# POMDP MULTI-ARMED BANDIT FORMULATION FOR ENERGY MINIMIZATION IN SENSOR NETWORKS

*Vikram Krishnamurthy*

Department of Electrical and Computer Engineering
University of British Columbia, Vancouver, V6T 1Z4, Canada.
email: vikramk@ece.ubc.ca

## ABSTRACT

In Network Centric Warfare, sensor platforms with active sensing equipment such as radars can betray their existence, by emitting energy that can be intercepted by enemy surveillance sensors thereby increasing the vulnerability of the entire combat system. To achieve the important tactical requirement of low probability of intercept (LPI), requires dynamically controlling the emission energy of sensors. In this paper we propose computationally efficient dynamic emission control and management algorithms for multiple networked heterogenous sensors. By formulating the problem as a partially observed Markov decision process with an on-going multi-armed bandit structure, near optimal sensor management algorithms are developed for controlling the active sensor emission to minimize the threat.

## 1. INTRODUCTION

The Joint Vision 2010 [1] is the conceptual template for US Armed forces. One of the fundamental themes underlying the Joint Vision 2010 is the concept of Network Centric Warfare (NCW). The tenets of NCW are [1]: (i) A robustly networked force improves information sharing; (ii) Information sharing enhances the quality of information and shared situational awareness; (iii) Shared situational awareness enables collaboration and self-synchronization, and enhances sustainability and speed of command; (iv) These, in turn, dramatically increase mission effectiveness.

The information for generating battlespace awareness in NCW is provided by numerous sources, for example, stand-alone intelligence, surveillance, and reconnaissance platforms and sensors employed on weapons platforms. In the fundamental shift to network-centric operations, sensor networks emerge as a key enabler of increased combat power. Networked sensors have several advantages including decreased time to engagement, increased ability to detect low signature targets, improved track accuracy and continuity, improved target detection and identification and reduced sensor detectability to the enemy [2].

In this paper, we focus on this reduced sensor detectability aspect of NCW. We present *decentralized sensor management* algorithms for reducing the detectability of networked sensor platforms to the enemy. We consider the problem of how to dynamically manage and control the emission energy of active sensors in multiple platforms to minimize the threat posed to these platforms in combat situations. In the defense literature the acronym EMCON is used for *emission control*. Emission management/control is emerging in importance due to the essential tactical necessity of sensor platforms satisfying a *low probability of intercept* (LPI) requirement. This LPI requirement is in response to the increase in capability of modern intercept receivers to detect and locate platforms that radiate active sensors. The emission management system needs to dynamically plan and react to an uncertain dynamic battlefield environment.

The aim of this paper is to answer the following question: *How should the sensor manager achieve EMCON by dynamically deciding which platforms (or group of platforms) are to radiate active sensors at each time instant in order to minimize the overall threat posed to all the platforms while simultaneously taking into account the cost of radiating these sensors and the quality of service they provide?* Unlike platform centric warfare where scheduling of sensors is carried out within a platform, the above aim is consistent with the philosophy of network centric warfare where given a network of several platforms, the sensor manager dynamically makes a local decision as to which platforms should radiate active sensors.

## 2. MULTI-PLATFORM EMISSION CONTROL (EMCON) PROBLEM

The network centric multi-platform system we consider in this paper consists of three sub-systems: networked sensor platforms, a sensor manager which decides which platform (or group of platforms) should radiate active sensors, and a threat evaluator which yields information about the threat

This research was supported by NSERC and the British Columbia Advanced Systems Institute (BCASI)

posed to the active platform, see Fig.1.

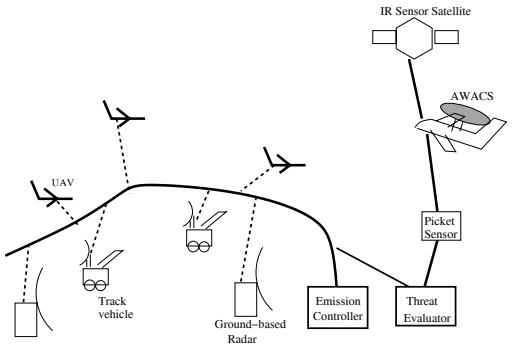


**Fig. 1**. Schematic setup consisting of 3 types of networked platforms (unmanned aerial vehicles (UAVs), Track vehicles and Ground based Radar), Threat Evaluator (IR Sensor Satellite, AWACS, Picket sensors) and EMCON.

### 2.1. Heterogeneous Networked Sensor Platforms

Consider $P$ heterogeneous sensor platforms indexed by $p = 1, \ldots, P$. Active sensors (e.g., radar) are typically linked with the deployment of weapon systems whereas passive sensors (e.g., ESM, ELINT (electronic intelligence), FLIR (forward looking infra-red radar), imagers) are often used for surveillance. We assume that at each time instant only one platform (or group of platforms) is allowed to radiate active sensors and the other $P-1$ platforms can only use passive sensors. This assumption is justified since in a network of sensor platforms, certain groups of sensor platforms are always operated together. For example, in multi-static radar sensor groups, alternately one radar sensor transmits while all of the other distributed networked sensors are used as receivers simultaneously.

### 2.2. Emission Level Impact (ELI)

Let $k = 0, 1, 2, \ldots$, denote discrete time. At each time instant $k$ the sensor manager decides which platform to activate. Let $u_k \in \{1, \ldots, P\}$ denote the platform that is activated by the sensor manager at time $k$. Denote the *emission level impact* (ELI) of platform $p$ at time $k$ as $s_k^{(p)}$. The ELI of platform $p$ is the cumulative received emission registered by the enemy sensors from platform $p$ until time $k$:

$$s_{k+1}^{(p)} = s_k^{(p)} + e_{k+1}^{(p)}, \quad p \in \{1, \ldots, P\}. \tag{1}$$

Here, $e_k^{(p)}$ denotes the *instantaneous (incremental) emission* registered at the enemy from platform $p$ at time $k$. The ELI is a surrogate measure for the effectiveness of the LPI feature of the sensor platform - the larger the ELI $s_k^{(p)}$, the worse the LPI feature of the sensor platform. Due to the uncertainty in modelling of how the enemy registers the ELI, $\{e_k^{(p)}\}$ and hence $\{s_k^{(p)}\}$ are assumed to be random processes. Assume that the ELI $s_k^{(p)}$ is quantized to a finite set $\{1, 2, \ldots, \mathcal{N}_p\}$ where the values in the finite set correspond to physical ELI values, e.g., 1 is low, 2 is medium and 3 is high. Given that the ELI $s_k^{(p)}$ is finite state and at any time instant $k$ depends on the ELI at the previous time instant (1), it is natural to model the evolution of $\{s_k^{(p)}\}$ probabilistically as a finite state Markov chain. It is clear from (1) that the ELI $s_k^{(u_k)}$ of the platform (or group of sensors) radiating active sensors at time $k$, evolves with time. The ELI of the platforms that only use passive sensors remain approximately constant since the sensors do not emit energy that can be intercepted by the enemy, i.e, $e_k^{(p)}$ is small when $p \neq u_k$. Thus: If $u_k = p$, the ELI $s_k^{(p)}$ evolves according to an $\mathcal{N}_p$-state homogeneous Markov chain with transition probability matrix $A^{(p)} = (a_{ij}^{(p)})_{i,j \in \mathcal{N}_p} = \mathbf{P}\left(s_{k+1}^{(p)} = j \mid s_k^{(p)} = i\right)$. The states of all the other $(P-1)$ platforms using passive only sensors are unaffected, i.e., $s_{k+1}^{(p)} = s_k^{(p)}$, if platform $p$ only uses passive sensors at time $k$, or equivalently $A^{(p)} = I \quad$ if $p \neq u_k$.

### 2.3. Threat Evaluator

In battlefield environments, the ELI $\{s_k^{(p)}\}$, $p = 1, \ldots, P$, registered by the enemy is not directly available to our sensor manager. We assume that local sensors on each platform $p$ together with a centralized threat evaluation system share information over the network to compute an *observed threat level* posed to each platform $p = 1, \ldots, P$ – which is a probabilistic function of the ELI as described below. The centralized threat evaluation system typically comprises of an infrared (IR) sensor satellite satellite, ground based picket sensors, surveillance sensor network and AWACS (Airborne Warning and Control System) aircraft that observe the behaviour of the enemy. Fig.1 shows the schematic setup.

Let $z_k^{(p)}$ denote the *observed cumulative threat* posed to platform $p$ at time $k$. Then the process $\{z_k^{(p)}\}$ evolves with time for each platform $p$ as $z_{k+1}^{(p)} = z_k^{(p)} + y_{k+1}^{(p)}, \quad p \in \{1, \ldots, P\}$ where $y_k^{(p)}$ denotes the observed *instantaneous (incremental) threat* posed to platform $p$ at time $k$. Clearly the threat posed to any platform $p$ is a function of the ELI of the platform. Thus the instantaneous threat $y_k^{(p)}$ is a probabilistic function of the instantaneous emission $e_k^{(p)}$. For example, one possible model for the instantaneous threat is $y_k^{(p)} = s_k^{(p)} - s_{k-1}^{(p)} + t_k^{(p)} + w_k^{(p)}$ where $t_k^{(p)}$ is a positive valued *incremental trend* process – which could be deterministic, e.g., $t_k^{(p)} = 1$ for all time $k$, or a stationary process

that is statistically independent of $w_k^{(p)}$ (defined below) and $s_k^{(p)}$. Hence the cumulative threat $z_k^{(p)}$ posed to platform $p$ typically monotonically increases with time $k$. $w_k^{(p)}$ denotes the observation noise and takes into account several factors such as measurement errors in the surveillance sensors and incomplete knowledge and uncertainty about the enemy.

We assume $y_k^{(p)}$ is quantized to a finite set $\{1, 2, \ldots, \mathcal{M}_p\}$ where, for example, 1 denotes a small increment, 2 a medium increment, and 3 a large increment in the threat level. The observed threat $y_k^{(p)}$ is a probabilistic function of the instantaneous emission $e_k^{(p)} = s_k^{(p)} - s_{k-1}^{(p)}$. This probabilistic relationship is summarized by the $(\mathcal{N}_p \times \mathcal{N}_p)$ likelihood matrices $B^{(p)}(1), \ldots, B^{(p)}(\mathcal{M}_p)$,

$$B^{(p)}(m) = (b_{ijm}^{(p)})_{i,j \in \mathcal{N}_p}, \quad b_{ijm}^{(p)} \stackrel{\triangle}{=} \mathbf{P}(y_{k+1}^{(p)} = m | s_k^{(p)} = i, s_{k+1}^{(p)} = j)$$

denotes the conditional probability (symbol probability) of the threat evaluator generating an observed threat symbol of $m$ when the instantaneous emission is $e_k^{(p)} = s_{k+1}^{(p)} - s_k^{(p)}$. If the platform $p$ is inactive, i.e., $p \neq u_k$, then since the emission $e_k^{(p)} = s_k^{(p)} - s_{k-1}^{(p)}$ is zero it follows that $b_{ijm}^{(p)} = 0$ for $i \neq j$. Thus $B^{(p)}(m) = I \quad$ if $p \neq u_k$.

Let $Y_k = (y_1^{(u_0)}, \ldots, y_k^{(u_{k-1})})$ denote the observed threat history and $U_k = (u_0, \ldots, u_k)$ denote the sequence of past decisions made by the EMCON functionality of the sensor manager. The above formulation captures the essence of a network centric system – the sensor manager controls different sensors in different platforms. This is in contrast to the older concept of platform centric systems where individual platforms have their own sensor managers.

### 2.4. Network Sensor Manager and Cost

The above probabilistic model for the sensor platform, emission level impact (ELI) and threat evaluator together constitute a controlled Hidden Markov Model (HMM) [3]. Here we address the fundamental issue of how the sensor manager should dynamically decide which platform (or group of platforms) should radiate active sensors at each time instant to minimize a suitable cost function that encompasses all the platforms. The EMCON functionality of the sensor manager decides which platform to activate at time $k$, based on the optimization of a discounted cost function which we now detail: The instantaneous cost incurred at time $k$ due to all the deployed platforms (both active and passive ) is

$$C_k = c(s_k^{(u_k)}, s_{k-1}^{(u_k)}, y_k^{(u_k)}, u_k) + \sum_{p \neq u_k} r(s_k^{(p)}, s_{k-1}^{(p)}, y_k^{(p)}, p) \tag{2}$$

where $c(s_k^{(u_k)}, s_{k-1}^{(u_k)}, y_k^{(u_k)}, u_k)$ denotes the cost of radiating active sensors in the platform $u_k$, and $r(s_k^{(p)}, s_{k-1}^{(p)}, y_k^{(p)}, p)$ denotes the cost of using only *passive* sensors in platform $p$. Based on the observed threat history $Y_k = (y_1^{(u_0)}, \ldots, y_k^{(u_{k-1})})$, and the history of decisions $U_{k-1} = (u_0, \ldots, u_{k-1})$, the sensor manager needs to decide which sensor platform to activate at time $k$. The sensor manager decides which platform to activate at time $k$ based on the stationary policy $\mu : (Y_k, U_{k-1}) \rightarrow u_k$. Here $\mu$ is a function that maps the history $Y_k$ and past decisions $U_{k-1}$ to the choice of which platform $u_k$ is to radiate active sensors at time $k$. Let $\mathcal{U}$ denote the class of admissible stationary policies, i.e., $\mathcal{U} = \{\mu : u_k = \mu(Y_k, U_{k-1})\}$. The total expected discounted reward over an infinite time horizon is given by

$$J_\mu = \mathbf{E}\left[\sum_{k=0}^{\infty} \beta^k C_k\right] \tag{3}$$

where $\beta \in (0,1)$ denotes the discount factor, $C_k$ is defined in (2) and $\mathbf{E}$ denotes mathematical expectation. *The aim of the sensor manager is to determine the optimal stationary policy $\mu^* \in \mathcal{U}$ which minimizes the cost in (3).*

It is well known, [4, pp.31] that by defining $c(i,p) = \sum_{j=1}^{\mathcal{N}_p} \sum_{m=1}^{\mathcal{M}_p} c(i,j,m.p) a_{ij}^{(p)} b_{ijm}^{(p)}$, etc. we use the equivalent cost $C_k = c(s_k^{(u_k)}, u_k) + \sum_{p \neq u_k} r(s_k^{(p)}, p)$ in (3) since this has the same expectation as $C_k$ in (2). Therefore, since the ELIs $s_k^{(p)}$ of the passive platforms $p \neq u_k$ remain constant, their cost $r(s_k^{(p)}, p)$ is also constant. Of course the cost $c(s_k^{(u_k)}, u_k)$ of the active platform evolves with time, since $s_k^{(u_k)}$ evolves with time. To minimize the overall threat to all platforms one can choose $c(s_k^{(p)}, s_{k+1}^{(p)}, y_k^{(p)}, p)$ $= r(s_k^{(p)}, s_{k+1}^{(p)}, y_k^{(p)}, p) = y_k^{(p)}$ leading to the infinite horizon cost (3) $\sum_{k=0}^{\infty} \beta^k \sum_{p=1}^{P} \mathbf{E}\{y_k^{(p)}\}$ which is the total discounted cumulative threat posed to all the $P$ platforms. Typically the cost includes the Quality of service (QoS) and sensor usage costs of the sensors in platform $p$ – usually, the QoS from radiating active sensors in a platform is much higher than using only passive sensors.

### 2.5. Information State Formulation

The above stochastic control problem (3) is an infinite horizon Partially Observed Markov Decision Process (POMDP). We convert it to a fully observed problem in terms of the *information state*, (see [5] for a textbook exposition) as follows: For each sensor platform $p$, the information state at time $k$, denote by $x_k^{(p)}$ $x_k^{(p)}(i) \stackrel{\triangle}{=} \mathbf{P}\left(s_k^{(p)} = i \mid Y_k, U_{k-1}\right)$, $i = 1, \ldots, \mathcal{N}_p$. The information state can be computed recursively by the HMM state filter [3]) as given in (5) below.

Using the smoothing property of conditional expectations, the EMCON cost (3) can be re-expressed in terms of the information state as follows:

$$J_\mu = \mathbf{E}\left[\sum_{k=0}^{\infty} \beta^k \left( c'(u_k) x_k^{(u_k)} + \sum_{p \neq u_k} r'(p) x_k^{(p)} \right)\right] \tag{4}$$

where $c(u_k)$ denotes the $\mathcal{N}_{u_k}$ dimensional reward vector $[c(s_k^{(p)} = 1, u_k), \ldots, c(s_k^{(p)} = \mathcal{N}_{u_k}, u_k)]'$, and $r(p)$ is the $\mathcal{N}_{u_k}$ dimensional reward vector $[r(s_k^{(p)} = 1, p), \ldots, c(s_k^{(p)} = \mathcal{N}_p, p)]'$. The aim of the EMCON problem is to compute the optimal policy $\arg\min_{\mu \in \mathcal{U}} J_\mu$.

In terms of the above information state formulation, the EMCON problem described above can be viewed as the following dynamic scheduling problem: Consider $P$ parallel HMM state filters, one for each sensor platform. The $p$th HMM filter computes the ELI (state) estimate (filtered density) $x_k^{(p)}$ of the $p$th platform, $p \in \{1, \ldots, P\}$. At each time instant, only one of the $P$ platforms radiates active sensors, say platform $p$. Let $y_{k+1}^{(p)}$ be its observed threat level. This is processed by the $p$th HMM state filter which updates its estimate of the sensor platform's ELI as

$$x_{k+1}^{(p)} = \frac{B^{(p)\prime}(y_{k+1}^{(p)}) \square A^{(p)\prime} x_k^{(p)}}{\mathbf{1}' B^{(p)}(y_{k+1}^{(p)}) A^{(p)\prime} x_k^{(p)}} \qquad \text{if } p = u_k \qquad (5)$$

where $\square$ denotes Hadamard product[1], and $\mathbf{1}$ is an $\mathcal{N}_p$-dimensional column unit vector. The ELI estimates of the other $P-1$ platforms that use only passive sensors remain unaffected, i.e., since $B^{(q)}(m) = I$ and $A^{(q)} = I$ if $q \neq u_k$, we have

$$x_{k+1}^{(q)} = x_k^{(q)} \qquad \text{if platform } q \text{ only uses passive sensors .}$$

## 3. MAIN IDEAS

Given the above formulation, we briefly comment on the solution procedure - see [6] for a complete exposition.

1. In general, POMDPs are known as PSPACE hard problems [7] requiring exponential memory and computation.are computationally intractable apart from examples with small state and action spaces. For realistic EMCON problems involving several tens or hundreds of sensor platforms, the POMDP has an underlying state space that is exponential in the number of platforms – which is prohibitively expensive to solve. The main point in the above formulation of the EMCON problem is that it is a POMDP with a very special structure – called an *on-going multi-armed bandit* [8]. This multi-armed bandit problem structure implies that the optimal EMCON policy can be found by a so-called *Gittins index* rule, [8]. As a result, the multi-platform EMCON problem simplifies to a finite number of single-platform optimization problems. Hence the optimal EMCON policy is *indexable* – meaning that at each time instant it is optimal to activate the sensors on the platform (or group of platforms) with the highest Gittins index. There are numerous applications of multi-armed bandit problems in the operations research and stochastic control literature, see [8] and [9].

[1]For square matrices $A$, $B$, $C$, the Hadamard product $C = A \square B$ has elements $c_{ij} = a_{ij} b_{ij}$

2. Given the multi-armed bandit POMDP formulation and the indexable nature of the optimal EMCON policy, the main issue is how to compute the Gittins index for the individual sensor platforms. While there are several algorithms available for computing the Gittins indices for fully observed Markov decision process bandit problems [5], our POMDP bandit problem is more difficult since underlying finite state Markov chain (actual threat level) is not directly observed – instead the observations (observed threat levels) are a probabilistic function of the unobserved finite state Markov chain. The main contribution of [6] is to present finite dimensional algorithms for computing the Gittins index of a POMDP bandit.

4. A key feature of the multi-armed bandit formulation is that the EMCON algorithm for selecting which platforms should radiate active sensors can be fully decentralized. Given the indexable nature of the problem, we present in [6] a scalable decentralized optimal EMCON algorithm whose computational complexity is linear in the number of platforms. A sub-optimal version of the multi-armed bandit based EMCON algorithm is presented using Lovejoy's approximation [10]. Also in [6] a two time scale controller that can deal with slowly time varying parameters is presented.

## 4. REFERENCES


[1] "Network Centric Warfare: Department of Defense Report to U.S. Congress," http://www.defenselink.mil/nii/NCW/, March 2001.

[2] Greg Gagnon, "Network-centric special operations– exploring new operational paradigms," *Air and Space Power Chronicles*, Feb. 2002, http://www.airpower.maxwell.af.mil/.

[3] M.R. James, V. Krishnamurthy, and F. LeGland, "Time discretization of continuous-time fi lters and smoothers for HMM parameter estimation," *IEEE Trans Info Theory*, vol. 42, no. 2, pp. 593–605, March 1996.

[4] A. R. Cassandra, *Exact and Approximate Algorithms for Partially Observed Markov Decision Process*, Ph.D. thesis, Brown University, 1998.

[5] D.P. Bertsekas, *Dynamic Programming and Optimal Control*, vol. 1 and 2, Athena Scientifi c, 1995.

[6] V. Krishnamurthy, "Emission management for low probability intercept sensors in network centric warfare," *IEEE Trans. Aerospace and Electronic Systems*, Jan. 2005.

[7] C.H. Papadimitriou, *Computational Complexity*, Addison Wesley, 1995.

[8] J.C.Gittins, *Multi–armed Bandit Allocation Indices*, Wiley, 1989.

[9] P. Whittle, "Multi-armed bandits and the Gittins index," *J. R. Statist. Soc. B*, vol. 42, no. 2, pp. 143–149, 1980.

[10] W.S. Lovejoy, "Computationally feasible bounds for partially observed Markov decision processes," *Operations Research*, vol. 39, no. 1, pp. 162–175, January–February 1991.