Optimal Policies Search for Sensor Management

Thomas Bréhard, Emmanuel Duflos, Philippe Vanheeghe, Pierre-Arnaud Coquelin

Optimal Policies Search for Sensor Management : Application to the ESA Radar

Thomas Bréhard
R&D Department
Huge Corporation
Gigantica, France

Emmanuel Duflos
Philippe Vanheeghe
Ecole Centrale de lille
LAGIS UMR CNRS 8146
INRIA Lille - Nord Europe
Project Team SequeL
Cité Scientifique - BP 46
59851, Villeneuve d'Ascq Cedex, France
Email: emmanuel.duflos@ec-lille.fr

Pierre-Arnaud Coquelin
Predict and Control
INRIA Lille - Nord Europe
Parc Scientifique de la Haute Borne
40 Avenue Halley
59650, Villeneuve d'Ascq Cedex, France

***Abstract*— This paper introduces a new approach to solve sensor management problems. Classically sensor management problems can be well formalized as Partially-Observed Markov Decision Processes (POMPD). The original approach developped here consists in deriving the optimal parameterized policy based on stochastic gradient estimation. We assume in this work that it is possible to learn the optimal policy off-line (in simulation) using models of the environement and of the sensor(s). The learned policy can then be used to manage the sensor(s). In order to approximate the gradient in a stochastic context, we introduce a new method to approximate the gradient, based on Infinitesimal Approximation (IPA). The effectiveness of this general framework is illustrated by the managing of an Electronically Scanned Array Radar.**

Keywords: Sensor(s) Management, Partially Observable Markov Decision Process, Stochastic Gradient Estimation, AESA Radar.

Sensor(s) Management Special Session

I. Introduction

Years after years the complexity and the performances of many sensors have increased leading to more and more complex sensor(s)-based systems which supply the decision centers with an increasing amount of data. The number, the types and the agility of sensors along with the increased quality of data far outstrip the ability of a human to manage them: it is often difficult to compare how much information can be gained by way of a given management scheme [1]. It results from this the necessity to derive unmanned sensing platforms that have the capacity to adapt to their environment [2]. This problem is often refered as the *Sensor(s) Management Problem*. In more simple situations, the operational context may lead to works on sensor(s) management like in the *radar - infrared sensor* case [3]. A general definition of this problem could then be : sensor management is the effective use of available sensing and database capabilities to meet the mission goals. Many applications deal with military applications,a classical one being to detect, to track and tp identify smart targets (a smart target can change its way of moving or its way of sensing when it detects it is under analysis) with several sensors. The questions are then the following at each time: how must we group the sensors, how long, in which direction, and with which functioning mode? The increasing complexity of the targets to be detected, tracked and identified, makes the management even more difficult and led to the development of researches on the definition of an optimal sensor management scheme in which the targets and the sensors are treated altogether in a complex dynamic system [4].

Sensor Management has become very popular this last years and many approaches can be found in the litterature. In [5] and [6] the authors use a the modelling of the detection process of an Electronically Scanned Array (ESA) Radar to propose management scheme during the detection step. In [7]–[9] an information-based approach is use to manage a set of sensors. From a theorical point of view the sensor management can be modelled as a Partially Observable Markov Decision Process (POMDP) [10]–[12]. Whatever the underlying application, the sensor management problem consists in choosing at each time t an action A_t within the set $\mathcal{A}$ of available actions. The choice of A_t is generally based on the density state vector X_t describing the environment of the system and variables of the system itself. It is generally assumed that the state or at least a part of this state is Markovian. Moreover in most of the applications, we only have access to a partial information of the state and X_t must be estimated from the measurements $\{Y_s\}_{1\leq s\leq t}$. This estimation process is often derived within a Bayesian framework where we use state-dynamics and observation models such as:

$$X_{t+1} = F(X_t, A_t, N_t) \tag{1}$$

$$Y_t = H(X_t, W_t) \tag{2}$$

where N_t, W_t, F and H respectively stands for the state noise, the measurements noise, the state-dynamics and the measurement function. F and H are generally time varying functions. The control problem consists in finding the scheduling policy π i.e. select A_t given the past and the possible

futures. However, this control problem may have a theorical solution, it is generally untractable in practice. However few works propose optimal solution in the frame of POMDPs like [12]. Beside, several works have been carried out to find sub-optimal policies like for instance myopic policies. Reinforcement Learning and Q-Learning have also been used to propose a solution ([13], [14]).

We propose in this paper to look for a policy within a class of parametrized policy π_θ and to learn it which means learn the optimal value of θ. Funding our work on the approach described in [15] we assume that it is possible to learn this policy *in simulation* using models of the overall system. Once the optimal parameter has been found it is used to manage the sensor(s). The frame of this work being the detection and localization of targets, we show in the last part of this paper how it may be applied the the management of an ESA radar.

The section II described the modelling of a sensor management problem using a POMDP approach. In the section III we derive the algorithm to learn the parameter of the policy.In section IV we show how this method may be used for the tasking of an ESA radar. Finally section V exhibits firts simulations results.

II. Modelling

A. POMDP Modelling

Let us consider three measurable continuous spaces denoted by $\mathcal{X}$, $\mathcal{A}$ and $\mathcal{Y}$, $\mathcal{X}$ is called the *state space*, $\mathcal{Y}$ the observation space and $\mathcal{A}$ the *action space*. We call $\mathcal{M}(\mathcal{X})$ the set of all the measures defined on $\mathcal{X}$. A Partially-Observable Decision Proceess is defined by a *state process* $(X_t)_{t\geq 0} \in \mathcal{X}$, an *observation process* $(Y_t)_{t\geq 1} \in \mathcal{Y}$ and a set of actions $(A_t)_{t\geq 1} \in \mathcal{A}$. In these definitions t stands usually for the *time*. The state process is an homogeneous Markov chain with initial probability measure $\mu(dx_0) \in \mathcal{M}(\mathcal{X})$ and with Markov transition kernel $K(dx_{t+1}|x_t)$ ([16]):

$$\forall t \geq 0, X_{t+1} \sim K(\cdot|X_t) \tag{3}$$

$$X_0 \sim \mu \tag{4}$$

$(Y_t)_{t\geq 1}$ is called the observation is linked with the state process by the conditional probability measure:

$$\mathcal{P}(Y_t \in dy_t|X_t = x_t) = g(x_t, y_t)\, dy_t \tag{5}$$

where $g : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ is the marginal density function of Y_t given X_t. In a general way, the state process evolves continuously with respect to time t whereas the observations are made at sampled time t_n. A new observation is used to derive a new action. We will therefore consider in the following the processes $(X_t)_{t\geq 0}$, $(Y_n)_{n\in\mathbb{N}}$, $(A_n)_{n\in\mathbb{N}}$ where n stands for the index of the observation. We also assume that there exists two generative functions $F_\mu : U \to \mathcal{X}$ and $F : \mathcal{X} \times U \to \mathcal{X}$, where $(U, \sigma(U), \nu)$ is a probability space, such that for any measurable *test function* f defined over $\mathcal{X}$ we have:

$$\int_{\mathcal{X}} f(x_t)K(dx_t|x_{t-1}) \quad = \quad \int f(F(x_{t-1}, u))\nu(du) \tag{6}$$

and

$$\int_{\mathcal{X}} f(x_0)\mu(dx_0) \quad = \quad \int f(F_\mu(u))\nu(du). \tag{7}$$

In many practical situations, $U = [0, 1]^{n_U}$, and u is a n_U-uple of pseudo random numbers generated by a computer. For sake of simplicity, we adopt the notations $K(dx_0|x_{-1}) \triangleq \mu(dx_0)$ and $F(x_{-1}, u) \triangleq F_\mu(u)$. Under this framework, the Markov Chain $(X_t)_{t\geq 0}$ is fully specified by the following dynamical equation:

$$X_{t+1} = F(X_t, U_t),\ U_t \overset{i.i.d.}{\sim} \nu \tag{8}$$

The observation process $(Y_n)_{n\in\mathbb{N}}$, defined on the measurable space $(\mathcal{Y}, \sigma(Y))$, is actually linked with the state process by the following conditional probability measure

$$\mathbb{P}(Y_n \in dy_n|X_{t_n} = x_{t_n}, A_n) = g(y_n, x_{t_n}, A_n)\lambda(dy_t) \tag{9}$$

where $A_n \in \mathcal{A}$ is defined on the measurable space $(\mathcal{A}, \sigma(A))$ and λ is a fixed probability measure on $(\mathcal{Y}, \sigma(Y))$. As we assume that observations are conditionally independent given the state process we cab write $\forall 1 \leq i, j \leq t,\ i \neq j$:

$$\begin{aligned}\mathbb{P}(Y_i \in dy_i, Y_j \in dy_j|X_{0:t}, A_i, A_j) = \\ \mathbb{P}(Y_i \in dy_i|X_{0:t}, A_i)\mathbb{P}(Y_j \in dy_j|X_{0:t}, A_j)\end{aligned} \tag{10}$$

where we have adopted the usual notation $z_{i:j} = (z_k)_{i\leq k\leq j}$.

B. Filtering distribution in a Partially-Observable Markov Decision Process

Given a sequence of action $A_{1:n}$ and a sample trajectory of the observation process $y_{1:n}$ and indices $\{n_1, n_2, t_1, t_2\}$ such that $1 \leq n_1 \leq n_2 \leq n$ and $0 \leq t_1 \leq t_{n_1} \leq t_{n_2} \leq t_2 \leq t_n$, we define, using the , the posterior probability distribution $M_{t_1:t_2|n_1:n_2}(dx_{t_1:t_2})$ by ([17]):

$$\mathbb{P}(X_{t_1:t_2} \in dx_{t_1:t_2}|Y_{n_1:n_2} = y_{n_1:n_2}, A_{n_1:n_2}) \tag{11}$$

Using the Feynman-Kac framework, the probabilty 11 can be written:

$$\frac{\prod_{t=t_1}^{t_2} K(dx_t|x_{t-1}) \prod_{j=n_1}^{n_2} G_{t_j}(x_{t_j})}{\int_{\mathcal{X}^{t_2-t_1}} \prod_{t=t_1}^{t_2} K(dx_t|x_{t-1}) \prod_{j=n_1}^{n_2} G_{t_j}(x_{t_j})}, \tag{12}$$

where for simplicity's sake, $G_{t_n}(x_{t_n}) \triangleq g(y_n, x_{t_n}, A_n)$ and $G_0(x_0) \triangleq 0$. One of the main interest here is to estimate the state at time t from noisy observations $y_{1:n_t}$ with n_t the index of the last observation just before time t. From a bayesian point of view this information is completely contained in the so-called *filtering distribution* $M_{t:t|1:n_t}$. In the following, the filtering distribution will simply be denoted as M_t.

C. Numerical methods for estimating the filtering distribution

Given a measurable test function $f : \mathcal{X} \to \mathbb{R}$, we want to evaluate

$$M_t(f) = \mathbb{E}[f(X_t)|Y_{1:n_t} = y_{1:n_t}, A_{1:n_t}] \tag{13}$$

which is equal, using the Feynman Kac framework, to:

$$\frac{\mathbb{E}[f(X_t)\prod_{j=1}^{n_t} G_{t_j}(X_{t_j})]}{\mathbb{E}[\prod_{j=1}^{n_t} G_{t_j}(X_{t_j})]} \tag{14}$$

In general, it is impossible to find $M_t(f)$ exactly except for simple cases such as linear/gaussian (using Kalman filter) or for finite state space Hidden Markov Models. In the general dynamics, continuous space case considered here, possible numerical methods for computing $M_t(f)$ include the Extended Kalman filter, quantization methods, Markov Chain Monte Carlo methods and Sequential Monte Carlo methods (SMC), also called particle filtering. The basic SMC method, called Bootstrap Filter, approximates $M_t(f)$ by an empirical distribution $M_t^N(f) = \frac{1}{N}\sum_{i=1}^N f(x_i^N)$ made of N so-called *particles* ([18]). It consists in a sequence of transition/selection steps: at time t, given observation y_t ([15]):

- **Transition step:** (also called **importance sampling** or **mutation**) a successor particles population $\widetilde{x}_t^{1:N}$ is generated according to the state dynamics from the previous population $x_{t-1}^{1:N}$. The (importance sampling) weights $w_t^{1:N} = \frac{g(\widetilde{x}_t^{1:N}, y_t)}{\sum_{j=1}^N g(\widetilde{x}_t^j, y_t)}$ are evaluated.
- **Selection step:** Resample (with replacement) N particles $x_t^{1:N}$ from the set $\widetilde{x}_t^{1:N}$ according to the weights $w_t^{1:N}$. We write $x_t^{1:N} = \widetilde{x}_t^{k_t^{1:N}}$ where $k_t^{1:N}$ are the selection indices.

Resampling is used to avoid the problem of degeneracy of the algorithm, i.e. that most of the weights decreases to zero. It consists in selecting new particle positions such as to preserve a consistency property :

$$\sum_{i=1}^N w_t^i \phi(\widetilde{x}_t^i) = \mathbf{E}[\frac{1}{N}\sum_{i=1}^N \phi(x_t^i)] \tag{15}$$

The simplest version introduced in [19] consists in choosing the selection indices $k_t^{1:N}$ by an independent sampling from the set $1:N$ according to a multinomial distribution with parameters $w_t^{1:N}$, i.e. $\mathbb{P}(k_t^i = j) = w_t^j$, for all $1 \le i \le N$. The idea is to replicate the particles in proportion to their weights. The reader can find some convergence results of $M_t^N(f)$ to $M_t(f)$ (e.g. Law of Large Numbers or Central Limit Theorems) in [17], but for our purpose we note that under weak conditions on the test function and on the HMM dynamics, we have the asymptotic consistency property in probability, i.e. $\lim_{N\to\infty} M_t^N(f) \stackrel{\mathbb{P}}{=} M_t(f)$.

III. Policy Learning Algorithm

A. Optimal Parameterized Policy for Partially-Observable Markov Decision Process

Let R_t be a real value reward function

$$R_t \triangleq R(X_t, M_t(f)) \,. \tag{16}$$

The goal is to find a policy

$$\pi : \mathcal{A}^n \times \mathcal{Y}^n \to \mathcal{A} \tag{17}$$

that maximizes the criterion performance :

$$J_\pi = \int_0^T \mathbb{E}[R_t]dt \tag{18}$$

where T is the duration of the scenario. Designing in practice policies that depend on the whole trajectory of the past observations/actions is unrealistic. It has been proved that the class of stationary policies that depend on the filtering distribution conditionally to past observations/actions M_t contains the optimal policy. In general the filtering distribution is an infinite dimensional object, and it cannot be represented in a computer and so is the policy. We therefore propose to look for the optimal policy in a class of parameterized policies $(\pi_\alpha)_{\alpha\in\Gamma}$ that depend on a statistic of the filtering distribution :

$$A_{n+1} = \pi_\alpha(M_{t_n}(f)) \tag{19}$$

where f is any test function. As the policy π is parameterized by α, the performance criterion now depends only on α. Thus we can maximize it by achieving a stochastic gradient ascent with respect to α :

$$\alpha_{k+1} = \alpha_k + \eta_k \nabla J_{\alpha_k}, \quad k \ge 0 \tag{20}$$

where ∇J_{α_k} denotes the gradient of J_{α_k} w.r.t α_k. By convention ∇J_{α_k} is column vector whose i-th component is the partial derivative with respect to α_i. $(\eta_k)_{k\ge 0}$ is a non-increasing positive sequence tending to zero. We present in the two following subsection a possible approach to estimate ∇J_{α_k} based on Infinitesimal Perturbation Analysis (IPA).

B. Infinitesimal Perturbation Analysis for gradient estimation

We assume that we can write the following equality at each k:

$$\nabla J_\alpha = \int_0^T \nabla_\alpha \mathbb{E}[R_t]dt \tag{21}$$

Proposition 1: We have the following decomposition of the gradient

$$\begin{aligned} \nabla_\alpha \mathbb{E}[R_t] &= \mathbb{E}[M_t(fS_t)\nabla_{M_t(f)} R_t] \\ &- \mathbb{E}[M_t(f)M_t(S_t)\nabla_{M_t(f)} R_t] \\ &+ \mathbb{E}[R_t S_t] \end{aligned} \tag{22}$$

where

$$S_t = \sum_{j=1}^{p_t} \frac{\nabla_\alpha G_{t_j}(X_{t_j})}{G_{t_j}(X_{t_j})} \tag{23}$$

Proof: First let us rewrite $\nabla_\alpha \mathbb{E}[R_t]$ as following:

$$\nabla_\alpha \mathbb{E}[R_t] = \nabla_\alpha \int_{\mathcal{X}^t \times \mathcal{Y}^{n_t}} R_t U_t V_t \prod_{j=1}^{n_t} \lambda(dy_j) \tag{24}$$

where:

$$\left\{ \begin{array}{ll} U_t & = \prod_{i=0}^{t} K(dx_i|x_{i-1}) \ , \\ V_t & = \prod_{j=1}^{n_t} G_{t_j}(x_{t_j}) \end{array} \right. . \quad (25)$$

Remarking that only R_t and V_t depends on α so that we obtain

$$\left\{ \begin{array}{ll} \nabla_\alpha V_t & = S_t V_t \ , \\ \nabla_\alpha R_t & = \nabla_\alpha M_t(f) \nabla_{M_t(f)} R_t \end{array} \right. \quad (26)$$

where S_t is given by eq.(23). Incorporating (24) in (26), we obtain

$$\nabla_\alpha \mathbb{E}[R_t] = \mathbb{E}[\nabla_\alpha M_t(f) \nabla_{M_t(f)} R_t] \ . + \mathbb{E}[R_t S_t] \ . \quad (27)$$

Now using one more time (26), we have

$$\begin{aligned} \nabla_\alpha M_t(f) &= \nabla_\alpha \mathbb{E}\Big[f(X_t)\frac{V_t}{\mathbb{E}[V_t]}\Big] \\ &= \mathbb{E}\Big[f(X_t)\frac{\nabla_\alpha V_t}{\mathbb{E}[V_t]}\Big] - \mathbb{E}\Big[f(X_t)\frac{V_t \mathbb{E}[\nabla_\alpha V_t]}{\mathbb{E}[V_t]^2}\Big] \\ &= \mathbb{E}\Big[f(X_t) S_t \frac{V_t}{\mathbb{E}[V_t]}\Big] - M_t S_t \mathbb{E}\Big[\frac{V_t}{\mathbb{E}[V_t]}\Big] \\ &= M_t(fS_t) - M_t(f) M_t(S_t) \end{aligned} \quad (28)$$

so that we obtain (22) by incorporating (28) in (27). ∎

We can deduce directly Algorithm 1 from (22). It is important to note that we must deal with two time-scales. This first and the shorter one allows to simulate the continuous state X_t. On the contrary the observation and action process are updated only each time we get a new observation. These specific time is denoted t_n in the algorithm. That is the eason while there is an alternative to update the variables S_t and $\tilde{w}_{t-1}^{(i)}$. A new action A_n is also calculated each t_n as already explained above. One can also be surprised to calculate $R(X_t, M_t(f))$ using the sampled value of X_t. To well understand this algorithm we must remind that **the learning is made off-line** using a simulated process. It is therefore possible to use the *real* value of X_t in this case.

Algorithm 1 Policy Gradient in POMDP via IPA

Initialize $\alpha_0 \in \Gamma$
for $k = 1$ **to** ∞ **do**
  for $t = 1$ **to** T **do**
    Sample $u_t \sim \nu$
    Set $x_t = F(x_{t-1}, u_t)$,
    If $t = t_n$, sample $y_n \sim g(., x_t, a_n)\lambda(.)$
    Set $s_t = \left\{ \begin{array}{ll} s_{t-1} + \frac{\frac{\partial g}{\partial \alpha}(x_t, y_n, a_n)}{g(x_t, y_n, a_n)} & \text{if } t = t_n \\ s_{t-1} & \text{else} \end{array} \right.$
    Set $\forall i \in \{1 \dots, I\}$
      $\tilde{x}_t^{(i)} = F(x_{t-1}^{(i)}, a_{t-1}, u_t^{(i)})$ where $u^{(i)} \overset{iid}{\sim} \nu$
      $\tilde{s}_t^{(i)} = \left\{ \begin{array}{ll} s_{t-1}^i + \frac{\frac{\partial g}{\partial \alpha}(x_t^i, y_n, a_n)}{g(x_t^{(i)}, y_n, a_n)} & \text{if } t = t_n \\ s_{t-1}^i & \text{else} \end{array} \right.$
      $\tilde{w}_t^{(i)} = \left\{ \begin{array}{ll} \frac{g(x_t^{(i)}, y_n, a_n)\tilde{w}_{t-1}^{(i)}}{\sum_j g(x_t^{(j)}, y_n, a_n)\tilde{w}_{t-1}^{(j)}} & \text{if } t = t_n \\ \tilde{w}_{t-1}^{(i)} & \text{else} \end{array} \right.$
    Set $(x_t^{(i)}, s_t^{(i)})_{i \in \{1,\dots,I\}} = (\tilde{x}_t^{(i)}, \tilde{s}_t^{(i)})_{i \in \{k_1,\dots,k_I\}}$, $k_{1:I}$ are selection indices associated to $(\tilde{w}^{(i)})_{i \in \{1,\dots,I\}}$,
    $m_t(f) = \frac{1}{I}\sum_i f(x_t^{(i)})$, $m_t(s_t) = \frac{1}{I}\sum_i s^{(i)}$, $m_t(fs_t) = \frac{1}{I}\sum_i f(x_t^{(i)}) s_t^{(i)}$,
    $a_{n+1} = \pi_{\alpha_k}(m_t)$ if $t = t_n$
    $r_t = R(x_t, m_t(f))$
    $\nabla r_t = (m_t(fs_t) - m_t(f) m_t(s_t)) \frac{\partial R}{\partial m_t(f)}(x_t, m_t(f)) + r_t s_t$
    $\nabla J_{\alpha_k} = \nabla J_{\alpha_k} + \nabla r_t$
  end for
  $\alpha_{k+1} = \alpha_k + \eta_k \nabla J_{\alpha_k}$
end for

IV. Application to the ESA Radar

The ESA is an agile beam radar which means that it is able to point its beam in any direction of the environnement almost instantaneously without inertia. However, the targets in the environement are detected w.r.t a probability of detection which depends on the direction of the beam and the time of observation in this direction. In the following, we precise first the nature of an action, then the influence of the action onto the probability of detection and finally the nature of the observations.

Definition of the action: The main property of an ESA is that it can point its beam without mechanically adjusting the antenna. An ESA radar provides measurements in a direction θ. We note δ, the time of observation in this direction. In this work the the n-th action is :

$$A_n = \begin{bmatrix} \theta_n & \delta_n \end{bmatrix}^T \quad (29)$$

with

$$\left\{ \begin{array}{l} \theta_n \in [-\frac{\pi}{2}, \frac{\pi}{2}] \ , \\ \delta_n \in \mathbb{R}^+ \end{array} \right. \quad \forall n \geq 0 \ . \quad (30)$$

This is a simple possible action. One could increase the number of componenets of an action by adding the emitted frequency for instance. The action does not influence directly the observation produced by the ESA but the probability of detection of a target.

The probability of detection P_d: It refers to the probability to detect a target and therefore to the probability to obtain an estimation of the state of a target p at time t_n denoted $X_{t_n,p}$ with action A_n. In this work, $X_{t_n,p}$ is composed of the localisation and velocity components of the target p at time t_n in the x-y plane:

$$X_{t_n,p} = \begin{bmatrix} rx_{t_n,p} & ry_{t_n,p} & vx_{t_n,p} & ry_{t_n,p} \end{bmatrix}^T \quad (31)$$

where the subscript T stands for *matrix transpose*. The terms $rx_{t_n,p}$ and $ry_{t_n,p}$ refers here to the position and $vx_{t_n,p}$ and $vy_{t_n,p}$ the velocity of target p at time t_n. We also denote $D_{n,p}$ the random variable which takes values 1 if the radar produces a detection (and therefore an estimation) for target p and 0 else :

$$D_n = \begin{bmatrix} D_{n,1} & \dots & D_{n,P} \end{bmatrix}^T \ . \quad (32)$$

As said previously, this probability also depends on the time of observation δ_n. Aerial targets being considering here, the reflectivity of a target can be modelled using a Swerling I model [20]. We then have the following relation between the probability of detection and the probability of false alarm P_{fa} (i.e. the probability that the radar produce a detection knowing that there is no target) ([5], [21]):

$$P_d(x_{t_n,p}, A_n) = P_{fa}^{\frac{1}{1+\rho(x_{t_n,p},A_n)}} \tag{33}$$

where $\rho(x_{t_n,p}, A_n)$ is the target signal-to-noise ratio. In the case of an ESA radar, it is equal to :

$$\rho(x_{t_n,p}, A_n) = \alpha\delta_n \frac{cos^2\theta_n}{r_{t_n,p}^4} e^{-\frac{(\beta_{t_n,p}-\theta_n)^2}{2B^2}} \tag{34}$$

where $r_{t_n,p}$ is the target range and $\beta_{t_n,p}$ the azimuth associated to target p at instant time t_n. α is a coefficient which includes all the parameters of the sensor and B is the beamwidth of the radar. It is reminded in Appendix A how the equations 33 and 34 may be derived. If we make the assumption that all the detections are independant, we can write :

$$\mathbb{P}(D_n = d_n | X_{t_n} = x_{t_n}, A_n) = \prod_p^P \mathbb{P}(D_{n,p} = d_{n,p} | X_{t_n,p} = x_{t_n,p}, A_n) \tag{35}$$

where

$$\mathbb{P}(D_{n,p} = d_{n,p} | X_{t_n,p} = x_{t_n,p}, A_n) = P_d(x_{t_n,p}, A_n)\delta_{d_{n,p}=1} + (1 - P_d(x_{t_n,p}, A_n))\delta_{d_{n,p}=0} \tag{36}$$

Observation equation: At instant time t_n, the radar produces a raw observation Y_n composed of P measurements :

$$Y_n = \begin{bmatrix} Y_{n,1} & \dots & Y_{n,P} \end{bmatrix}^T . \tag{37}$$

where $Y_{n,p}$ is the observation related to target of state value $x_{t_n,p}$ obtained with action A_n (we do not consider here the problem of measurement-target association). Moveover, we assume that the number of targets P is known. Each of these measurements has the following formulation :

$$Y_{n,p} = \begin{bmatrix} r_{n,p} & \beta_{n,p} & \dot{r}_{n,p} \end{bmatrix}^T \tag{38}$$

where $r_{n,p}$, $\beta_{n,p}$, $\dot{r}_{n,p}$ are range, azimuth and range rate. The equation observation can be written

$$\mathbb{P}(Y_n \in dy_n | X_{t_n} = x_{t_n}, A_n) = \tag{39}$$

$$\prod_p^P \mathbb{P}(Y_{n,p} \in dy_{n,p} | X_{t_n,p} = x_{t_n,p}, A_n) \tag{40}$$

where

$$\mathbb{P}(Y_{n,p} \in dy_{n,p} | X_{t_n,p} = x_{t_n,p}, A_n) \tag{41}$$

$$= g(y_{n,p}, x_{t_n,p}, A_n)\lambda(dy_{n,p}) \tag{42}$$

$$g(y_{n,p}, x_{t_n,p}, A_n) = \begin{pmatrix} \mathcal{N}\left(h_t(x_{t_n,p}), \Sigma_y\right) P_d(x_{t_n,p}, A_n) \\ 1 - P_d(x_{t_n,p}, A_n) \end{pmatrix}^T \tag{43}$$

and

$$\lambda(dy_{n,p}) = \lambda_{cont}(dy_{n,p}) + \lambda_{disc}(dy_{n,p}) \tag{44}$$

The relation between the state and the raw observations is given by :

$$Y_{n,p} = h_{t_n}(X_{t_n,p}) + W_{n,p} \tag{45}$$

with $h_{t_n}(x_{t_n,p})$ equals to:

$$\begin{pmatrix} \sqrt{(rx_{t_n,p} - rx_{t_n}^{obs})^2 + (ry_{t_n,p} - ry_{t_n}^{obs})^2} \\ \text{atan}\left\{\frac{ry_{t_n,p} - ry_{t_n}^{obs}}{rx_{t_n,p} - rx_{t_n}^{obs}}\right\} \\ \frac{(rx_{t_n,p} - rx_{t_n}^{obs})(vx_{t_n,p} - vx_{t_n}^{obs}) + (ry_{t_n,p} - ry_{t_n}^{obs})(vy_{t_n,p} - vy_{t_n}^{obs})}{\sqrt{(rx_{t_n,p} - rx_{t_n}^{obs})^2 + (ry_{t_n,p} - ry_{t_n}^{obs})^2}} \end{pmatrix} \tag{46}$$

and $W_{n,p}$ a gaussian noise the covariance matrix of which is given by :

$$\Sigma_y = diag(\sigma_r^2, \sigma_\beta^2, \sigma_{\dot{r}}^2) . \tag{47}$$

State equation: First let us introduce the definition of the unknown state X_t at time t and its evolution through time. $X_{t,p}$ is the state of the target p. It has been defined above. Let P be the known number of targets in the space under analysis at time t. X_t has the following form: .

$$X_t = \begin{bmatrix} X_{t,1} & \dots & X_{t,P} \end{bmatrix}^T \tag{48}$$

Based on [22] works, we classically assume that all the targets follow a nearly constant velocity model. We use a discretized version of this model ([23]) :

$$X_{t,p} = F(X_{t-1,p}, U_t) \text{ where } U_t \sim \mathcal{N}\left(0, \sigma^2 Q\right) \tag{49}$$

where

$$F = \begin{bmatrix} 1 & 0 & \beta & 0 \\ 0 & 1 & 0 & \beta \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \text{ and } Q = \begin{bmatrix} \frac{\beta^3}{3} & 0 & \frac{\beta^2}{2} & 0 \\ 0 & \frac{\beta^3}{3} & 0 & \frac{\beta^2}{2} \\ \frac{\beta^2}{2} & 0 & \beta & 0 \\ 0 & \frac{\beta^2}{2} & 0 & \beta \end{bmatrix} . \tag{50}$$

V. Simulations

VI. Conclusion

Appendix A

We show in this Appendix how the probability of detection is derived. First, the radar transmits a pulse expressed as follows

$$s(t) = \alpha(t)\cos(w_c t) \tag{51}$$

$$= \text{Re}\{\alpha(t)e^{jw_c t}\} \tag{52}$$

where $\alpha(t)$ is the envelope also called the transmitted pulse and w_c the carrier frequency. This pulse is modified by the process of reflection. A target is modelled as a set of elementary reflectors, each reflecting: time delayed, Doppler shift,

Phase shift and attenuated version of the transmitted signal. We usually assume that the reflection process is linear and frequency independent within the bandwidth of the transmitted pulse. The return signal has the following formulation:

$$s_r(t) = G\sum_i \alpha(t-\tau_i) g_i e^{j(w_c(t-\tau_i+\frac{2\dot{r}_i}{c}t)+\theta_i)} + n(t) \quad (53)$$

where

- g_i is the radar cross section associated to reflector i,
- θ_i is the phase shift associated to reflector i,
- $\dot{r}_i$ is the radial velocity between the antenna and the object (Doppler frequency shift),
- G: others losses heavily range dependent due to spatial spreading of energy,
- $n(t)$ is a thermal noise of the receiver such that $\text{Re}\{n(t)\}, \text{Im}\{n(t)\} \sim \mathcal{N}(0, \sigma_n^2)$.

We make the following approximations:

$$\begin{cases} \dot{r}_i \approx \dot{r} \\ \alpha(t-\tau_i) \approx \alpha(t-\tau) \end{cases} \quad (54)$$

where $\dot{r}$ is the mean radial velocity of the target τ is the mean time delay of the target. Using these approximations, the return signal can be rewritten as follows:

$$s_r(t) = \alpha(t-\tau) G e^{jw_D t} b + n(t) \quad (55)$$

where

$$\begin{cases} w_D &= w_c(1+\frac{2\dot{r}_i}{c}) \\ b &= \sum_i g_i e^{j(-w_c\tau_i+\theta_i)} \end{cases} . \quad (56)$$

The fluctuations of b are known and modelled using Swerling 1 model [20]. There are differents models availables (Swerling 1, 2, 3,...) corresponding to different types of targets. Swerling 1 given below is convenient for aircrafts. We can then write :

$$\text{Re}\{b\}, \text{Im}\{b\} \sim \mathcal{N}(0, \sigma_{RCS}^2) . \quad (57)$$

This modelling of b assumes that the phase shifts θ_i are independent and uniformly distributed and the magnitudes g_i are identically distributed. If the number of reflector is large, the central limit theorem gives that b is a complex-valued Gaussian random variable centered at the origin. Now, a matching filter is applied to our return signal

$$s_m(t) = \int_{-\infty}^{+\infty} s_r(t) h(s) ds \quad (58)$$

where $h(t)$ is a shifted, scaled and reversed copy of $s_r(t)$

$$h(s) = \alpha(\delta - t) e^{-jw_D(\delta - t)} . \quad (59)$$

We choose $t = \delta + \tau$ which yields the best signal to noise ratio where δ is the length of the transmitted pulse. The probability of detection is based on quantity $|s_m(\delta+\tau)|^2$. We can show that

$$s_m(\delta+\tau) = G e^{jw_D\tau} b + \int_{-\infty}^{+\infty} n(\delta+\tau-s) h(s) ds . \quad (60)$$

One can remark that $s_m(\delta+\tau)$ is the sum of two complex-value Gaussian variables. We look at the following statistic

$$\Lambda = \frac{|s_m(\delta+\tau)|^2}{2\sigma_n^2} \quad (61)$$

and we introduce the following notation

$$\sigma_s^2 = G^2 \sigma_{RCS}^2 \quad (62)$$

Now we construct the test

$$\begin{cases} \mathcal{H}_1 : \text{data generated by signal + noise} \\ \mathcal{H}_0 : \text{data generated by noise} \end{cases} \quad (63)$$

$$\begin{cases} \mathcal{H}_1 : p_\Lambda(x) = \frac{1}{\frac{\sigma_s^2}{\sigma_n^2}+1} e^{-\frac{x}{\frac{\sigma_s^2}{\sigma_n^2}+1}} \\ \mathcal{H}_0 : p_\Lambda(x) = e^{-x} \end{cases} \quad (64)$$

Then, we derive the probability of detection and false alarm.

$$\begin{cases} P_d = \int_\gamma^{+\infty} p_\Lambda(x|\mathcal{H}_1 \text{ is true}) = e^{-\frac{\gamma}{\frac{\sigma_s^2}{\sigma_n^2}+1}} \\ P_{fa} = \int_\gamma^{+\infty} p_\Lambda(x|\mathcal{H}_0 \text{ is true }) = e^{-\gamma} \end{cases} \quad (65)$$

Consequently

$$P_d = P_{fa}^{\frac{1}{\frac{\sigma_s^2}{\sigma_n^2}+1}} \quad (66)$$

The ratio $\frac{\sigma_s^2}{\sigma_n^2}$ is called the Signal-to-Noise Ration noted ρ. This SNR is related to the parameters of the system and the target. The classical radar equation is given by the following formula ([21]):

$$\rho = \frac{P_t G_t G_r \lambda^2 \sigma}{(4\pi)^3 r^4} \quad (67)$$

where P_t is the energy of the transmitted pulse, G_t is the gain of the transmitted antenna, G_r is the gain of the received antenna, σ is the radar cross section (for an aircraft between 0.1 and $1\ m^2$), r is the target range, γ is the system noise temperature and L is a general loss term. However, the above formula does not take into account for the sake of simplicity the losses due to atmospheric attenuation and to the imperfection of the radar. Thus , extra terms must be added :

$$\rho = \frac{P_t G_t G_r \lambda^2 \sigma}{(4\pi)^3 k b L \gamma r^4} \quad (68)$$

where b is the receiver noise bandwith (generally consider equal to the signal bandwidth so that $b = \frac{1}{\delta_t}$), k is Boltzmann's constant, γ is the temperature of the system and L some losses. Moreover, the gain reduces with the deviation of the beam from the antenna normal in an array antenna.

$$G_t = G_0 cos^\alpha(\theta_t) , \quad (69)$$

$$G_r = G_0 cos^\alpha(\theta_t) \quad (70)$$

where G_0 is the gain of the antenna. In [24], $\alpha = 2$, in [21], $\alpha = 2.7$. According [25], there is also a beam loss because the

radar beam is not pointing directly so that the radar equation is:

$$\rho = \frac{P_t G_0^2 \lambda^2 \sigma \delta_t \cos^2(\theta_t)}{(4\pi)^3 k L \gamma r^4} e^{-\frac{(\theta_t - \beta_t)^2}{2B^2}} \quad (71)$$

where is B is the beamwidth.

REFERENCES

[1] M. K. Kalandros, L. Trailović, L. Y. Pao, and Y. Bar-Shalom, "Tutorial on multisensor management and fusion algorithms for target tracking," in *Proceeding of the 2004 American Control Conference Boston, Massachusetts June 30 - July 2*, 2004, pp. 4734–4748. [Online]. Available: http://vehicle.me.berkeley.edu/~caveney/C3UV/papers/MSTrackingTutorialACC04.pdf

[2] S. Ji, R. Parr, and L. Carin, "Nonmyopic multiaspect sensing with partially observable markov decision processes," *IEEE Transactions on Signal Processing*, vol. 55, no. 6, pp. 2720–2730, June 2007.

[3] S. Blackman and R. Popoli, *Design and Analysis of Modern Tracking Systems*. Artech House Publishers, 1999.

[4] A. Doucet, B. Vo, C. Andrieu, and M. Davy, "Particle filtering for multi-target tracking and sensor management," *Proceedings of ISIF*, pp. 474–481, 2005.

[5] E. Duflos, M. deVilmorin, and P. Vanheeghe, "Time allocation of a set of radars in a multitarget environment," in *Proceedings of FUSION 2007 Conference*, I. S. on Information Fusion, Ed. Quebec (Canada): International Society on Information Fusion, July 2007.

[6] T. Huguerre, E. Duflos, T. Bréhard, and P. Vanheeghe, "An optimal detection strategy for esa radars," in *Proceedings of the COGnitive systems with Interactive Sensors Conference*, d. l. e. d. T. d. l. e. d. l. C. Société de l'Electricité, Ed. Société de l'Electricité, de l'Electronique et des Technologies de l'Information et de la Communication, November 2007.

[7] K. Kastella, "Discrimination gain to optimize detection and classification," *IEEE Transaction on Systems, Man and Cybernetics - Part A : Systems and Human*, vol. 27, no. 1, pp. 112–116, January 1997.

[8] M. Kolba and L. Collins, "Information based sensor management in the presence of uncertainty," *IEEE Transactions on Signal Processing*, vol. 55, no. 6, pp. 2731–2735, June 2007.

[9] A. Khodayari-Rostamabad and S. Valaee, "Information theoric enumeration and tracking of multiple sources," *IEEE Transactions on Signal Processing*, vol. 55, no. 6, pp. 2765–2773, June 2007.

[10] C. Kreucher, D. Blatt, A. Hero, and K. Kastella, "Adaptive multimodality sensor scheduling for detection and tracking of smart targets," in *The 2004 Defense Applications of Signal Processing Workshop (DASP), October 31 - November 5*, 2004. [Online]. Available: http://www.eecs.umich.edu/~hero/Preprints/2004DASP.pdf

[11] V. Krishnamurthy, "Algorithms for optimal scheduling and management of hidden markov model sensors," *IEEE Transactions on Signal Procesing*, vol. 50, no. 6, pp. 1382–1397, June 2002.

[12] V. Krishnamurthy and D. Djonin, "Structured threshold policies for dynamic sensor scheduling - a partially observed markov decision process approach," *IEEE Transactions on Signal Procesing*, vol. 55, no. 10, pp. 4938–4957, October 2007.

[13] R. S. Sutton and A. G. Barto, "Time-derivative models of pavlovian reinforcement," *Learning and Computational Neuroscience: Foundations of Adaptive Networks, M. Gabriel and J. Moore Eds.*, 1990.

[14] C. Kreucher and A. Hero., "Non-myopic approaches to scheduling agile sensors for multitarget detection, tracking, and identification," in *The Proceedings of the 2005 IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP) Special Section on Advances in Waveform Agile Sensor Processing, volume V, March 18 - 23*, 2005, pp. 885–888. [Online]. Available: http://www.eecs.umich.edu/~hero/Preprints/2005ICASSP_a.pdf

[15] P. Coquelin and R. Munos, "Particle filter - based policy gradient in pomdps," February 2008, submitted at ICML'08.

[16] P. Coquelin, R. Deguest, and R. Munos, "Numerical methods for sensitivity analysis of feynman-kac models," INRIA-Futurs, Tech. Rep., 2007.

[17] P. D. Moral, *Feynman-Kac Formulae Genealogical and Interacting Particle Systems with Applications*. Springer, 2004.

[18] A. Doucet, S. Godsill, and C. Andrieu, "On Sequential Monte Carlo Sampling Methods for Bayesian Filtering," Cambridge University Engineering Department, Tech. Rep., 2000.

[19] N. Gordon, D. Salmond, and A. Smith, "Novel approach to nonlinear and non-gaussian bayesian state," *Proceedings IEE-F*, pp. 107–113, 1993.

[20] G. Curry, *Radar System Performance Modeling, Second Edition*. Artech House, 2005.

[21] J. Wintenby, "Resource allocation in airborn surveillance radar," Ph.D. dissertation, Chalmers University of Technology, 2003.

[22] X. Rong Li and V. Jilkov, "A Survey of Maneuvering Target Tracking Part I: Dynamics Models," IEEE *Trans. on Aerospace and Electronic Systems*, vol. 39, no. 4, pp. 1333–1364, October 2003.

[23] J.-P. L. Cadre and O. Tremois, "Bearings-only tracking for maneuvering sources," IEEE *Trans. on Aerospace and Electronic Systems*, vol. 34, no. 1, pp. 179–193, January 1998.

[24] M. D. Vilmorin, "Contribution à la grstion optimale de capteurs: application à la tenue de situations aériennes," Ph.D. dissertation, Ecole Centrale de Lille et Université des Sciences et Technologie de Lille, 2002.

[25] G. V. Keuk and S. Blackman, "On Phased-Array Radar Tracking and Parameter Control," IEEE *Trans. on Aerospace and Electronic Systems*, vol. 1, no. 29, pp. 186–194, January 1993.