control: Toward safe learning in control," *Annu. Rev. of Control, Robot., and Auton. Syst.*, vol. 3, no. 1, pp. 269–296, 2020.

[13] V. Krishnamurthy, *Partially Observed Markov Decision Processes: From Filtering to Controlled Sensing*. Cambridge, UK: Cambridge University Press, 2016.

[14] K. Chua, R. Calandra, R. McAllister, and S. Levine, "Deep reinforcement learning in a handful of trials using probabilistic dynamics models," in *32nd Conf. on Neural Inf. Process. Syst. (NeurIPS 2018)*, 2018.

[15] C. M. Bishop, *Pattern recognition and machine learning*. Berlin, Heidelberg: Springer-Verlag, 2006.

[16] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, 2011.