Learning of scanning strategies for electronic support using predictive state representations

Hadrien Glaude, Cyrille Enderli, Jean-François Grandin, Olivier Pietquin

LEARNING OF SCANNING STRATEGIES FOR ELECTRONIC SUPPORT USING PREDICTIVE STATE REPRESENTATIONS

Hadrien Glaude[⋆†] *Cyrille Enderli*[†] *Jean-François Grandin*[†] *Olivier Pietquin*[⋆◇]

† Thales Airborne Systems, Elancourt, France
⋆ Univ. Lille, CRIStAL, UMR 9189, SequeL Team, Villeneuve d'Ascq, France
◇ Institut Universitaire de France (IUF)

ABSTRACT

In Electronic Support, a receiver must monitor a wide frequency spectrum in which threatening emitters operate. A common approach is to use sensors with high sensitivity but a narrow bandwidth. To maintain surveillance over the whole spectrum, the sensor has to sweep between frequency bands but requires a scanning strategy. Search strategies are usually designed prior to the mission using an approximate knowledge of illumination patterns. This often results in open-loop policies that cannot take advantage of previous observations. As pointed out in past researches, these strategies lack of robustness to the prior. We propose a new closed loop search strategy that learns a stochastic model of each radar using predictive state representations. The learning algorithm benefits from the recent advances in spectral learning and rank minimization using nuclear norm penalization.

***Index Terms*—** Electronic support, super heterodyne, sensor scheduling, predictive state representation, subspace identification

1. INTRODUCTION

In times of crisis, controlling the electromagnetic environment is a critical part of any modern military action. For example, radars are widely used to track aircrafts and military vehicles, to capture images, or to guide missiles. Electronic support (ES) is the branch of electronic warfare (EW) that refers to passive detection and analysis of electromagnetic radiation in order to shape the tactical situation and engage countermeasures like jamming or launching flares. An ES system involves a long signal processing chain whose main steps are interception of electromagnetic activities, radar pulses deinterleaving, reconstructed signals analysis, and finally emitters identification. This paper focuses on the intercept of signals. In our setting, electromagnetic activities are captured with a super heterodyne (SH) receivers. A SH receiver has a narrow bandwidth and a high sensitivity allowing a better detection of low power signals. Because of hardware costs, the number of SH receivers in an ES system is limited. Thus, only a small portion of the frequency spectrum can be monitored simultaneously. By sweeping across frequency bands, the SH receiver can cover the whole spectrum. A scanning strategy describes which frequency bands have to be visited as well as when and for how long.

A common approach to design scanning strategies is to use prior knowledge gathered by electronic intelligence about radars likely to be encountered during the mission. The scheduling algorithm is then tuned to achieve high performances measured by probabilities of interception [1, 2, 3] (PoI) or mean intercept times [4, 5, 6]. These strategies have two drawbacks. First, they strongly rely on the accuracy of the prior. Sometimes, a small shift in the prior can cause performances to collapse [7]. So that, according to [8], with an unknown prior, the best stationary strategies are stochastic. This phenomenon, known as synchronization, has been extensively studied [9, 10]. Secondly, these strategies are open-loop, meaning the scheduling algorithm does not depend on past measurements. Recently, in [11], authors studied what they called dynamic scheduling or equivalently closed loop strategies. Similarly, in [12], authors proposed a strategy able to learn through Bayesian filtering the radar sweeping period. In this paper, we model scanning strategies as a sequential decision problem where the goal is to optimize the total number of interceptions. Each radar illumination pattern is modeled by a predictive state representation (PSR) learnt from past observations. Thus, at each time step, the next frequency band is chosen to both increase the number of interceptions and improve radar models accuracy. Hence, we designed our strategy to deal with exploration and exploitation. The algorithm is finally evaluated on synthetic data.

2. BACKGROUND

2.1. Radar Signals and Search Strategies

In our setting, both the receiver and emitters are scanning. Interceptions happen when the main beam of an emitter is directed toward the receiver which, in turn, has to be listening in the right frequency band for a sufficient amount of time. The scanning pattern of an emitter depends on his type of antenna. Mechanical antennas produce periodic illuminations of the receiver, whereas more sophisticated emitters use electronic beam steering that can produce more or less random patterns. However, due to their scanning function electronic beam steering antennas produce roughly periodic illuminations. In addition, even for mechanical antennas, there are many sources of randomness in the period. So, we model illumination patterns as on/off periodic signals with a uniform jitter on the period. Here, the scanning strategy is viewed as planning into a partially observable stochastic decision process where a learning agent has to decide among actions given observations. In this framework, at each time step, an action stands for listening one frequency band and observations to interceptions or silences. In ES, a first key operational requirement is the ability to detect or intercept radars in the shortest possible time. This requirement is usually met by sweeping rapidly across frequency bands. Once intercepted for a first time, a second key operational requirement is to monitor threatening radars in order to maintain a good knowledge of the tactical situation. For the sensor, this second requirement is met if the scanning strategy is tuned to intercept every successive illuminations. Open-loop strategies focus

only on the first requirement and usually waste time in intercepting known nonthreatening radars to the detriment of new or threatening ones. In the sequel, time is discretized such that the time step is long enough to intercept any radar signals. We assume existence of a processing step that associates intercepted signals to radars.

2.2. Spectrum Opportunities in Cognitive Radio Networks

Opportunistic Spectrum Access (OSA) in cognitive radio networks is a closely related problem. In this problem, the network is compound of a certain number of channels corresponding to different frequency bands, each owned by a primary user to communicate. A set of secondary users is allowed to transmit at a particular time and location when and where the primary users are not active. In decentralized cognitive networks secondary users have to learn the emitting pattern of primary users to find opportunities in the spectrum. This problem can be formulated as a multi-user multi-armed bandit problem [23, 26], where each channel is modeled as a i.i.d. process. More complicated settings assumed that channels evolve in a Markovian way, [24]. The quality of a strategy is measured by the regret with respect to the same strategy that would have a perfect knowledge of the channels' statistics. However, these works focus mainly on fixed strategies choosing the best channel, while better results could be achieve with dynamic strategies allowed to switch between frequency bands. In [25, 27], authors derive the notion of strong regret corresponding to dynamic strategies. They propose algorithms achieving a low regret, assuming the channels evolve in Markov way. Although closely related, these works cannot be applied to our problem because in electronic support each frequency band is partially observable. For example, even with the knowledge of radars emitting patterns, if at a particular time in a frequency band no illumination is observed, one cannot predict when the next one will occur. To modeled these partially observable processes, we use linear predictive state representations, presented in the next section.

2.3. Linear Predictive State Representation

A predictive state is a sufficient statistic of past events to predict the future. Let $\mathcal{A}$ be a discrete set of actions and $\mathcal{O}$ a discrete set of observations, an history $h_t := (a_1, o_1, a_2, o_2, ..., a_t, o_t)$ is a succession of taking actions and observations received since system started up to time t. Let $\mathbb{P}\left(o_{t+1}|a_{t+1}, h_t\right)$ be the probability of observing o_{t+1} after taken action a_{t+1} and observing history h_t. We call a test $\tau_{t+1} := (a_{t+1}, o_{t+1}, a_{t+2}, o_{t+2}, ..., a_{t+p}, o_{t+p})$ of size p a succession of p action-observation pairs in the future. We write τao for a test built by concatenation of a test τ followed by an action a and an observation o. We denote by τ^A the sequence of actions and τ^O the sequence of observations of τ. Similarly, a history can be divided between observations h^O and actions h^A.

In linear Predictive State Representations (PSRs), introduced in [22], the probability of any future can be written as a linear combination of the occurrence probabilities of a small set of tests given the current history. These tests are called core tests. Thus, in the current history h_t, the predictive state, denoted by $\mathbf{m}_t$, can be defined from any set $\mathcal{Q} := \left\{\tau_1, ..., \tau_{|\mathcal{Q}|}\right\}$ of core tests, as a vector of conditional probabilities of these tests given the current history h_t :

$$\mathbf{m}_t := \left[\mathbb{P}\left(\tau_i^O \middle| h_t, \tau_i^A\right)\right]_{\tau_i \in \mathcal{Q}},$$

More generally, we say that a set of tests $\mathcal{T}$ is core if and only if we can define a predictive state such that for any test τ, there exists $\mathbf{r}_\tau^\mathcal{T}$ verifying, for any history h_t,

$$\begin{aligned}\mathbb{P}\left(\tau^O \middle| h_t, \tau^A\right) &= {\mathbf{r}_\tau^\mathcal{T}}^\top \mathbf{m}_t^\mathcal{T}, \\ \text{with } \mathbf{m}_t^\mathcal{T} &= \left[\mathbb{P}\left(\tau_i^O \middle| h_t, \tau_i^A\right)\right]_{\tau_i \in \mathcal{T}}.\end{aligned}$$

Let $M_{ao} := [\mathbf{r}_{\tau_i ao}^\top]_i = [\mathbb{P}\left(\tau_i^O | h_t, \tau_i^A\right)]_i$ be the matrix built by stacking the row vectors $\mathbf{r}_{\tau_i ao}^\top$ for each test in the core set, then we have

$$\left[\mathbb{P}\left(\tau_i^O o \middle| h_t, \tau_i^A a\right)\right]_{\tau_i \in \mathcal{Q}} = [\mathbf{r}_{\tau_i ao}^\top]_i \mathbf{m}_t = M_{ao}\mathbf{m}_t.$$

Let $\mathbf{m}_\infty^\top$ a normalization vector such that $\forall t\, \mathbf{m}_\infty^\top \mathbf{m}_t = 1$, then prediction of the next observation given the next action can be done by,

$$\mathbb{P}\left(o|a, h_t\right) = \mathbf{m}_\infty^\top \left[\mathbb{P}\left(\tau_i^O o \middle| h_t, \tau_i^A a\right)\right]_{\tau_i \in \mathcal{Q}} = \mathbf{m}_\infty^\top M_{ao}\mathbf{m}_t, \quad (1)$$

where the normalization vector is marginalizing out the probability of the state from the joint probability $\mathbb{P}\left(\tau_i^O o \middle| \tau_i^A a, h_t\right)$. We can now perform a state update through Bayesian filtering. After executing action a and observing o, we have,

$$\begin{aligned}\mathbf{m}_{t+1} &= \left[\mathbb{P}\left(\tau_i^O \middle| h_{t+1}, \tau_i^A\right)\right]_{\tau_i \in \mathcal{Q}} \\ &= \left[\mathbb{P}\left(\tau_i^O \middle| h_t, o, \tau_i^A a\right)\right]_{\tau_i \in \mathcal{Q}} \\ &= \left[\frac{\mathbb{P}\left(\tau_i^O o | h_t, \tau_i^A a\right)}{\mathbb{P}\left(o | h_t, \tau_i^A a\right)}\right]_{\tau_i \in \mathcal{Q}} \\ &= \left[\frac{\mathbb{P}\left(\tau_i^O o | h_t, \tau_i^A a\right)}{\mathbb{P}\left(o|a, h_t\right)}\right]_{\tau_i \in \mathcal{Q}} \text{ by causality} \\ &= \left[\frac{\mathbf{r}_{\tau_i ao}^\top \mathbf{m}_t}{\mathbf{m}_\infty^\top M_{ao}\mathbf{m}_t}\right]_{\tau_i \in \mathcal{Q}} = \frac{M_{ao}\mathbf{m}_t}{\mathbf{m}_\infty^\top M_{ao}\mathbf{m}_t}.\end{aligned} \quad (2)$$

We denote by $\mathbf{m}_\star$ the initial predictive state, where we assumed that the system starts in a random state drawn from its stationary distribution. The vectors $\mathbf{m}_\star$, $\mathbf{m}_\infty$ and the matrices M_{ao} define the PSR parameters. We say that a core set is *minimal* if the occurrence probability of each test cannot be written as a linear combination of probabilities of other tests in the set. In the sequel, $\mathcal{Q}$ stands for a minimal core set.

Note that predicting (1) and filtering (2) equations are invariant to any linear invertible transformation J of the parameters. Let,

$$\mathbf{b}_\infty := J\mathbf{m}_\infty, \quad B_{ao} := JM_{ao}J^{-1}, \quad \mathbf{b}_\star := J\mathbf{m}_\star,$$

we still have that

$$\mathbb{P}\left(o|a, h_t\right) = \mathbf{b}_\infty^\top B_{ao}\mathbf{b}_t, \qquad \mathbf{b}_{t+1} = \frac{B_{ao}\mathbf{b}_t}{\mathbf{b}_\infty^\top B_{ao}\mathbf{b}_t}.$$

These new parameters define a transformed version of the PSR, refereed to as the Transformed PSR (TPSR [13]). So, starting with a non minimal core set of tests $\mathcal{T}$, one can recover the PSR parameters up to a linear transformation by spectral decomposition [14]. First, let's define the following probability vector and joint probabilities matrices,

$$\begin{aligned}\mathcal{P}_\mathcal{H} &= \left[\mathbb{P}\left(h_i^O \middle| h_i^A\right)\right]_i \in \mathbb{R}^{|\mathcal{H}|}, \\ \mathcal{P}_{\mathcal{T},\mathcal{H}} &= \left[\mathbb{P}\left(h_j^O, \tau_i^O \middle| h_j^A, \tau_i^A\right)\right]_{i,j} \in \mathbb{R}^{|\mathcal{T}|\times|\mathcal{H}|}, \\ \forall a, o \quad \mathcal{P}_{\mathcal{T},ao,\mathcal{H}} &= \left[\mathbb{P}\left(h_j^O, \tau_i^O o \middle| h_j^A, \tau_i^A a\right)\right]_{i,j} \in \mathbb{R}^{|\mathcal{T}|\times|\mathcal{H}|}.\end{aligned}$$

In [14], authors proved that $\mathcal{P}_{\mathcal{T},\mathcal{H}}$ and $\mathcal{P}_{\mathcal{T},ao,\mathcal{H}}$ have rank at most $d := |\mathcal{Q}|$. In addition, they showed that if,

$$U, S, V = \mathcal{SVD}_d(\mathcal{P}_{\mathcal{T},\mathcal{H}}), \tag{3}$$

where $\mathcal{SVD}_d$ is the thin singular value decomposition (SVD), in a way that U contains only the columns associated to the d largest singular values, then, $J = U^\top R$ with $R = [\mathbf{r}_{\tau_i}^{\mathcal{T}\top}]_i$ is an invertible linear transformation. In addition, noting by $X^\dagger = (X^\top X)^{-1}X^\top$ the Moore pseudo-inverse of a matrix X, we have,

$$\mathbf{b}_\star = U^\top \mathcal{P}_{\mathcal{T},\mathcal{H}}\mathbf{1}, \tag{4}$$

$$\mathbf{b}_\infty^\top = \mathcal{P}_{\mathcal{H}}^\top (U^\top \mathcal{P}_{\mathcal{T},\mathcal{H}})^\dagger, \tag{5}$$

$$\forall ao \quad B_{ao} = U^\top \mathcal{P}_{\mathcal{T},ao,\mathcal{H}}(U^\top \mathcal{P}_{\mathcal{T},\mathcal{H}})^\dagger. \tag{6}$$

As $\mathbf{b}_t$ is a sufficient statistic to predict the outcome of any test, we called it the belief at time t. Equations (1) to (6) still hold when working with features of tests $\phi^{\mathcal{T}}$ and histories $\phi^{\mathcal{H}}$ [14]. The matrices $\mathcal{P}_{\mathcal{H}}$, $\mathcal{P}_{\mathcal{T},\mathcal{H}}$, and $\mathcal{P}_{\mathcal{T},ao,\mathcal{H}}$ will no longer contain probabilities but rather expected values of features or products of features. We can recover the probability matrices using indicator functions of tests and histories as features.

$$\mathcal{P}_{\mathcal{H}} = \mathbb{E}\left[\phi_t^{\mathcal{H}} \middle| h_t^A\right], \qquad \mathcal{P}_{\mathcal{T},\mathcal{H}} = \mathbb{E}\left[\phi_t^{\mathcal{T}} \phi_t^{\mathcal{H}\top} \middle| h_t^A, \tau_t^A\right],$$

$$\forall a, o \quad \mathcal{P}_{\mathcal{T},ao,\mathcal{H}} = \mathbb{E}\left[\phi_{t+1}^{\mathcal{T}} \phi_t^{\mathcal{H}\top} \middle| a_t = a, o_t = o\right].$$

3. THE LEARNING ALGORITHM

The underlying idea of spectral learning is to start with a large set of tests $\mathcal{T}$ such that it is core; then by spectral decomposition to recover the PSR parameters up to a linear transformation as described in the previous section. In practice, the set $\mathcal{T}$ is constructed from all fixed-size sequences of action-observation pairs appearing in the learning trajectory. Same is done to build the set of histories $\mathcal{H}$.

The algorithm works by building empirical estimators $\hat{\mathcal{P}}_{\mathcal{H}}$, $\hat{\mathcal{P}}_{\mathcal{T},\mathcal{H}}$ and $\hat{\mathcal{P}}_{\mathcal{T},ao,\mathcal{H}}$ of the matrices $\mathcal{P}_{\mathcal{H}}$, $\mathcal{P}_{\mathcal{T},\mathcal{H}}$ and $\mathcal{P}_{\mathcal{T},ao,\mathcal{H}}$. Usually, to gather independent samples, the system must allow to be reset. When reset is not available [15], approximate estimators can be built from a single trajectory by dividing it into subsequences. This approach, called the suffix-algorithm, produces still good estimators in practice. When samples are generated from a non-blind policy (that depends on past observations given past actions), unbiased estimators are obtained by importance sampling [16].

In previous algorithms [14], $\hat{\mathcal{P}}_{\mathcal{T},\mathcal{H}}$ is directly plugged into eq. (3). However, identifying the low dimensional subspace from the noisy matrix $\hat{\mathcal{P}}_{\mathcal{T},\mathcal{H}}$ can cause severe errors, if some dimensions needed to explain the dynamics are absent [17]. Moreover, eq. (3) needs an estimate of that subspace size. The lowest is the dimension, the most compact the model will be but one has to account for the quantity of noise in order to select a subspace with an appropriate size. If errors due to sampling are large, searching for a small subspace can lead to large approximation errors, defined as the distance between the true subspace and the learned one. Indeed, some dimensions can be used to model the noise instead of the system dynamics. In this case, a bigger subspace that maps both the system dynamics and a part of the noise will achieve better performance because the noise can be next reduced during regression by means of regularization. The trade off between approximation and estimation errors is handled by solving a rank minimization problem [18].

In contrast to the previous approaches, $\hat{\mathcal{P}}_{\mathcal{T},\mathcal{H}}$ and all $\hat{\mathcal{P}}_{\mathcal{T},ao,\mathcal{H}}$ matrices are used for subspace identification. First, we build matrices F and $\hat{F}$ of size $|\mathcal{T}| \times (|\mathcal{H}|\,(1 + |\mathcal{A}|\,|\mathcal{O}|))$ by stacking all these matrices as follows,

$$F = \left(\mathcal{P}_{\mathcal{T},\mathcal{H}}, \mathcal{P}_{\mathcal{T},a_1o_1,\mathcal{H}}, \ldots, \mathcal{P}_{\mathcal{T},a_{|\mathcal{A}|}o_{|\mathcal{O}|},\mathcal{H}}\right)$$

$$\hat{F} = \left(\hat{\mathcal{P}}_{\mathcal{T},\mathcal{H}}, \hat{\mathcal{P}}_{\mathcal{T},a_1o_1,\mathcal{H}}, \ldots, \hat{\mathcal{P}}_{\mathcal{T},a_{|\mathcal{A}|}o_{|\mathcal{O}|},\mathcal{H}}\right)$$

Notice that the linear property of the PSR implies that F is low rank. Precisely, $d := \mathrm{rank}(F) = \mathrm{rank}(\mathcal{P}_{\mathcal{T},\mathcal{H}})$.

Let $\|X\|_\star = \sum_i \sigma(X)_i$ be the nuclear norm and $\|X\|_{frob}$ the Frobenius norm of a matrix X, we define the following rank minimization problem

$$\begin{aligned} &\min_{F \in \mathbb{R}^{n\times m}} \mathrm{rank}(F) \\ &\text{subject to } \left\|F - \hat{F}\right\|_{frob} \leq \delta, \end{aligned} \tag{P_1}$$

which is an NP-hard problem. Minimizing the nuclear norm, defined as the sum of the singular values of a matrix, provides a tractable alternative. In some sense [19], problem (P_2) is the tightest convex relaxation of (P_1). Let $\tilde{F}$ be the solution of,

$$\min_{F \in \mathbb{R}^{n\times m}} \mu\|F\|_* + \frac{1}{2}\left\|F - \hat{F}\right\|_{frob}^2, \tag{P_2}$$

where μ highlights the trade off between ensuring that $\tilde{F}$ is low rank and sticks to observations $\hat{F}$. We propose a close form solution. Considering the SVD of the matrix X of rank r,

$$X = USV^\top, \qquad \Sigma = \mathrm{diag}(\{\sigma_i\}_{1\leq i\leq r}).$$

For each $\tau \geq 0$, we introduce the SVT operator,

$$\mathcal{D}_\tau(X) = U\mathcal{D}_\tau(\Sigma)V^\top, \qquad \mathcal{D}_\tau(\Sigma) = \mathrm{diag}(\{\sigma_i - \tau\}_+),$$

where $x_+ = \max\{0, x\}$. It has been shown [19] that $\tilde{F} = \mathcal{D}_\mu(\hat{F})$ achieves the minimum in (P_2). In the experiments, we took $\mu = \sigma_2(\hat{F})/2$, as the first dimension often corresponds to normalization.

Once the low rank estimators $\tilde{\mathcal{P}}_{\mathcal{T},\mathcal{H}}$, $\tilde{\mathcal{P}}_{\mathcal{T},ao,\mathcal{H}}$ are extracted from $\tilde{F}$, we can find the low dimension subspace by plugging them, with $\hat{\mathcal{P}}_{\mathcal{H}}$, into eqs. (3) to (6),

$$\tilde{U}, \tilde{S}, \tilde{V} = \mathcal{SVD}_l(\tilde{F}),$$

$$\tilde{\mathbf{b}}_1 = \tilde{U}\tilde{\mathcal{P}}_{\mathcal{T},\mathcal{H}}\mathbf{e},$$

$$\tilde{\mathbf{b}}_\infty^\top = \hat{\mathcal{P}}_{\mathcal{H}}^\top (\tilde{U}\tilde{\mathcal{P}}_{\mathcal{T},\mathcal{H}})^{\lambda\dagger},$$

$$\forall ao \quad \tilde{B}_{ao} = \tilde{U}\tilde{\mathcal{P}}_{\mathcal{T},ao,\mathcal{H}}(\tilde{U}\tilde{\mathcal{P}}_{\mathcal{T},\mathcal{H}})^{\lambda\dagger},$$

where σ_l is the smallest positive singular value of $\tilde{\mathcal{P}}_{\mathcal{T},\mathcal{H}}$ and $X^{\lambda\dagger} = (X^\top X + \lambda I)^{-1}X^\top$ the regularized Moore pseudo-inverse. In the experiments, we performed a grid search on λ and chose $\lambda = 10^{-7}$.

In an online setting, we want to regularly update the PSR parameters when new observations are coming in order to improve the quality of predictions. Ideally, one would like to perform one update at each time step. In the experiments, updates are scheduled at regular time intervals. Let superscripts t indicate the variables computed from observations up to time t. After each update, a new subspace is identified and the current belief has to be projected on it. So, in Bayesian filtering (eq. (2)), before multiplying the current predictive

state $\mathbf{b}_t^t$ with the new matrices $\tilde{B}_{ao}^{t+1}$ and $\tilde{\mathbf{b}}_\infty^{t+1\top}$, we need to project it to the new subspace and renormalizing it:

$$\tilde{\mathbf{b}}_t^{t+1} = \frac{\tilde{U}^{t+1\top}\tilde{U}^t\tilde{\mathbf{b}}_t^t}{\tilde{\mathbf{b}}_\infty^{t+1\top}\tilde{U}^{t+1\top}\tilde{U}^t\tilde{\mathbf{b}}_t^t}, \qquad \tilde{\mathbf{b}}_{t+1}^{t+1} = \frac{\tilde{B}_{ao}^{t+1}\tilde{\mathbf{b}}_t^{t+1}}{\tilde{\mathbf{b}}_\infty^{t+1\top}\tilde{B}_{ao}^{t+1}\tilde{\mathbf{b}}_t^{t+1}}. \quad (7)$$

Indeed, by left multiplying the current belief by $\tilde{U}^t$, we recover the probabilities of all the tests in $\mathcal{T}$. Then, by left multiplying by $\tilde{U}^{t+1\top}$, we combine linearly these probabilities to obtain the current belief in the new subspace of tests. However, at each update, a bit of information contained in the belief lies outside the new subspace and is then lost after the projection. That is why, we need to normalize in order to ensure that $\tilde{\mathbf{b}}_\infty^{t+1\top}\tilde{\mathbf{b}}_t^{t+1} = 1$. These procedures introduce some errors in the belief which will be propagated during filtering. However, Bayesian filtering is able to correct a wrong prior as long as new observations are made. These two effects compete with each other. Fortunately, errors introduced by re-projecting the current belief are related to $||\hat{\mathcal{P}}_{\mathcal{T},\mathcal{H}} - \mathcal{P}_{\mathcal{T},\mathcal{H}}||_{frob}$ and decrease with time. So, errors in Bayesian filtering will reduce with time. Note that the efficient update algorithm in [20] could be adapted to our case to speed up computation.

4. APPLICATION TO RADAR ILLUMINATIONS

In this section we explain how to learn a set (one per radar) of PSRs modeling random illumination patterns, predict the next observation and finally schedule the next frequency band to listen. First, we consider only one radar for the sake of clarity. The observed illumination pattern is modeled as a random process $\{O_t\}$, where $o_t = 1$ if an illumination is intercepted at time t and $o_t = 0$ otherwise. We assume that the time step duration is longer than an illumination. Actions are also binary variables. A positive action ($a_t = 1$) consists in listening at time t a frequency band covering the radar frequency. In order to get a core set of tests, we have to use long tests and histories. Theoretically, tests and histories longer than a period when concatenated is sufficient. In practice, we used tests and histories whose length are around three times the period. We denote by $l(\mathcal{H})$ (resp. $l(\mathcal{T})$) the length of histories (resp. tests). Working with long tests or histories increases exponentially the number of tests or histories and so the size of estimated matrices. To keep the problem tractable we use features of tests and histories,

$$\phi_t^{\mathcal{H}} = (o_{t-l(\mathcal{H})}, \ldots, o_t)^\top \in \{0,1\}^{l(\mathcal{H})},$$
$$\phi_t^{\mathcal{T}} = (o_{t+1}, \ldots, o_{t+l(\mathcal{T})+1})^\top \in \{0,1\}^{l(\mathcal{T})}.$$

Note that, we also removed past and future actions in the features as they represent very few information.

During an initialization phase, taken in the experiment of length $2(l(\mathcal{H}) + l(\mathcal{T}))$, no predictions are made and observations are collected through uniform random sampling to estimate the PSR parameters. This initialization phase helps to avoid stability issues with inaccurate estimates introduced by SVD. In practice this initialization could be replaced by using some prior information on the radar. Then, after each c steps, the PSR parameters are updated to include new observations. The new parameters are then used to predict and filter the current belief. During each update, the current belief has to be projected on to the new learnt subspace.

At time t, a prediction $\mathbb{P}(O_{t+1} = 1 | A_{t+1} = 1, b_t)$ is made for each radar corresponding to the probability of intercepting an illumination given that we listen at time $t+1$ to a right frequency band. Now, we consider K frequency bands. Let $\mathcal{B}_k$ be the set of radars emitting in the k-th frequency band. Let $p_t(i)$ be the prediction $\mathbb{P}(O_{t+1} = 1 | A_{t+1} = 1, b_t)$ for the i-th radar. We denote by $\tilde{a}_t \in \{1, ..., K\}$ the action corresponding to listening the $\tilde{a}_t$-th frequency band at time t. Thus, for any PSR corresponding to a radar the perceived action is

$$a_t = \begin{cases} 1, & \text{if the radar is covered by the frequency band } \tilde{a}_t \\ 0, & \text{otherwise.} \end{cases}$$

In the planning experiment, the goal is to intercept as many illuminations as possible. So the score is the total number of illuminations intercepted. In an operational context, we could associate weights to radars with little changes. In order to intercept illuminations both to maximize the score (exploiting the learned PSRs parameters) and to learn more accurate PSRs parameters (exploring), we built a stochastic random strategy inspired by works on adversarial bandits [21]. This strategy mixes with the parameter γ a uniform random exploration and an exponentially weighted one-step lookahead exploitation policy. Let T be a parameter, for M radars, the stochastic strategy is

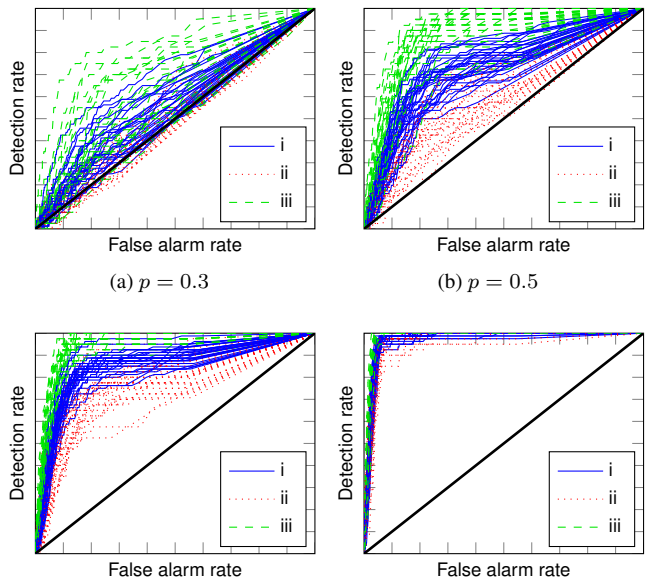
$$\mathbb{P}\left(\tilde{A}_t = k\right) = \frac{(1-\gamma)\exp\left(\frac{\gamma}{T}\sum_{m\in\mathcal{B}_k} p_t(m)\right)}{\sum_{l=1}^K \exp\left(\frac{\gamma}{T}\sum_{m\in\mathcal{B}_l} p_t(m)\right)} + \frac{\gamma}{K}. \quad (8)$$

5. EXPERIMENTS

We designed two experiments. In the first one, we consider only one radar whose illuminations are partially intercepted through a random uniform sampling strategy. At each time step, the radar is sensed with probability p and a prediction on the next observation is made. After the initialization phase, PSR parameters are updated every $c = 50$ steps. The radar signal is a binary signal with a period of 50 time step. In addition, we corrupt the period with a jitter of 10%. We measure the quality of one-step-ahead predictions learned online from past observations. Results are given in Figure 1 using empirical Receiver Operating Characteristic (ROC) curves which represents the detection ratio depending on the false alarm ratio. Detections and false alarms are computed on three parts of the trajectory : (i) on the first half of the trajectory, (ii) on the whole trajectory, (iii) on the second half of the trajectory. The trajectory length is 4000 and the initialization phase equals 300.

The curves show that predictions accuracy improves over time by considering the difference between (ii) and (iii). This experiment highlights that detection rate increases when fewer illuminations are missed. This is encouraging for online learning, because we can expect that, by closing the learning planning loop, a good stochastic policy will reduce the number of missed illuminations.

In the second experiment, $M = 9$ radars equally distributed in $K = 3$ frequency bands are considered. The period of each radar is randomly drawn in $[40; 80]$ according to a uniform distribution. The jitter is set to be 5% of the period. Tests and histories length is 240. During the initialization phase of 480 time steps, bands are drawn uniformly. After, we used the policy described in eq. (8). The length of the trajectory is 10000. The Figure 2 shows the average on radars of the number of detected illuminations on the total number of illuminations depending on the time. Results are averaged on 30 simulations and presented with 95% asymptotic confidence intervals computed from a Normal distribution. The red dotted line correspond to the performance of a uniform strategy. The figure shows that online learning allows a dynamic scheduling strategy that uses past interceptions to both catch more illuminations and learn a



(a) $p = 0.3$ (b) $p = 0.5$

(c) $p = 0.8$ (d) $p = 1$

Fig. 1. ROC curves for different parameters p. Each curve represents one simulation over the 30 conducted.

stochastic model of the environment. In our oversimplified settings, it results in a scanning strategy whose performance improves with time over the uniform random scheduling.

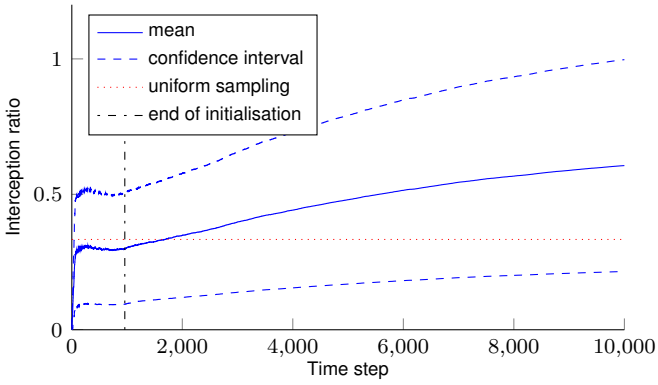


Fig. 2. Detection ratio. The stochastic policy is the one of eq. (8), with $\gamma = 0.2$ and $T = 0.001$.

6. DISCUSSION

In this paper, we modeled a sensor management problem, encountered in ES, a sequential decision making problem in a partially observable environment. We especially focused on online learning and planning. Combining recent advances in nuclear norm minimization and consistent spectral learning algorithms, we proposed a new algorithm to learn PSRs and a scanning strategy. Nuclear norm minimization allows adapting the model size to the underlying process and collected samples which is very useful to ensure good performances over time in an online setting. We detailed how to perform Bayesian filtering with evolving PSR parameters. The first experiment demonstrates the learning performance of our algorithm. While, the second one runs an online learning and planning problem, occurring in ES. Our scanning strategy closed loop improves over time compared with the usual random scheduling.

7. REFERENCES

[1] B. Dutertre, "Dynamic scan scheduling," in *Real-Time Systems Symposium, 2002. RTSS 2002. 23rd IEEE*, 2002, pp. 327–336.

[2] Emin Koksal, *Periodic Search Strategies For Electronic Countermeasure Receivers With Desired Probability Of Intercept For Each Frequency Band*, Ph.D. thesis, Middle East Technical University, 2010.

[3] C. Winsor and E.J. Hughes, "Optimisation and evaluation of receiver search strategies for electronic support," *Radar, Sonar Navigation, IET*, vol. 6, no. 4, pp. 233–240, April 2012.

[4] Richard G. Wiley, *ELINT: The Interception and Analysis of Radar Signals*, Artech House Publishers, 2006.

[5] I.V.L. Clarkson, "Optimal periodic sensor scheduling in electronic support," *Proceedings of the Defence Applications of Signal Processing (DASP'05)*, 2005.

[6] I. V. L. Clarkson, E. D. El-Mahassni, and S. D. Howard, "Sensor scheduling in electronic support using markov chains," *Radar, Sonar and Navigation, IEEE Proceedings on*, vol. 153, no. 4, pp. 325 –332, Aug. 2006.

[7] I.V.L. Clarkson, "Synchronisation in scan-on-scan-on-scan problems," *Proceedings of the Defence Applications of Signal Processing (DASP'09)*, 2009.

[8] I.V.L. Clarkson and A.D. Pollington, "Performance limits of sensor-scheduling strategies in electronic support," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 43, no. 2, pp. 645 –650, Apr. 2007.

[9] S.W. Kelly, G.P. Noone, and J.E. Perkins, "Synchronization effects on probability of pulse train interception," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 32, no. 1, pp. 213 –220, Jan. 1996.

[10] E. D. El-Mahassni and G. P. Noone, "A new way of estimating radar pulse intercepts," *ANZIAM J.*, vol. 45, pp. C448–C460, June 2004.

[11] Y. Xun, M.M. Kokar, and K. Baclawski, "Control based sensor management for a multiple radar monitoring scenario," *Information Fusion*, vol. 5, no. 1, pp. 49–63, 2004.

[12] Yong Xun, M.M. Kokar, and K. Baclawski, "Using a task-specific qos for controlling sensing requests and scheduling," in *Network Computing and Applications, 2004. (NCA 2004). Proceedings. Third IEEE International Symposium on*, Aug 2004, pp. 269–276.

[13] Matthew Rosencrantz, Geoff Gordon, and Sebastian Thrun, "Learning low dimensional predictive representations," in *Proceedings of the twentyfirst International Conference on Machine Learning (ICML-04)*. ACM, 2004, p. 88.

[14] Byron Boots, Sajid M Siddiqi, and Geoffrey J Gordon, "Closing the learning-planning loop with predictive state representations," *The International Journal of Robotics Research*, vol. 30, no. 7, pp. 954–966, 2011.

[15] Britton Wolfe, Michael R James, and Satinder Singh, "Learning predictive state representations in dynamical systems without reset," in *Proceedings of the Twentysecond International Conference on Machine Learning (ICML-05)*. ACM, 2005, pp. 980–987.

[16] Michael Bowling, Peter McCracken, Michael James, James Neufeld, and Dana Wilkinson, "Learning predictive state representations using non-blind policies," in *Proceedings of the Twentythird international Conference on Machine learning (ICML-06)*. ACM, 2006, pp. 129–136.

[17] Alex Kulesza, N Raj Rao, and Satinder Singh, "Low-rank spectral learning," in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS-14)*, 2014, pp. 522–530.

[18] Hadrien Glaude, Olivier Pietquin, and Cyrille Enderli, "Subspace identification for predictive state representation by nuclear norm minimization," in *Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL-14)*, 2014.

[19] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[20] Byron Boots and Geoffrey J Gordon, "An online spectral learning algorithm for partially observable nonlinear dynamical systems," in *Proceedings of the Twentyfifth AAAI Conference on Artificial Intelligence (AAAI-11)*, 2011.

[21] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire, "The nonstochastic multiarmed bandit problem," *SIAM Journal on Computing*, vol. 32, no. 1, pp. 48–77, 2002.

[22] Britton Wolfe, Michael R James, and Satinder Singh, "Predictive representations of state," in *Proceedings of the Fifteenth Conference on Neural Information Processing Systems (NIPS-01)*, pp. 1555–1561, 2001.

[23] A Anandkumar, N Michael, A K Tang, and A Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," in *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, pp. 731-745, 2011.

[24] Cem Tekin, and Liu Mingyan, "Online learning of rested and restless bandits," in *IEEE Transactions on Information Theory*, vol. 58, no. 8, pp. 5588–5611, 2012.

[25] Cem Tekin, and Liu Mingyan, ""Adaptive learning of uncontrolled restless bandits with logarithmic regret," in *the Forty-ninth Annual Conference on Communication, Control, and Computing*, 2011.

[26] K. Liu, and Q. Zhao "Distributed Learning in Multi-Armed Bandit with Multiple Players," in *IEEE Transactions on Signal Processing*, vol. 58, no. 11, pp. 5667–5681, 2010.

[27] Ronald Ortner, Daniil Ryabko, Peter Auer, and Rémi Munos "Regret bounds for restless Markov bandits," in *Theoretical Computer Science*, vol. 558, pp. 62–76, 2014.