A Novel Resource Allocation for Anti-jamming in Cognitive-UAVs: an Active Inference Approach

Ali Krayani, *Graduate Student Member, IEEE,* Atm S. Alam, *Member, IEEE,* Lucio Marcenaro, *Senior Member, IEEE,* Arumugam Nallanathan, *Fellow, IEEE,* and Carlo Regazzoni, *Senior Member, IEEE*

***Abstract*—This work proposes a novel resource allocation strategy for anti-jamming in Cognitive Radio using Active Inference (*AIn*), and a cognitive-UAV is employed as a case study. An Active Generalized Dynamic Bayesian Network (Active-GDBN) is proposed to represent the external environment that jointly encodes the physical signal dynamics and the dynamic interaction between UAV and jammer in the spectrum. We cast the action and planning as a Bayesian inference problem that can be solved by avoiding surprising states (minimizing abnormality) during online learning. Simulation results verify the effectiveness of the proposed *AIn* approach in minimizing abnormalities (maximizing rewards) and has a high convergence speed by comparing it with the conventional Frequency Hopping and Q-learning.**

***Index Terms*—Active Inference, Resource Allocation, Generalized Bayesian Filtering, Anti-jamming, Cognitive Radio.**

I. Introduction

With the integration of Unmanned Aerial Vehicles (UAVs), Wireless Communications (WCs) are more prone to terrestrial jammers due to the high heterogeneity and dominant Line-of-Sight (LoS) links [1]. Jammers cause damage to communication and degrade the system's performance. Therefore, it is crucial to develop an anti-jamming strategy to reach robust connectivity and improve communication security.

Cognitive Radio is a key technology to accomplish intelligent resource management in jamming scenarios. In detecting the existence of the jammers and avoiding jamming attacks, conventional anti-jamming solutions that use fixed transmission patterns can be used. However, they are unable to deal with dynamic jamming patterns in complicated radio environments with high uncertainty, and unpredictable jamming behaviours [2]. Recently, Reinforcement Learning (RL) has attracted much attention in WCs to design anti-jamming solutions in complex environments. RL methods such as Q-learning (QL) [3] are used to deal with different types of jammers. However, they suffer from slow convergence if the state and action spaces are large, which leads to anti-jamming performance degradation. Deep-QL has been proposed in [4] to overcome that issue and learn efficient defence policy. RL methods are based on a reward signal coming from the environment as a feedback to evaluate the performed action. However, defining a proper reward function in complex and dynamic environments is a big challenge [5]. **Active Inference (*AIn*)** [6] can overcome this challenging task by replacing reward functions with prior beliefs about desired sensory signals received from the environment. Thus, *AIn* agent can learn to describe how it expects itself to behave without getting a feedback from the environment. *AIn* is a promising emerging theory from cognitive neuroscience; it provides a theoretical Bayesian framework that supports how biological agents perceive and act in the real world through the free-energy principle and offers an alternative to RL.

This letter proposes an *AIn* framework as a novel resource allocation strategy for anti-jamming and studies the Cognitive-UAV based scenario. Under the *AIn* framework, the Cognitive-UAV is endowed with a joint internal representation (generative model) of the external environment, encoding the physical signal and the available physical resources jointly. This enables encoding the dynamic interaction between the UAV and the jammer in the spectrum. The objective is to learn the best set of actions performed by the UAV as interaction with a jammer that leads to the minimum surprise (positive reward). Such a representation goes over the necessity of mapping actions to signals' states directly (unlike the RL approach) and modelling them over a continuous state-space, which can be a complicated task in RL. There are four main rationals to use *AIn* approach over RL ([3], [4]): *i*) *AIn* operates in a pure belief-based setting allowing one to seek information about the environment and resolve uncertainty in a Bayesian-optimal fashion. *ii*) *AIn* enables speeding up the learning process by performing multiple updates simultaneously while adapting to the dynamic changes in the spectrum. *iii*) There is a dynamic balance between the exploration and exploitation due to the pure belief-based mode, while RL is driven by a value function that updates a single state action at each step. *iv*) In *AIn* the reliance on an explicit reward signal coming from the environment is not necessary; the reward is substituted by Generalized Errors that can be treated as self-information to avoid surprising states (i.e., states under attack) and reach the equilibrium. *To our best knowledge, this is the first work that adopts AIn for anti-jamming in WCs.*

II. System Model and Problem Formulation

Consider a cellular-connected UAV communicating with its respective Ground Base Station (GBS) to receive the tele-commands during a given mission of duration T over the Command and Control (C2) link which does not exceed a data rate of 100 Kbps [7], while a malicious terrestrial jammer transmits jamming signals with the intention of disturbing the legitimate UAV communications. The jammer may adopt

Ali Krayani is with DITEN, University of Genoa, 16145 Genoa, Italy, and also with EECS, Queen Mary University of London, London E1 4NS, U.K. (e-mail: ali.krayani@edu.unige.it, a.krayani@qmul.ac.uk). Lucio Marcenaro and Carlo Regazzoni are with DITEN, University of Genoa, 16145 Genoa, Italy, and the Italian National Consortium for Telecommunications (CNIT). (e-mails: {lucio.marcenaro, carlo.regazzoni}@unige.it). Atm S. Alam and Arumugam Nallanathan are with EECS, Queen Mary University of London, London E1 4NS, U.K. (e-mails: {a.alam, a.nallanathan}@qmul.ac.uk).

constant, random or sweep jamming patterns during a certain experience. The UAV, GBS and jammer are denoted as u, g and j, respectively. The 3D coordinate of GBS and jammer are fixed at $\boldsymbol{o^g} = [x^g, y^g, z^g]$ and $\boldsymbol{p^j} = [x^j, y^j, z^j]$, respectively, while the time-varying coordinate of UAV at time instant t is defined as $\boldsymbol{q_t^u} = [x_t^u, y_t^u, z_t^u]$. The path-loss model from the ground equipment (i.e., GBS or jammer) to UAV follows the cellular to UAV path-loss model, which can be expressed according to [8] as: $\mathrm{PL_t^{e,u}}(d_t, \theta_t) = \mathrm{PL^{ter}}(d_t) + \eta(\theta_t) + \chi(\theta_t)$, where $e \in \{g, j\}$, $\mathrm{PL_t^{ter}}(d_t) = 10\alpha \log(d_t)$ is the terrestrial path-loss of the point beneath the UAV, α is the terrestrial path-loss exponent that depends on the propagation environment and $d_t = \sqrt{(x_t^u - x^e)^2 + (y_t^u - y^e)^2}$ is the 2D distance between e and u. In addition, $\eta(\theta_t) = C(\theta_t - \theta_0)\exp\left(-\frac{\theta_t - \theta_0}{D}\right) + \eta_0$ is the excess aerial path-loss and $\chi(\theta_t)$ is a zero-mean Gaussian variable with an angle-dependent standard deviation describing the shadowing effect such that $\chi(\theta_t) \sim \mathcal{N}(0, \sigma(\theta_t){=}a\theta_t + \sigma_0)$, where C is the excess path-loss scaler, D is the angle scaler, θ_0 is the angle offset, η_0 is the excess path-loss offset, a is the UAV shadowing slope, $\theta_t = \arctan\left(\frac{z_t^u - z_t^e}{d_t}\right)$ is the depression angle and σ_0 is the UAV shadowing offset. The GBS assigns one Physical Resource Block (PRB) to the UAV each t where C2 data are transmitted [9]. The set of available links is denoted as $\mathcal{RB}{=}\{f_1, \ldots, f_n, \ldots, f_N\}$, $1 \leq \mathrm{n} \leq N$, where $|\mathcal{RB}|{=}N$ is the total number of available PRBs that depends on the channel bandwidth BW. To cope with the malicious jamming, the UAV aims to learn the best allocation strategy online by selecting the proper PRBs that are not targeted by the jammer while interacting with the environment and sending updated information to GBS to adapt to the environmental dynamic changes. Denote $\mathcal{H}_0$ and $\mathcal{H}_1$ as the hypotheses of the absence (i.e., UAV and jammer selected different PRBs) and presence (i.e., UAV and jammer selected the same PRB) of the jammer, respectively. The complex signal that is received at the UAV at time instant t and over f_n is given as $r_{t,f_n} = h_{t,f_n}^{g,u} x_{t,f_n}^u + v_t$ and $r_{t,f_n} = h_{t,f_n}^{g,u} x_{t,f_n}^u + h_{t,f_n}^{j,u} x_{t,f_n}^j + v_t$ at hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, respectively, where x_{t,f_n}^u denotes the C2 signal, $h_{t,f_n}^{g,u} = 1/\mathrm{PL_t}^{g,u}$ is the channel gain from GBS to UAV, x_{t,f_n}^j stands for the jammer's signal, $h_{t,f_n}^{j,u} = 1/\mathrm{PL_t}^{j,u}$ is the channel gain from jammer to UAV and v_t is the random noise. The corresponding SINR at the UAV is given by $\gamma_t = P_t^u h_{t,f_n}^{g,u} / (\alpha P_t^j h_{t,f_n}^{j,u} + \sigma^2)$, where P_t^u is the transmitted power, P_t^j is the jammer power, whose presence is denoted by α which is equal to 0 under $\mathcal{H}_0$ and equals to 1 under $\mathcal{H}_1$.

The anti-jamming defense problem can be formulated as a partially observable Markov decision process (POMDP) since the spectrum is only partially observable to the UAV. A discrete-time POMDP that models the relationship between the UAV and its environment can be described as 7-element tuple $(\boldsymbol{S}, \boldsymbol{X}, \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{P}_\tau^u}, \boldsymbol{\mathcal{P}_\tau^j}, \boldsymbol{\Pi_\tau^{a^u}}, \boldsymbol{\tilde{Z}_{t,f_n}})$, where $\boldsymbol{S}$ and $\boldsymbol{X}$ are sets of the environmental hidden states, $\boldsymbol{\mathcal{A}}$ is a set of actions where action is PRB selection ($a_t \in \mathcal{RB}$), $\boldsymbol{\mathcal{P}_\tau^u}$ and $\boldsymbol{\mathcal{P}_\tau^j}$ are the time-varying transition models for UAV and jammer, respectively. $\boldsymbol{\Pi_\tau^{a^u}}$ is the *AIn*-table that encodes the state-action couple and $\boldsymbol{\tilde{Z}_{t,f_n}}$ are the observations received at each t over f_n. During the offline training, UAV learns a dynamic model $\mathcal{M}$ encoding the dynamic rules that generate desired sensory signals (i.e., without jamming interference). During the active inference process (i.e., online learning), UAV predicts the environmental hidden states characterized by the posterior distributions $\mathrm{P}(s_t^* {\in} \boldsymbol{S} | z_t {\in} \boldsymbol{\tilde{Z}_{t,f_n}}, \mathcal{M})$ and $\mathrm{P}(x_t^* \in \boldsymbol{X} | z_t \in \boldsymbol{\tilde{Z}_{t,f_n}}, \mathcal{M})$ based on a prior belief (encoded in $\mathcal{M}$) and infers the actions most likely to generate preferred sensory signals (i.e., clean signals without jamming interference). Then, UAV can evaluate the situation after receiving the current observation z_t and calculate the similarity between predictions and observations using a probabilistic distance $\mathcal{D}$ (i.e., abnormality indicator). If the similarity is high (i.e., $\mathcal{H}_0$), UAV can understand that the selected action has led to desired states and to the reception of desired signals. If the similarity is low (i.e., $\mathcal{H}_1$), UAV can understand that the selected action is a bad action and updates $\boldsymbol{\Pi_\tau^{a^u}}$ accordingly to avoid selecting actions that lead to surprising states (i.e., high abnormality). Therefore, while acting and sensing the spectrum, the UAV aims to minimise the cumulative abnormality:

$$\min_{a_t^u} \sum_{t=1}^{\mathrm{T}} \mathcal{D}\Big(\mathrm{P}(s_t^* | z_t, \mathcal{M}), \mathrm{P}(z_t | s_t^*, \mathcal{M})\Big). \tag{1}$$

It is to note that (1) is equivalent to maximize the SINR.

III. Proposed Anti-jamming method

A. Radio Environment Representation

We assume that the environment is described by a Generalized-state-space model, comprised of:

$$\tilde{S}_{t,fn}^u = \mathrm{F}(\tilde{S}_{t-1,f_n}^u) + \tilde{w}_{t,f_n}, \tag{2}$$

$$\tilde{X}_{t,f_n}^u = A\tilde{X}_{t-1,f_n}^u + BU_{\tilde{S}_{t,f_n}^u} + \tilde{w}_{t,f_n}, \tag{3}$$

$$\tilde{Z}_{t,f_n} = H\tilde{X}_{t,f_n}^u + H\tilde{X}_{t,f_n}^j + \tilde{v}_{t,f_n}, \tag{4}$$

In (2), $\tilde{S}_{t,f_n}^u$ are discrete random variables (or Generalized superstates GSS) describing the discrete clusters of the UAVs' C2 signals that evolve according to (2) where F(.) is a non-linear function describing the signals' dynamic transitions among the discrete variables and its evolution over time at a specific PRB (f_n) and $\tilde{w}_{t,f_n}$ is a Generalized process noise such that, $\tilde{w}_{t,f_n} {\sim} \mathcal{N}(0, \Sigma_{\tilde{w}_{t,f_n}})$. The dynamic model in (3) explains the dynamic evolution of the continuous random variables $\tilde{X}_{t,f_n}$ (or Generalized states GS) where $A {\in} \mathbb{R}^{2d,2d}$, $B {\in} \mathbb{R}^{2d,2d}$ are the dynamic model and control model matrices, respectively, and $U_{\tilde{S}_{t,f_n}^u}$ is the control vector. The observation model is given in (4) where $\tilde{Z}_{t,f_n} {\in} \mathbb{R}^{2d}$ is the generalized observations including the signals' features in terms of I and Q components and the 1^{st}-order temporal derivatives ($\dot{I}$, $\dot{Q}$) where d is the space dimensionality. We assume that each sensory signal is a linear combination of one hidden GS ($\tilde{X}_{t,f_n}^u$) affected by additive random noise in a normal situation (i.e., under $\mathcal{H}_0$) and by additional interference ($\tilde{X}_{t,f_n}^j$) caused by the jammer in an abnormal situation (i.e., under $\mathcal{H}_1$). $\tilde{X}_{t,f_n}^u$ and $\tilde{X}_{t,f_n}^j$ are the UAV's GS and the jammer's GS (that is caused by $\tilde{S}_{t,f_n}^j$), respectively. $H {\in} \mathbb{R}^{2d,2d}$ maps hidden states to observations, f_n is the n-th PRB where $f_n {\in} \mathcal{RB}$ and $\tilde{v}_{t,f_n} {\sim} \mathcal{N}(0, \Sigma_{\tilde{v}_{t,f_n}})$.

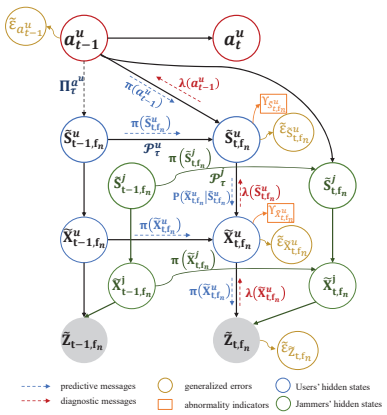


Fig. 1. Graphical representation of the proposed Active-GDBN. The top-level of the hierarchy stands for the active states (a^u_{t-1}) representing the actions that the UAV can perform. The UAV can predict the consequences of the performed actions that affect the hidden environmental states ($\tilde{S}_{t,f_n}$, $\tilde{X}_{t,f_n}$) causing sensory signals ($\tilde{Z}_{t,f_n}$). $\tilde{S}_{t,f_n}$ are discrete variables representing the clusters and $\tilde{X}_{t,f_n}$ are continuous variables representing the dynamics of the physical signal inside a certain cluster. Edges represent the conditional dependencies among random variables at multiple levels. Each level of the hierarchy holds beliefs about the variables of the level below. Beliefs are signalled via predictive messages in a top-down manner and compared against sensory signals, resulting in multi-level abnormality indicators and generalized errors that are fed back via diagnostic messages in a bottom-up manner.

B. Offline learning of desired observations

During training, we assume that the jammer is absent and the UAV aims to learn the dynamics of the desired observations (i.e., C2 signals without jamming interference) while sensing the spectrum. UAV starts perceiving the surroundings by partially sensing the spectrum, supposing that no signals are present and observations are subject to a stationary noise process that evolves according to static rules. UAV relays on (3) to predict the continuous signal's state where the force at sensing PRB (f_n) is $U_{\tilde{S}_{t,f_n}}=0$, as no rules have been discovered yet. In case of active transmissions in f_n, UAV detects abnormalities all the time and calculates the Generalized Errors (GEs) projected on the GS space as follows: $\tilde{\mathcal{E}}_{\tilde{\mathbf{X}}^{\mathbf{u}}_{\mathbf{t},\mathbf{f_n}}}=\left[\tilde{X}^u_{t,f_n}, \mathrm{P}(\dot{\mathcal{E}}_{\tilde{X}^u_{t,f_n}})\right]=\left[\tilde{X}^u_{t,f_n}, H^{-1}\tilde{\mathcal{E}}_{\tilde{Z}_{t,f_n}}\right]$, where $\dot{\mathcal{E}}_{\tilde{X}^u_{t,f_n}}$ is the difference between predictions and observations that capture the dynamics of the signals present inside the spectrum and should be applied to $\tilde{X}^u_{t,f_n}$ and $\tilde{\mathcal{E}}_{\tilde{Z}_{t,f_n}}=\tilde{Z}_{t,f_n}-H\tilde{X}^u_{t,f_n}$. GEs can be clustered in an unsupervised manner using the Growing Neural Gas (GNG) to learn the top level of abstraction (semantic level). GNG produces a set of GSS (or clusters) encoding the GEs into discrete regions described by the set $\tilde{\boldsymbol{S}}^{\boldsymbol{u}}_{\boldsymbol{f_n}}$, such that: $\tilde{\boldsymbol{S}}^{\boldsymbol{u}}_{\boldsymbol{f_n}}=\{\tilde{S}^u_{1,f_n}, \tilde{S}^u_{2,f_n}, \ldots, \tilde{S}^u_{M,f_n}\}$, where M is the total number of clusters associated with a specific PRB. Analysing the signal's dynamic transitions among the GSS and how they vary with time allows estimating the time-varying transition probabilities $\pi^u_{i f_n|j f_n,\tau}=\mathrm{P}(\tilde{S}^u_{t,f_n}=i|\tilde{S}^u_{t-1,f_n}=j,\tau)$ which is encoded in the time-varying transition matrix $\Pi^u_{f_n,\tau}$ where $i,j\in\tilde{\boldsymbol{S}}^{\boldsymbol{u}}_{\boldsymbol{f_n}}$. Moreover, each discrete variable $\tilde{S}^u_{m,f_n}\in\tilde{\boldsymbol{S}}^{\boldsymbol{u}}_{\boldsymbol{f_n}}$ is associated with statistical proprieties as generalized mean $\tilde{\mu}_{\tilde{S}^u_{m,f_n}}$ and covariance $\Sigma_{\tilde{S}^u_{m,f_n}}$. During offline learning, UAV has been trained to learn and encode the dynamic rules that generate desired sensory signals (i.e., without jamming attacks) using multiple observations (over multiple RBs).

C. Active Inference stage (online learning)

The hierarchical dynamic models formulated in terms of stochastic processes as defined in (2),(3),(4) are structured in an Active Generalized Dynamic Bayesian Networks (Active-GDBN) depicted in Fig.1. The Active-GDBN allows to solve the POMDP to find the best set of actions by predicting the situation the UAV could encounter in the future, conditioned on the actions it executes. Thus, *AIn* provides a way, through planning as inference, to form beliefs about the future and describe the causal relationship among actions, hidden states and outcomes at multiple levels.

*1) **Initialization**:* $\boldsymbol{\mathcal{P}^u_\tau}$ and $\boldsymbol{\mathcal{P}^j_\tau}$ are the $N\times N$ time-varying matrices encoding the possible transitions among the N available resources performed by the UAV and encoding the UAV's belief about the possible actions that the jammer can perform, respectively. Since there is no a priori information concerning the jammer's behaviour inside the spectrum, the probability entries in both $\boldsymbol{\mathcal{P}^u_\tau}$ and $\boldsymbol{\mathcal{P}^j_\tau}$ are initially assigned equal values:

$$\boldsymbol{\mathcal{P}^u_\tau}=\begin{bmatrix}\mathrm{P}(\Pi^u_{f_1|f_1,\tau}) & \cdots & \mathrm{P}(\Pi^u_{f_1|f_N,\tau})\\ \vdots & \ddots & \vdots\\ \mathrm{P}(\Pi^u_{f_N|f_1,\tau}) & \cdots & \mathrm{P}(\Pi^u_{f_N|f_N,\tau})\end{bmatrix},\boldsymbol{\mathcal{P}^j_\tau}=\begin{bmatrix}\mathrm{P}(\Pi^j_{f_1|f_1,\tau}) & \cdots & \mathrm{P}(\Pi^j_{f_1|f_N,\tau})\\ \vdots & \ddots & \vdots\\ \mathrm{P}(\Pi^j_{f_N|f_1,\tau}) & \cdots & \mathrm{P}(\Pi^j_{f_N|f_N,\tau})\end{bmatrix}, \tag{5}$$

where $\mathrm{P}(\Pi^u_{f_r|f_q,\tau})=\frac{1}{N}$, $\mathrm{P}(\Pi^j_{f_r|f_q,\tau})=\frac{1}{N}$ $\forall r,q\in\mathcal{RB}$. $\boldsymbol{\Pi^{a^u}_\tau}\in\mathbb{R}^{N,N}$ is a time-varying matrix encoding the probabilistic dependencies between states and actions representing the link $a^u_{t-1}\rightarrow\tilde{S}^u_{t-1,f_n}$ in the Active-GDBN that describes $\mathrm{P}(a^u_{t-1}=f_i|\tilde{S}^u_{t-1,f_k})$ and defined as:

$$\boldsymbol{\Pi^{a^u}_\tau}=\begin{bmatrix}\mathrm{P}(a_1=f_1|\tilde{S}^u_{t-1,f_1}) & \cdots & \mathrm{P}(a_N=f_N|\tilde{S}^u_{t-1,f_1})\\ \vdots & \ddots & \vdots\\ \mathrm{P}(a_1=f_1|\tilde{S}^u_{t-1,f_N}) & \cdots & \mathrm{P}(a_N=f_N|\tilde{S}^u_{t-1,f_N})\end{bmatrix}, \tag{6}$$

where $\mathrm{P}(a^u_{t-1}=f_i|\tilde{S}^u_{t-1,f_k})=\frac{1}{N}$ $\forall i,k\in\mathcal{RB}$. UAV's action depends on the state-action couple encoded in $\Pi^{a^u}_\tau$ and on its belief about the presence of the jammer in the radio spectrum encoded in $\mathcal{P}^j_\tau$.

*2) **Action selection process**:* Initially, UAV performs random sampling to select the actions during the 1^{st} iteration as every possible action has the same probability ($\frac{1}{N}$) of being chosen. The selected action a^u_{t-1} indicates what will be the next hidden state $\tilde{S}^u_{t,f_n}$ according to $\mathrm{P}(\tilde{S}^u_{t,f_n}|\tilde{S}^u_{t-1,f_n},a^u_{t-1})$. $\tilde{S}^u_{t,f_n}$ encodes the predicted cluster of the model and the activated PRB (f_n).

In the successive iterations, first, UAV predicts the future activity of the jammer implicitly according to $\mathcal{P}^u_\tau$. Then, it can adjust the action selection step by skipping the risky resources (i.e., resources expected with high probability to be targeted by the jammer in the near future). The action selection procedure depends on a certain policy adopted by the UAV according to:

$$a^{u*}_{t-1}=\mathrm{argmax}_{\tilde{S}^u_{t-1,f_k},\mathcal{P}^u_\tau(\tilde{S}^u_{t-1,f_k})}\,\pi(a^u_{t-1}), \tag{7}$$

where $\pi(a^u_{t-1})=\mathrm{P}(a^u_{t-1}|\tilde{S}^u_{t-1,f_k})$ is a specific row in $\Pi^{a^u}_\tau$ and $\mathcal{P}^u_\tau(\tilde{S}^u_{t-1,f_k})$ is a specific row selected from $(\mathcal{P}^u_\tau)$ representing the dynamic model associated with $(\tilde{S}^u_{t-1,f_k})$ where the jammer's transitions are implicitly encoded. The model has prior belief about how a certain state $(\tilde{S}^u_{t-1,f_k})$ will evolve

into another ($\tilde{S}^{u*}_{t,f_k}$) depending on the chosen action (a^{u*}_{t-1}) according to: $\mathrm{P}(\tilde{S}^{u*}_{t,f_k}|a^{u*}_{t-1}, \tilde{S}^{u}_{t-1,f_k})$, where $\tilde{S}^{u*}_{t,f_k}$ is the expected state associated with the selected action.

*3) **Perception and joint state-prediction**:* After selecting the action that indicates the chosen PRB, UAV can rely on the corresponding transition matrix ($\Pi^{u}_{f_r|f_q,\tau}$) to perform the predictions by employing the Modified Markov Jump Particle Filter (M-MJPF) [9], that uses a combination of Particle Filter (PF) and a bank of Kalman Filters (KFs). PF starts by propagating L particles equally weighted based on the proposal density encoded in $\Pi^{u}_{f_r|f_q,\tau}$, such that: $<\tilde{S}^{u,l}_{t,f_n}, W^l_t> \sim <\pi^{u}_{if_n|jf_n,\tau}, \frac{1}{L}>$. For each particle $\tilde{S}^{u,l}_{t,f_n}$, a KF is employed to predict $\tilde{X}^{u}_{t,f_n}$. The prediction at this level is driven by the higher level as pointed out in (3) (where $U_{\tilde{S}^{u}_{t,fn}}=\tilde{\mu}_{\tilde{S}^{u,l}_{t,fn}}$) which can be expressed as $\mathrm{P}(\tilde{X}^{u}_{t,f_n}|\tilde{X}^{u}_{t-1,f_n}, \tilde{S}^{u}_{t,f_n})$. The posterior probability associated with $\tilde{X}^{u}_{t,f_n}$ is given by: $\pi(\tilde{X}^{u}_{t,f_n})=\mathrm{P}(\tilde{X}^{u}_{t,f_n}, \tilde{S}^{u}_{t,f_n}|\tilde{Z}_{t-1,f_n})$.

Once a new sensory signal is received, diagnostic messages propagate in bottom-up to adjust the expectations and update belief in hidden variables. Thus, the posterior can be updated using: $\mathrm{P}(\tilde{X}^{u}_{t,f_n}, \tilde{S}^{u}_{t,f_n}|\tilde{Z}_{t,f_n})=\pi(\tilde{X}^{u}_{t,f_n})\lambda(\tilde{X}^{u}_{t,f_n})$. In addition, the likelihood message $\lambda(\tilde{S}^{u}_{t,f_n})$ can be used to update the particles' weights according to: $W^l_t=W^l_t\lambda(\tilde{S}^{u}_{t,f_n})$, where: $\lambda(\tilde{S}^{u}_{t,f_n})=\lambda(\tilde{X}^{u}_{t,f_n})\mathrm{P}(\tilde{X}^{u}_{t,f_n}|\tilde{S}^{u}_{t,f_n})=\mathrm{P}(\tilde{Z}^{u}_{t,f_n}|\tilde{X}^{u}_{t,f_n})\mathrm{P}(\tilde{X}^{u}_{t,f_n}|\tilde{S}^{u}_{t,f_n})$, and $\mathrm{P}(\tilde{X}^{u}_{t,f_n}|\tilde{S}^{u}_{t,f_n})\sim\mathcal{N}(\mu_{\tilde{S}^{u}_{m}}, \Sigma_{\tilde{S}^{u}_{m}})$ denotes a multivariate Gaussian distribution. Also, GE ($\tilde{\mathcal{E}}_{\tilde{S}^{u}_{t,fn}}$) at the superstate level conditioned on transiting from $\tilde{S}^{u}_{t-1,f_n}$ can be expressed as: $\tilde{\mathcal{E}}_{\tilde{\mathbf{S}}^{\mathbf{u}}_{\mathbf{t,fn}}}=\left[\tilde{S}^{u}_{t-1,f_k}, \dot{\mathcal{E}}_{\tilde{S}^{u}_{t,fn}}\right]$, where $\dot{\mathcal{E}}_{\tilde{S}^{u}_{t,fn}}$ is an aleatory variable whose probability density function is given by $\mathrm{P}(\dot{\mathcal{E}}_{\tilde{S}^{u}_{t,fn}})=\lambda(\tilde{S}^{u}_{t,f_n}) - \pi(\tilde{S}^{u}_{t,f_n})$ representing the new force that can be used to update $\mathcal{P}^{u}_{\tau}$ and thus improve future predictions.

*4) **Abnormality measurements**:* In order to evaluate to what extent the current signal's evolution at the discrete level matches the predicted one based on the learned and encoded dynamics in the model, we used an abnormality indicator ($\Upsilon_{\tilde{\mathbf{S}}^{\mathbf{u}}_{\boldsymbol{t,fn}}}$) based on the Symmetric Kullback-Leibler ($\mathcal{SKL}$) Divergence (D_{KL}) [9]. $\Upsilon_{\tilde{\mathbf{S}}^{\mathbf{u}}_{\boldsymbol{t,fn}}}$ calculates the similarity between the two messages that represent discrete probability distributions entering to node $\tilde{S}^{u}_{t,f_n}$, namely, $\pi(\tilde{S}^{u}_{t,f_n})$ and $\lambda(\tilde{S}^{u}_{t,f_n})$, it is associated with $\tilde{\mathcal{E}}_{\tilde{\mathbf{S}}^{\mathbf{u}}_{\mathbf{t,fn}}}$ and formulated as:

$$\Upsilon_{\tilde{\mathbf{S}}^{\mathbf{u}}_{\boldsymbol{t,fn}}} = \sum_{i\in\mathcal{S}} \mathrm{P}_r(\tilde{S}^{u}_{t,fn}=i) D_{KL}\big(\pi(\tilde{S}^{u}_{t,fn})||\lambda(\tilde{S}^{u}_{t,fn})\big) + \sum_{i\in\mathcal{S}} \mathrm{P}_r(\tilde{S}^{u}_{t,fn}=i) D_{KL}\big(\lambda(\tilde{S}^{u}_{t,fn})||\pi(\tilde{S}^{u}_{t,fn})\big), \tag{8}$$

where $\mathrm{P}_r(\tilde{S}^{u}_{t,f_n})$ is the probability of occurrence of each superstate picked from the histogram at time instant t and calculated as follows: $\mathrm{P}_r(\tilde{S}^{u}_{t,f_n})=\frac{fr(\tilde{S}^{u}_{t,fn}=i)}{N}$, where $fr(.)$ is the frequency of occurrence of a specific superstate i, N is the total number of particles propagated by PF, and $\mathcal{S}$ is the set consisting of all winning particles, such that: $\mathcal{S}=\left\{i|\mathrm{P}_r(\tilde{S}^{u}_{t,f_n})>0\right\}$, $i\in\boldsymbol{S}^{\boldsymbol{u}}_{\boldsymbol{f_n}}$.

Likewise, it is possible to understand how much the observation supports the predictions at the GS level using:

$$\Upsilon_{\tilde{\boldsymbol{X}}^{\boldsymbol{u}}_{\boldsymbol{t,fn}}} = -\ln\left(\mathcal{BC}\big(\pi(\tilde{X}^{u}_{t,f_n}), \lambda(\tilde{X}^{u}_{t,f_n})\big)\right), \tag{9}$$

where $\mathcal{BC}(.) = \int\sqrt{\pi(\tilde{X}^{u}_{t,f_n})\lambda(\tilde{X}^{u}_{t,f_n})}d\tilde{X}^{u}_{t,f_n}$ is the Bhattacharyya coefficient and $\Upsilon_{\tilde{\boldsymbol{X}}^{\boldsymbol{u}}_{\boldsymbol{t,fn}}}$ is associated with $\tilde{\mathcal{E}}_{\tilde{\mathbf{X}}^{\mathbf{u}}_{\mathbf{t,fn}}}$.

*5) **Updating of action selection process**:* After acting in the environment, UAV can save the consequence of the chosen action (i.e., the transition from $\tilde{S}^{u}_{t-1,f_k}$ to $\tilde{S}^{u*}_{t,f_k}$) in $\mathcal{P}^{u}_{\tau}$ and evaluate how much the sensory outcomes support predictions and thus evaluate if the performed action was good or bad by using the abnormality measurements defined in (8) and (9). In addition, it is possible to calculate the GE ($\tilde{\mathcal{E}}_{a^{u}_{t-1}}$) during abnormal situations to adapt UAV's strategy in selecting actions and understand how it should behave in the future to avoid the jammer. $\tilde{\mathcal{E}}_{a^{u}_{t-1}}$ is the difference between observation and expectation which can be expressed as: $\tilde{\mathcal{E}}_{a^{u}_{t-1}}=\left[a^{u*}_{t-1}, \dot{\mathcal{E}}_{a^{u}_{t-1}}\right]$, where $\dot{\mathcal{E}}_{a^{u}_{t-1}}$ depicts an aleatory variable representing the new force that should be applied to update $\boldsymbol{\pi}(a^{u}_{t-1})$ and its probability density function is given by $\mathrm{P}(\dot{\mathcal{E}}_{a^{u}_{t-1}})=\lambda(a^{u}_{t-1}) - \boldsymbol{\pi}(a^{u}_{t-1})$ that can be used as a metric alternative to the reward in RL. $\lambda(a^{u}_{t-1})$ is the diagnostic message travelling from $\tilde{S}^{u}_{t,f_n}$ towards a^{u}_{t-1} and defined as: $\lambda(a^{u}_{t-1})=\lambda(\tilde{S}^{u}_{t,f_n})\mathrm{P}(\tilde{S}^{u}_{t,f_n}|a^{u}_{t-1})$ representing a discrete probability distribution that holds information about the observed sensory signal and encoding the probabilities about how the states $\tilde{S}^{u}_{t,f_n}$ belonging to the available frequencies change based on the evidence, it is given by:

$$\lambda(a^{u}_{t-1}) = \begin{cases} \mathcal{P}_{\tau-1}(\tilde{S}^{u}_{t-1,f_n}) - \gamma^*, & \text{if } a^{u}_{t-1} = a^{u*}_{t-1}, \\ \mathcal{P}_{\tau-1}(\tilde{S}^{u}_{t-1,f_n}) + \frac{\gamma^*}{N-1}, & \text{if } a^{u}_{t-1} \neq a^{u*}_{t-1}, \end{cases} \tag{10}$$

where γ depends on the GE ($\tilde{\mathcal{E}}_{\tilde{\mathbf{S}}^{\mathbf{u}}_{\mathbf{t,fn}}}$), that is: $\gamma=\gamma^*$ *if* $\tilde{\mathcal{E}}_{\tilde{S}_{t,f_k}} \geq th$, and $\gamma=0$ *if* $\tilde{\mathcal{E}}_{\tilde{S}_{t,f_k}} < th$, where th is the threshold indicating whether the radio situation is normal or abnormal and the value of γ^* depends on the abnormality indicators defined in (8) and (9). Hence, GE ($\tilde{\mathcal{E}}_{a^{u}_{t-1}}$) is proportional to $\tilde{\mathcal{E}}_{\tilde{\mathbf{S}}^{\mathbf{u}}_{\mathbf{t,fn}}}$ due to the messages propagated from lower level towards the higher levels, such that $\tilde{\mathcal{E}}_{a^{u}_{t-1}}=f(\tilde{\mathcal{E}}_{\tilde{S}^{u}_{t,fn}})$. When the UAV get surprised by the sensory outcomes after performing a certain action, it can use the prediction error signal to update its belief about the jammer's transition model to improve future actions. The core idea is that the user occupying a piece of the spectrum should minimize the abnormality (surprise) associated with finding itself in unlikely states (states under attack). Jammer's dynamic model ($\mathcal{P}^{j}_{\tau}$) can be updated following:

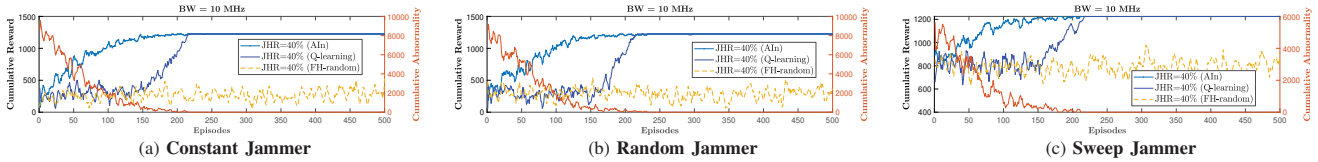
$$\mathcal{P}^{j}_{\tau}(.\,, \tilde{S}^{j}_{t,f_n}) = \mathcal{P}^{j}_{\tau-1}(.\,, \tilde{S}^{j}_{t,f_n}) - \mathrm{P}(\dot{\mathcal{E}}^{u}_{a_{t-1}}), \tag{11}$$

In an abnormal situation, the user and jammer share the same RB, which means they performed the same action. Thus, the user should update $\Pi^{a^u}_{\tau}$ by decreasing the probability of selecting that action as follows:

$$\pi^*(a^{u}_{t-1}) = \pi(a^{u}_{t-1}) + \mathrm{P}(\dot{\mathcal{E}}^{u}_{a_{t-1}}), \tag{12}$$

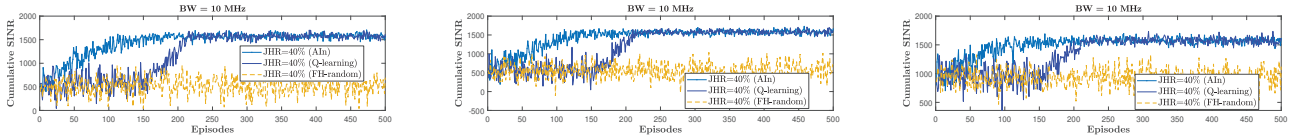
and update $\mathcal{P}^{u}_{\tau}$ by decreasing the probability of transiting to $\tilde{S}^{u}_{t,f_k}$ from $\tilde{S}^{u}_{t-1,f_k}$ after choosing action a^{u*}_{t-1} using the GE ($\tilde{\mathcal{E}}_{\tilde{\mathbf{S}}^{\mathbf{u}}_{\mathbf{t,fn}}}$) following:

$$\mathcal{P}^{u}_{\tau}(\tilde{S}^{u}_{t-1,f_k}, \tilde{S}^{u}_{t,f_n}) = \mathcal{P}^{u}_{\tau-1}(\tilde{S}^{u}_{t-1,f_k}, \tilde{S}^{u}_{t,f_n}) + \mathrm{P}(\dot{\mathcal{E}}_{\tilde{S}^{u}_{t,fn}}). \tag{13}$$



(a) **Constant Jammer** (b) **Random Jammer** (c) **Sweep Jammer**

Fig. 2. Performance comparison of cumulative reward and abnormality ($\mathcal{SKL}$) with the proposed *AIn*, FH and QL under different jamming strategies.



(a) **Constant Jammer** (b) **Random Jammer** (c) **Sweep Jammer**

Fig. 3. Performance comparison of cumulative SINR with the proposed *AIn*, FH and QL under different jamming strategies.

IV. Results and Discussion

To evaluate the performance of the proposed *AIn* approach for anti-jamming, following three types of jammers are considered in the simulation: *1)* Constant jammer that acts on statistically pre-configured channels; *2)* Sweep jammer that attacks by sweeping among the available PRBs at each time slot; and *3)* Random jammer that selects uniformly random actions to attack the available PRBs. The simulation settings are as: BW=10MHz; FDD; sub-carrier spacing of 15 KHz; number of PRBs per BW is 50; sampling frequency of 1.92 MHz; N_{FFT} of 128; 7 OFDM symbols per slot; normal CP; SNR of 15dB; QPSK for C2 and jamming signal; jamming to signal power ratio (JSR) of 6dB; and a total of 200 radio frames. In addition, the propagation environment is a typical suburban, mean aerial speed is 4.8m/s, BS height is 30m, UAV height is 60m and the channel model parameters [8] are α=3.04, σ_0=8.52, C=−23.29, η_0=20.70, θ_0=−3.61, D=4.14, a=−0.41, σ_0=5.86, where a perfect CSI is assumed. Also, we consider a jamming hit rate (JHR) of JHR=40%. C2 data, jamming signals and UAV trajectory are generated as in [9].

Let us compare the performance of *AIn* in terms of cumulative abnormality (defined in (8)) and cumulative reward with that of random Frequency Hopping (FH-random) and Q-Learning (QL), as illustrated in Fig. 2. Here, the objective of *AIn* is to minimize abnormality while that of QL is to maximize reward. Thus, the reward is considered in *AIn* approach just for the sake of comparison with QL. We consider a binary reward which is equal to −1 under $\mathcal{H}_1$ and +1 under $\mathcal{H}_0$. Nevertheless, the relationship of these metrics is opposites to one another. For a fair comparison with QL, we use time-varying q-tables to deal with the dynamic environmental changes. The exploration process in QL follows the ϵ-greedy policy with $\epsilon = 1$ decaying to 0. It can be seen from the figure that *AIn* outperforms QL and FH-random under different jamming strategies while AIn converges faster than QL due to its capability in discovering jammer's policy and performing multiple updates. Fig. 3 depicts the cumulative SINR under different jamming patterns achieved by the proposed *AIn* and compared it with FH-random and QL. By observing Fig. 2 and Fig. 3, we can notice that minimizing the abnormality (or maximizing the reward) leads to maximizing the SINR where the time needed to reach the convergence is equivalent to that in Fig. 2 and *AIn* beats both the FH-random and QL. This means that avoiding surprising states minimizes the abnormality and maximises reward and SINR. *AIn* outperforms FH and QL due to its ability to characterize the jammer and discover its attacking strategy, explaining how the UAV should act in the environment. Since *AIn* operates in a pure belief-based setting. It can evaluate whether the action was correct or wrong and also understand how to correct those actions using the errors by performing multiple updates to the *AIn*-table, which speeds up the learning process and reach convergence faster. In contrast, QL performs single updates to the q-table without being able to explain how to correct the wrong actions, hindering the learning process. While FH can not reach convergence as it is always selecting random actions.

V. Conclusion

This letter has proposed a novel resource allocation strategy using Active Inference for anti-jamming in a Cognitive-UAV scenario. Simulated results have indicated that the proposed method outperforms conventional Frequency Hopping and Q-Learning in terms of learning speed (convergence). Further research will explore performance improvements by facing smart reactive jammers in fully-observable environments.

References

[1] Q. Wu, W. Mei, and R. Zhang. Safeguarding Wireless Network with UAVs: A Physical Layer Security Perspective. *IEEE Wireless Communications*, 26(5):12–18, 2019.

[2] Q. Qiu, H. Li, H. Zhang, and J. Luo. Bandit based Dynamic Spectrum Anti-jamming Strategy in Software Defined UAV Swarm Network. In *2020 IEEE 11th International Conference on Software Engineering and Service Science (ICSESS)*, pages 184–188, 2020.

[3] S. Machuzak *et al.* Reinforcement learning based anti-jamming with wideband autonomous cognitive radios. In *2016 IEEE/CIC International Conference on Communications in China (ICCC)*, pages 1–5, 2016.

[4] G. Han, L. Xiao, and H. V. Poor. Two-dimensional anti-jamming communication based on deep reinforcement learning. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2087–2091, 2017.

[5] Z. Hu, K. Wan, X. Gao, and Y. Zhai. A Dynamic Adjusting Reward Function Method for Deep Reinforcement Learning with Adjustable Parameters. *Mathematical Problems in Engineering*, 2019, 2019.

[6] K. Friston *et al.* Cognitive Dynamics: From Attractors to Active Inference. *Proceedings of the IEEE*, 102(4):427–445, 2014.

[7] S. R. Sabuj, A. Ahmed, Y. Cho, K. Lee, and H. Jo. Cognitive UAV-Aided URLLC and mMTC Services: Analyzing Energy Efficiency and Latency. *IEEE Access*, 9:5011–5027, 2021.

[8] Akram Al-Hourani and Karina Gomez. Modeling Cellular-to-UAV Path-Loss for Suburban Environments. *IEEE Wireless Communications Letters*, 7(1):82–85, Feb 2018.

[9] A. Krayani *et al.* Self-Learning Bayesian Generative Models for Jammer Detection in Cognitive-UAV-Radios. In *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, pages 1–7, 2020.