2011 American Control Conference
on O'Farrell Street, San Francisco, CA, USA
June 29 - July 01, 2011

UAV Perimeter Patrol Operations Optimization using Efficient Dynamic Programming

K. Krishnamoorthy, M. Pachter, P. Chandler, D. Casbeer and S. Darbha

Abstract— **A reduced order Dynamic Programming (DP) method that efficiently computes the optimal policy and value function for a class of controlled Markov chains is developed. We assume that the Markov chains exhibit the property that a subset of the states have a single (default) control action associated with them. Furthermore, we assume that the transition probabilities between the remaining (decision) states can be derived from the original Markov chain specification. Under these assumptions, the suggested reduced order DP method yields significant savings in computation time and also leads to faster convergence to the optimal solution. Most importantly, the reduced order DP has been shown analytically to give the exact same solution that one would obtain via performing DP on the original full state space Markov chain. The method is illustrated via a multi UAV perimeter patrol stochastic optimal control problem.**

I. Introduction

The Dynamic Programming (DP) [1] approach to solving infinite horizon Markov decision problems (MDPs) has a long and illustrious history. However the *curse of dimensionality* has rendered it impossible to use exact DP to obtain optimal solutions for large scale problems. This has motivated the development of several approximate techniques that give tractable sub-optimal solutions instead [2], [3]. In particular, for the UAV perimeter patrol problem discussed herein, state aggregation based techniques have been employed to derive sub-optimal solutions [4], [5]. When using approximate DP methods, optimality is compromised for tractability. If the problem exhibits a certain structure, one can exploit this and perhaps still obtain optimal solutions in reasonable time. This paper exploits a feature common to many Markov chains derived from continuous time models, i.e., a reasonably large sized subset of the states have only a single (default) control action associated with them (henceforth called the non-decision states). If this is the case and if one can readily obtain the transition probabilities between the remaining decision states, then the proposed reduced order DP method can be used to efficiently compute the optimal value function and policy.

Corresponding author: K. Krishnamoorthy krishnak@ucla.edu

This research was performed while the corresponding author held a National Research Council Research Associateship Award

K. Krishnamoorthy is a Visiting Scientist with the Control Design & Analysis Branch, Air Force Research Laboratory, Wright-Patterson AFB, OH 45433, USA

M. Pachter is with the Department of Electrical Engineering, Air Force Institute of Technology, Wright-Patterson AFB, OH 45433, USA

P. Chandler is with the Control Design & Analysis Branch, Air Force Research Laboratory, Wright-Patterson AFB, OH 45433, USA

D. Casbeer is with the Control Design & Analysis Branch, Air Force Research Laboratory, Wright-Patterson AFB, OH 45433, USA

S. Darbha is with the Department of Mechanical Engineering, Texas A&M University, College Station, TX 77843, USA

II. UAV Perimeter Alert Patrol

The perimeter patrol problem arose out of the Cooperative Operations in Urban Terrain (COUNTER) project at the Air Force Research Laboratory (AFRL) [6]. In this scenario, there is a closed perimeter which must be monitored by a team of UAVs (we will consider only 2 here). Along the perimeter are m Unattended Ground Sensors (UGSs) and for the sake of simplicity, incursions into the perimeter can only occur in the vicinity of the UGS. The UGS flags an alert when there is an incursion. A UAV then investigates the alert by flying to the alert site, loitering there (we refer to this time as the dwell time) and taking video of the vicinity of the site. This video is transmitted to a remotely located operator. The operator, upon receiving and examining the transmitted video, will make the call as to whether the alert is a nuisance or a real threat. For the transmitted video to be relevant, the UAVs must service an alert within a certain response time. So it is imperative to develop control policies that minimize the response time. Previous efforts were focused on computing optimal policies for a single UAV perimeter patrol problem [7]. In this paper, we solve a multi-UAV perimeter patrol problem via the proposed reduced order DP method since traditional DP methods are rendered intractable.

A. Model Description

The patrolled perimeter is a simple closed curve with $N(\geq m)$ nodes which are uniformly distributed, of which m correspond to the UGS/alert stations. At time instant k, $x_j(k)$ is the position of the j^{th} UAV on the perimeter ($x_j \in \{0, \ldots, N-1\}$), $d_j(k)$ is the dwell time (number of loiters completed if at an alert site). $A_i(k)$ is a binary variable indicating the status of the alert at the i^{th} station and $\tilde{Y}_i(k)$ is another binary, but random variable indicating the arrival of an alert at the i^{th} station. We assume that the statistics associated with the random variable $\tilde{Y}_i(k)$ are known and that $\tilde{Y}_i, i = 1, \ldots, m$ are independent. We model the arrival of alerts as follows: Each station has an independent Poisson arrival stream of alerts at a rate of α alerts per unit time. Once a station has an alert waiting, no new alerts can arrive there until the current one is serviced. Hence, there are 2^m possibilities for the configuration of the vector of alerts $y(k) = [\tilde{Y}_1(k) \ \cdots \tilde{Y}_m(k)]$ ranging from the binary equivalent of 0 to $2^m - 1$. The control decision for the j^{th} UAV is indicated by the binary variable u_j. If $u_j = 0$, then the UAV moves to the next node and if $u_j = 0$, the UAV

dwells at the current alert station. We assume that a UAV moves by one unit every time step if $u_j = 0$. Also, we assume that the time to complete one loiter is equal to the time step. Let the m distinct station locations be indicated by the node numbers $X_1, \ldots, X_m$ (where $X_i \in \{0, \ldots, N-1\}$). One may write the discrete-time state update equations for the j^{th} UAV as follows:

$$\begin{aligned} x_j(k+1) &= (x_j(k) + 1 - u_j(k)) \bmod N \\ d_j(k+1) &= (d_j(k) + 1)u_j(k). \end{aligned} \quad (1)$$

The alert status flag at station i, $(i = 1, \ldots, m)$ is updated according to,

$$\begin{aligned} A_i(k+1) &= \prod_{j=1}^{2}(1 - \delta(x_j(k) - X_i)u_j(k)) \\ &\quad \max\{A_i(k), \tilde{Y}_i(k)\}, \end{aligned} \quad (2)$$

i.e., any alert at station i is cleared when a UAV decides to loiter there. In the above equation, δ denotes the Kronecker delta function. Also we have the constraints,

$$u_j(k) \leq \sum_{i=1}^{m} \delta(x_j(k) - X_i), \quad j = 1, 2, \quad (3)$$

i.e., UAVs can only loiter at alert stations and

$$\begin{aligned} d_j(k) &\leq D, \\ d_j(k) &= D \Rightarrow u_j(k) = 0. \end{aligned} \quad (4)$$

This constraint imposes a maximum (to keep the state space finite) on the number of allowed loiter orbits. If at any time k, $d_j(k) = D$, then UAV j is forced to leave the station it is loitering at. Now we (arbitrarily) order the states of the system and set up a one-to-one correspondence with the set $\mathcal{S} = \{1, 2, \ldots, |\mathcal{S}|\}$. If we denote $s \in \mathcal{S}$ to denote the s^{th} state of the system, we may then express the state evolution equations (1), (2), (3) and (4) compactly as,

$$s(k+1) = f(s(k), u(k), y(k)), \quad (5)$$

where,

$$\begin{aligned} u(k) &= [u_1(k)\ u_2(k)] \quad \text{and} \\ y(k) &= [\tilde{Y}_1(k)\ \tilde{Y}_2(k)\ \cdots \tilde{Y}_m(k)]. \end{aligned} \quad (6)$$

We denote the 2^m possible values (from the m digit binary representation of 0 to $2^m - 1$) that $y(k)$ can take by the row vector $\tilde{y}_j \in \mathbb{R}^m, j = 1, \ldots, 2^m$. Given that the alert arrival process is Poisson with parameter α, the probability that there is no alert in a unit time interval, $p = e^{-\alpha}$ and hence, the probability that $y(k)$ takes any one of 2^m possible values is given by,

$$p_j := Prob\{y(k) = \tilde{y}_j\} = p^{(m-n_j)}(1-p)^{n_j}, \quad (7)$$

for $j \in \{1, \ldots, 2^m\}$, where $n_j = \sum_{i=1}^{m} \tilde{y}_j(i)$ denotes the number of stations with alerts for the alert arrival configuration indicated by $\tilde{y}_j$. We model the reward (stage cost) $R^u(s)$ to be a function of the current dwell state d, alert status A and control action u, i.e.,

$$R^u(s) = \begin{cases} \sum_{j=1}^{2} \left[\mathcal{I}(d_j + 1) - \mathcal{I}(d_j)\right] u_j \\ -\beta \sum_{i=1}^{m} A_i, x_1 \neq x_2, \\ \left[\mathcal{I}(d_{\bar{j}} + 1) - \mathcal{I}(d_{\bar{j}})\right] u_{\bar{j}} - \beta \sum_{i=1}^{m} A_i, \\ x_1 = x_2, \bar{j} = \arg\max_j d_j, \end{cases} \quad (8)$$

where $\mathcal{I}$ is the information gain function (see Fig. 1) based on an operator error model [7]. The parameter $\beta > 0$, is a constant weighing the incremental information gained upon loitering once more against the number of stations with active alerts. Note that if both UAVs were to loiter at the same location, we only reward the UAV that got there first (i.e., the one with $\max_j d_j(k)$). This is to prevent multiple UAVs from servicing the same alert site (based on the notion that collecting multiple streams of video from the same site is redundant). With the reward structure defined as above, one can pose the perimeter alert patrol problem as a Markov Decision Problem (*MDP*) with the goal of maximizing the expected infinite horizon discounted reward at every state. However, the large number of *MDP* states,

$$|\mathcal{S}| = \sum_{i=0}^{m} \binom{m}{i} (N + (m-i)D)^q, \quad (9)$$

for the q UAV scenario, makes the problem intractable. Upon inspecting the expression for the number of states (9), one immediately ponders: is there a way to eliminate all those states that do not have a decision associated with them? Clearly this would significantly reduce the size of the state space thereby reducing the computational burden. From the problem description (for the 2 UAV scenario), this is equivalent to eliminating all the states wherein both UAVs are at nodes/locations that are not alert stations. Even if one UAV were to be at a station, a decision whether or not it should loiter or move on has to be made. If this elimination were possible, then we would end up with only the decision

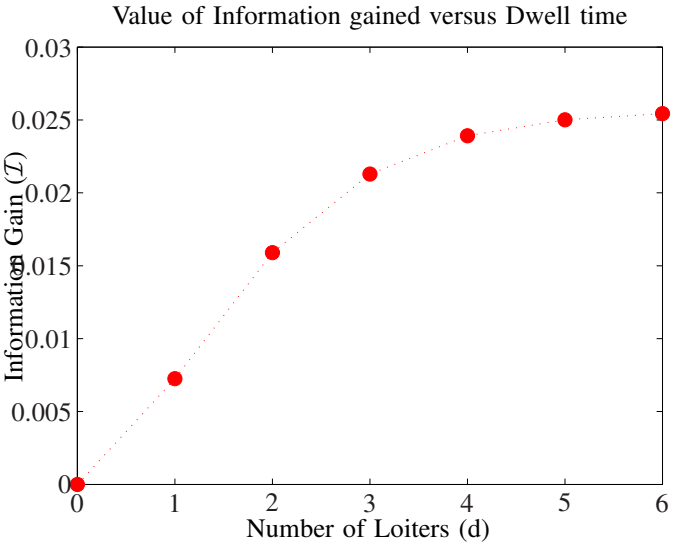

Fig. 1. Value of Information gained vs Dwell Time

states (states where a decision has to be made for at least one UAV). The number of such decision states,

$$|\mathcal{D}| = \sum_{i=0}^{m} \binom{m}{i} \left\{ (N + (m-i)D)^q - (N-m)^q \right\}. \quad (10)$$

Notice that for $q = 2$, $|\mathcal{D}|$ is linear in N compared to the original problem size $|\mathcal{S}|$, which is quadratic in N. So, given the advantage in eliminating the non-decision states, the next logical question is: how does one solve for the optimal value function without involving the value function entries associated with the non-decision states. To answer this question, we first set up a stochastic optimal control problem (for a generic *MDP*) that has a natural partitioning of the state space into decision and non-decision states and derive a reduced order DP method for the same. We then illustrate a particular instance of the reduced order DP method that solves the perimeter alert patrol optimization problem.

III. Discrete time Markov Decision Problem

Consider a discrete-time Markov decision process (MDP) with a finite state space $\mathcal{S} = \{1, 2, \ldots, |\mathcal{S}|\}$. At each state $s \in \mathcal{S}$, there is a finite set $\mathcal{A}_s$ of admissible actions. If the current state is s and an action u is taken, the agent gains a reward of $R^u(s)$ and the system transitions to a state $\bar{s}$ with probability $P^u(s, \bar{s})$ where the vector $R^u \in \mathbb{R}^{|\mathcal{S}|}$ and matrix $P^u \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$. A (deterministic) stationary policy is a mapping π that assigns an action u to each state s. We are interested in a policy that maximizes the value (or cost-to-go) function $V^\pi \in \mathbb{R}^{|\mathcal{S}|}$ i.e., the expected infinite horizon discounted cost associated with each state $\bar{s} \in \mathcal{S}$,

$$V^\pi(\bar{s}) = \mathbf{E}_\pi \left\{ \sum_{k=0}^{\infty} \lambda^k R^\pi(s(k)) | s(0) = \bar{s} \right\}, \quad (11)$$

where, with some abuse of notation, $R^\pi(i) = R^{\pi(i)}(i)$. In the above equation, k indicates time and the temporal discount factor $\lambda \in (0, 1)$. To obtain the optimal policy that maximizes the value function, we use Bellman's equation [1]: $\forall s \in \mathcal{S}$, solve for the optimal value function $V^* \in \mathbb{R}^{|\mathcal{S}|}$ given by,

$$V^*(s) = \max_{u \in \mathcal{A}_s} \left\{ R^u(s) + \lambda \sum_{\bar{s} \in \mathcal{S}} P^u(s, \bar{s}) V^*(\bar{s}) \right\}. \quad (12)$$

The optimal policy π^* is given by,

$$\pi^*(s) = \arg\max_{u \in \mathcal{A}_s} \left\{ R^u(s) + \lambda \sum_{\bar{s} \in \mathcal{S}} P^u(s, \bar{s}) V^*(\bar{s}) \right\}. \quad (13)$$

A standard DP method for solving (12), *value iteration* [8], generates a sequence V_l converging to V^* according to $V_{l+1} = T(V_l)$, where T is the DP operator, defined by

$$(TV)(s) = \max_{u \in \mathcal{A}_s} \left\{ R_u(s) + \lambda \sum_{\bar{s} \in \mathcal{S}} P^u(s, \bar{s}) V(\bar{s}) \right\}, \quad (14)$$

for all $s \in \mathcal{S}$. This sequence converges to V^* for any initialization, $V_0 \in \mathbb{R}^{|\mathcal{S}|}$ (for proof see Sec 7.2 of Bertsekas [9]). However evaluating (14) has complexity $\mathcal{O}(|\mathcal{S}|^2)$ per step. This makes the problem intractable when the size of the state space is unmanageably large.

A. State-space Partitioning into Decision and Non-Decision States

We assume that the state space can be partitioned as follows: $\mathcal{S} = \{1, 2, \ldots, |\mathcal{D}|, |\mathcal{D}| + 1, \ldots, |\mathcal{S}|\}$ where $\mathcal{D} \subset \mathcal{S}$ is the subset of states that have a decision associated with them. By definition, for all non-decision states, the control action is fixed i.e.,

$$\mathcal{A}_s = \{\phi\}, \quad \forall s \in \{|\mathcal{D}| + 1, \ldots, |\mathcal{S}|\}. \quad (15)$$

In particular, for the perimeter patrol problem, ϕ would be the default action: *both UAVs move on* $\rightarrow u_j = 0, \quad \forall j$. Now for any policy π, we have from (11)

$$V^\pi = R^\pi + \lambda P^\pi V, \quad (16)$$

where again with some abuse of notation, $P^\pi(i, j) = P^{\pi(i)}(i, j)$. We partition the vectors and transition probability matrix above into decision and (intermediate) non-decision states consistent with the partitioning of the state space:

$$V^\pi = \begin{bmatrix} V_d^\pi \\ V_n^\pi \end{bmatrix}, R^\pi = \begin{bmatrix} R_d^\pi \\ R_n \end{bmatrix}, P^\pi = \begin{bmatrix} P_{dd}^\pi & P_{dn}^\pi \\ P_{nd} & P_{nn} \end{bmatrix},$$

where the subscript d denotes decision states and n denotes non-decision (intermediate) states. Notice that by definition, R_n, P_{nd} and P_{nn} are independent of control action u, since the default action ϕ is always chosen when the system is in any of the non-decision states. Then we can write,

$$\begin{aligned} V_n^\pi &= R_n + \lambda \left(P_{nd} V_d^\pi + P_{nn} V_n^\pi \right), \\ \Rightarrow V_n^\pi &= \left(I - \lambda P_{nn} \right)^{-1} \left(R_n + \lambda P_{nd} V_d^\pi \right), \\ V_d^\pi &= R_d^\pi + \lambda \left(P_{dd}^\pi V_d^\pi + P_{dn}^\pi V_n^\pi \right). \end{aligned} \quad (17)$$

Now substituting for V_n^π in (17) we get,

$$\begin{aligned} V_d^\pi &= \left(R_d^\pi + \lambda P_{dn}^\pi \left(I - \lambda P_{nn} \right)^{-1} R_n \right) \\ &+ \lambda \left(P_{dd}^\pi + \lambda P_{dn}^\pi \left(I - \lambda P_{nn} \right)^{-1} P_{nd} \right) V_d^\pi. \end{aligned} \quad (18)$$

Since $\lambda < 1$ and P_{nn} is a right stochastic matrix (all rows consist of nonnegative numbers with each row summing to 1), one can write $\left(I - \lambda P_{nn} \right)^{-1} = \sum_{j=0}^{\infty} \lambda^j P_{nn}^j$. Also since the system cannot indefinitely remain in the set of non-decision states, there exists a $\mathcal{K} < \infty$ such that $P_{nn}^{\mathcal{K}+1}$ is the zero matrix. Now for a particular decision state $s \in \mathcal{D}$, we can expand the matrices in (18) and write,

$$\begin{aligned} V_d^\pi(s) &= R_d^\pi(s) + \sum_{j=0}^{\mathcal{K}} \lambda^{j+1} \sum_{i=1}^{|\mathcal{S}\setminus\mathcal{D}|} P_{dn}^\pi(s, i) \\ &\quad \sum_{l=1}^{|\mathcal{S}\setminus\mathcal{D}|} P_{nn}^j(i, l) R_n(l) \\ &+ \lambda \sum_{\bar{s} \in \mathcal{D}} \left(P_{dd}^\pi(s, \bar{s}) + \sum_{j=0}^{\mathcal{K}} \lambda^{j+1} \sum_{i=1}^{|\mathcal{S}\setminus\mathcal{D}|} P_{dn}^\pi(s, i) \right. \\ &\quad \left. \sum_{l=1}^{|\mathcal{S}\setminus\mathcal{D}|} P_{nn}^j(i, l) P_{nd}(l, \bar{s}) \right) V_d^\pi(\bar{s}). \end{aligned} \quad (19)$$

Now upon selection of action $u = \pi(s)$, let the next possible decision state $\bar{s}$ the system transitions to from s be $\mathcal{T}_{s,u}$ time steps away. We note that although the system is stochastic, $\mathcal{T}_{s,u}$ is a deterministic quantity that depends only on the current state, s and the action taken, u. If the system transitions immediately to another decision state, then $\mathcal{T}_{s,u} = 1$. Else, if the system transitions to a non-decision state from s upon taking action u, $\mathcal{T}_{s,u} \geq 2$. One can rewrite (19) in compact form as,

$$V_d^{\pi}(s) = \bar{R}^{\pi}(s) + \lambda^{\mathcal{T}_{s,u}} \sum_{\bar{s} \in \mathcal{D}} \bar{P}^{\pi}(s, \bar{s}) V_d^{\pi}(\bar{s}), \tag{20}$$

where we define the reduced order reward,

$$\bar{R}^u(s) = \begin{cases} R_d^u(s), \mathcal{T}_{s,u} = 1, \\ R_d^u(s) + \sum_{j=0}^{\mathcal{T}_{s,u}-2} \lambda^{j+1} \sum_{i=1}^{|\mathcal{S}\setminus\mathcal{D}|} P_{dn}^u(s,i) \\ \sum_{l=1}^{|\mathcal{S}\setminus\mathcal{D}|} P_{nn}^j(i,l) R_n(l) \ , \mathcal{T}_{s,u} \geq 2, \end{cases} \tag{21}$$

and the reduced order transition probability,

$$\bar{P}^u(s, \bar{s}) = \begin{cases} P_{dd}^u(s, \bar{s}), \mathcal{T}_{s,u} = 1, \\ \sum_{i=1}^{|\mathcal{S}\setminus\mathcal{D}|} P_{dn}^u(s,i) \\ \sum_{l=1}^{|\mathcal{S}\setminus\mathcal{D}|} P_{nn}^{\mathcal{T}_{s,u}-2}(i,l) P_{nd}(l, \bar{s}) \ , \mathcal{T}_{s,u} \geq 2. \end{cases} \tag{22}$$

This crucial step can be easily understood via the schematic

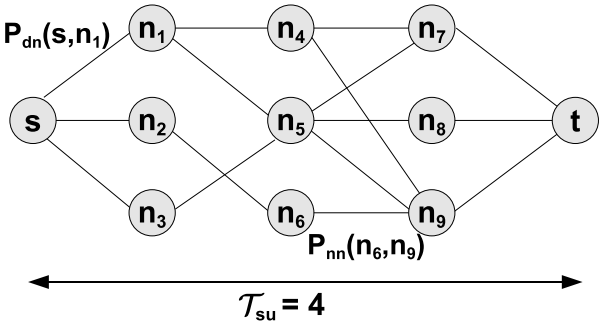

Fig. 2. Schematic of State Transition from one Decision State to another

shown in fig. 2 where upon taking action u, the system transitions from decision state s to decision state t after $\mathcal{T}_{s,u} = 4$ steps. In this instance, $P_{dd}^u(s, \bar{s}) = 0, \forall \bar{s} \in \mathcal{D}$ since the system does not transition immediately to another decision state from s and also,

$$\sum_{i=1}^{|\mathcal{S}\setminus\mathcal{D}|} P_{dn}^{\pi}(s,i) \sum_{l=1}^{|\mathcal{S}\setminus\mathcal{D}|} P_{nn}^j(i,l) P_{nd}(l, \bar{s}) = 0,$$

for $j = 0, \ldots, \mathcal{K}; j \neq \mathcal{T}_{s,u} - 2$. So the only non-zero entry is the one corresponding to $j = \mathcal{T}_{s,u} - 2$ that appears in (22). It is intuitive to think of $\bar{R}(u)$ as defined in (21) to be the expected discounted reward gained by traversing through $\mathcal{T}_{s,u} - 1$ intermediate steps before the system reaches the next decision state.

Now that we have a reduced order relation for computing the value function associated with any policy π, one can immediately write the DP equation for the optimal value function: solve for all $s \in \mathcal{D}$,

$$V_d^*(s) = \max_{u \in \mathcal{A}_s} \left\{ \bar{R}^u(s) + \lambda^{\mathcal{T}_{s,u}} \sum_{\bar{s} \in \mathcal{D}} \bar{P}^u(s, \bar{s}) V_d^*(\bar{s}) \right\}. \tag{23}$$

Again, value iteration can be used to generate a sequence V_l converging to V_d^* according to $V_{l+1} = \bar{T}(V_l)$, where $\bar{T}$ is the DP operator, defined by

$$(\bar{T}V)(s) = \max_{u \in \mathcal{A}_s} \left\{ \bar{R}^u(s) + \lambda^{\mathcal{T}_{s,u}} \sum_{\bar{s} \in \mathcal{D}} \bar{P}^u(s, \bar{s}) V(\bar{s}) \right\}, \tag{24}$$

for all $s \in \mathcal{D}$. This sequence converges to V_d^* for any initialization, $V_0 \in \mathbb{R}^{|\mathcal{D}|}$. Evaluating (24) has complexity $\mathcal{O}(|\mathcal{D}|^2)$ per step. Notice the differences between the reduced (23) and the original DP equation (12). First, the size of the vector one needs to solve for is less i.e., $|\mathcal{D}| < |\mathcal{S}|$. Second, one would expect faster convergence for the values which correspond to decision states with high $\mathcal{T}_{s,u}$ since $\lambda \in (0,1)$ and the value of future states are discounted by $\lambda^{\mathcal{T}_{s,u}}$ as opposed to just λ in the original equation. As expected, when there are no non-decision states,

$$\mathcal{D} = \mathcal{S} \Rightarrow \mathcal{T}_{s,u} = 1, \quad \forall u \in \mathcal{A}_s, \quad \forall s \in \mathcal{S}$$

the reduced (23) and the original (12) DP equations are identical and hence there is no computational savings.

Thus far, we have not restricted ourselves to the perimeter alert patrol problem. Hence, for any generic *MDP* with the stated assumptions, one can use the reduced order DP (23) to solve for the optimal value function and policy. But this requires time consuming and probably memory intensive matrix manipulations involved in constructing $\bar{R}^u$ and $\bar{P}^u$ especially when the state space is large. Hence the reduced order method is beneficial only if there is an efficient way of computing $\bar{R}^u$ and $\bar{P}^u$ for different control actions. We shall now illustrate the reduced order DP method via the patrol problem since we do have a simple way of computing the transition probabilities between the decision states therein.

B. Reduced order DP applied to Perimeter Patrol Problem

As established earlier, the original number of states $\mathcal{S}$ (9) and the reduced number of states $\mathcal{D}$ (10) for the alert patrol problem are quadratic and linear resp. in the number of locations N around the perimeter. Given that the computational complexity per *value iteration* step is of $\mathcal{O}(|\mathcal{S}|^2)$ and $\mathcal{O}(|\mathcal{D}|^2)$ resp. for the original and reduced order DP methods, we achieve significant savings in terms of CPU time, especially for large N. Before we establish the reduced order DP method for the perimeter patrol problem, we first set up some preliminary identities. The transition probability between different states is given by,

$$P^u(s, \bar{s}) = \begin{cases} 0, \text{ if } \bar{s} \neq f(s,u,\tilde{y}_j) \text{ for any } j, \\ \sum_{j \in \mathcal{C}} p_j, \text{ where } \mathcal{C} = \{j | \bar{s} = f(s,u,\tilde{y}_j)\}, \end{cases} \tag{25}$$

from the system evolution equation (5) and probabilities (7) established earlier. Hence the DP equation (12) for the

perimeter patrol problem can be written as: solve for all $s \in \mathcal{S}$,

$$V^*(s) = \max_{u \in \mathcal{A}_s} \left\{ R^u(s) + \lambda \sum_{j=1}^{2^m} p_j V^*(f(s,u,\tilde{y}_j)) \right\}, \quad (26)$$

with the reward $R^u(s)$ defined earlier (8). We note that for the patrol problem, the only stochastic component of the state is the alert status A_i at each alert station i. Hence the transition probability between decision states s and $\bar{s}$ is purely determined by the alert status component of these states, the control actions chosen and the time steps that elapsed going from state s to $\bar{s}$, $\mathcal{T}_{s,u}$. Computing $\mathcal{T}_{s,u}$ for the alert patrol problem is straightforward. For any given state s, we look at the location component of the state $(x_j(s); j = 1, 2)$. For each UAV j, under control action u_j, let the distance to the closest alert station location be denoted by $\zeta(x_j(s), u_j)$. Then we have $\mathcal{T}_{s,u} = \min_j \zeta(x_j(s), u_j)$. Now only when both UAVs decide to move on $(u_j = 0; j = 1, 2)$, the system transitions to a non-decision state and $\mathcal{T}_{s,u} \geq 2$. For all other control choices u (where at least one UAV decides to loiter), $\mathcal{T}_{s,u} = 1$ since the system transitions immediately to another decision state. Let us define $q_s = \sum_{i=1}^{m} A_i(s)$ to be the number of stations with an active alert flag status in state s and $\mathcal{P}\{r, i|q\}$ to be the probability that exactly i additional stations are active after r time steps given that q stations are active at the current time. At the non-decision states, the reward defined earlier (8), simplifies to

$$R_n(s) = R^{\phi}(s + |\mathcal{D}|) = -\beta q_{s+|\mathcal{D}|}, \quad s \in \{1, \ldots, |\mathcal{S}\backslash\mathcal{D}|\}, \quad (27)$$

because the UAVs are not allowed to loiter at these states and hence there is no information gain.

Having set up the preliminaries, we can now derive compact expressions for the reduced order reward vector (21) and transition probability matrix (22) as follows: First, we note that the expected discounted reward obtained for traversing through $\mathcal{T}_{s,u}-1$ intermediate (non-decision) stages is given by,

$$\sum_{j=0}^{\mathcal{T}_{s,u}-2} \lambda^{j+1} \sum_{i=1}^{|\mathcal{S}\backslash\mathcal{D}|} P^{\phi}_{dn}(s,i) \sum_{l=1}^{|\mathcal{S}\backslash\mathcal{D}|} P^{j}_{nn}(i,l) R_n(l) = -\beta \sum_{j=1}^{\mathcal{T}_{s,u}-1} \lambda^j \sum_{z=0}^{m-q_s} (q_s + z) \mathcal{P}\{j, z|q_s\}. \quad (28)$$

This is so because we start with q_s active stations and end up with $q_{\bar{s}} \geq q_s$ after $\mathcal{T}_{s,u}$ steps. In between, the number of active stations can either stay the same or go up. It can never go down in a (intermediate) non-decision state because a UAV can reset an active alert at a station, thereby reducing the number of active stations by one, only by loitering there. But this constitutes a decision state! In conjunction with (27), we see that the *LHS* of (28) is the expected (stage-discounted) reward i.e., the expected number of active stations multiplied by the weighing factor β.

Second, we have the probability that the system transitions from decision state s to decision state $\bar{s}$ in $\mathcal{T}_{s,u} \geq 2$ steps given by,

$$\sum_{i=1}^{|\mathcal{S}\backslash\mathcal{D}|} P^{\phi}_{dn}(s,i) \sum_{l=1}^{|\mathcal{S}\backslash\mathcal{D}|} P^{\mathcal{T}_{s,u}-2}_{nn}(i,l) P_{nd}(l,\bar{s}) = \frac{1}{\binom{m-q_s}{q_{\bar{s}}-q_s}} \mathcal{P}\{\mathcal{T}_{s,u}, q_{\bar{s}} - q_s | q_s\}. \quad (29)$$

We interpret the *LHS* of (29) as the probability of first going from decision state s to some non-decision state i, followed by $\mathcal{T}_{s,u} - 2$ transitions in the set of non-decision states $\mathcal{S}\backslash\mathcal{D}$ and finally going from non-decision state l to the decision state $\bar{s}$ (see fig. 2 for clarity). This relation can be clarified via a simple example: in state s, say we have stations 1 and 3 (where the m stations are ordered arbitrarily) currently active and after $\mathcal{T}_{s,u}$ time steps, station 2 also becomes active. Now, $\mathcal{P}\{\mathcal{T}_{s,u}, 1|2\}$ denotes the probability that any one of the remaining $m - 2$ inactive stations became active after $\mathcal{T}_{s,u}$ steps. Hence for the particular instance that station 2 becomes active, we have the probability given by $\frac{1}{\binom{m-2}{1}} \mathcal{P}\{\mathcal{T}_{s,u}, 1|2\}$.

Finally, using (28) and (29), we can write the reduced order reward vector,

$$\bar{R}^u(s) = \begin{cases} R^u(s), \mathcal{T}_{s,u} = 1, \\ R^u(s) - \beta \sum_{j=1}^{\mathcal{T}_{s,u}-1} \lambda^j \\ \sum_{z=0}^{m-q_s} (q_s + z) \mathcal{P}\{j, z|q_s\} \end{cases}, \mathcal{T}_{s,u} \geq 2, \quad (30)$$

and the reduced order transition probability,

$$\bar{P}^u(s,\bar{s}) = \begin{cases} P^u(s,\bar{s}), \mathcal{T}_{s,u} = 1, \\ \frac{1}{\binom{m-q(s)}{q_{\bar{s}}-q_s}} \mathcal{P}\{\mathcal{T}_{s,u}, q_{\bar{s}} - q_s | q_s\}, \mathcal{T}_{s,u} \geq 2, \end{cases} \quad (31)$$

for decision states s and $\bar{s}$. With the definitions (30) and (31), we write the reduced order DP equation for the perimeter patrol problem: solve for all $s \in \mathcal{D}$,

$$V^*_d(s) = \max_{u \in \mathcal{A}} \left\{ \bar{R}^u(s) + \lambda^{\mathcal{T}_{s,u}} \sum_{\bar{s} \in \mathcal{D}} \bar{P}^u(s,\bar{s}) V^*_d(\bar{s}) \right\}. \quad (32)$$

The only missing link is an expression to compute the probability $\mathcal{P}\{r, i|q\}$ for arbitrary number of stages r, that appears in both (30) and (31). For this, we first establish $\mathcal{P}\{1, i|q\}$ and then derive a recursive relationship for higher number of intermediate stages, $r \geq 2$.

C. Transition Probability between Decision States

As defined earlier, $\mathcal{P}\{r, i|q\}$ is the probability that exactly i additional stations are active after r time steps given that q stations are active at the current time. Now the probability that exactly i of the inactive stations flag an alert after one time step is given by,

$$\mathcal{P}\{1, i|q\} = \binom{m-q}{i} (1-p)^i p^{m-q-i}. \quad (33)$$

This follows from the definition of p made earlier (7). So we have the identity,

$$\sum_{i=0}^{m-q} \mathcal{P}\{1, i|q\} = (1 - p + p)^{m-q} = 1. \quad (34)$$

The above relation can be extended to any number of time steps $r \geq 2$. First we note that for $q+i$ stations to flag an alert after $r \geq 2$ steps, $i-j$ stations will have flagged an alert after $r-1$ steps and an additional j stations will have flagged an alert in the r^{th} time interval. So we have for $r \geq 2$,

$$\begin{aligned}
\mathcal{P}\{r,i|q\} &= \sum_{j=0}^{i} \mathcal{P}\{1,j|q+i-j\}\mathcal{P}\{r-1,i-j|q\} \\
&= \sum_{j=0}^{i} \binom{m-q-i+j}{j}(1-p)^j p^{m-q-i} \\
&\quad \mathcal{P}\{r-1,i-j|q\} \qquad (35) \\
&= p^{m-q-i}\sum_{j=0}^{i} \binom{m-q-i+j}{j}(1-p)^j \\
&\quad \mathcal{P}\{r-1,i-j|q\}, \qquad (36)
\end{aligned}$$

where we have used (33) to substitute for $\mathcal{P}\{1,j|q+i-j\}$. We have the identity (see Appendix I for proof),

$$\sum_{i=0}^{m-q} \mathcal{P}\{r,i|q\} = 1. \qquad (37)$$

So one can use the recursive update equation (36), in conjunction with (33), to compute the probability $\mathcal{P}\{r,i|q\}$, for an arbitrary number of stages r. In summary, we have an efficient DP method for computing the optimal value function for the perimeter alert patrol optimization problem given by (32) with the reduced order reward vector and transition probability matrix computed via relations (30), (31) and (36).

IV. Conclusions

We have established a reduced order DP method that can be used to efficiently compute the optimal value function for a controlled Markov chain under the assumptions that a reasonably large number of the states are non-decision states and that the transition probabilities between decision states be readily computable. The proposed reduced order method has the potential to yield significant savings in computation time and also faster convergence to the optimal solution. The exact amount of savings one gets depends on the size of the set of non-decision states in the given problem. Since the reduced order DP was derived analytically from the original DP equation, it yields the optimal solution. The method has been illustrated on a multi UAV perimeter patrol optimization problem.

Appendix I

Theorem 1: The transition probabilities satisfy

$$\sum_{i=0}^{m-q} \mathcal{P}\{r,i|q\} = 1. \qquad (38)$$

Proof:

We shall prove the above by induction. First, we notice that (38) readily holds for $r=1$ by (34). Now, assume (38) holds for some $r=\bar{r}$ i.e.,

$$\sum_{i=0}^{m-q} \mathcal{P}\{\bar{r},i|q\} = 1. \qquad (39)$$

Then for $r=\bar{r}+1$ we have,

$$\begin{aligned}
\sum_{i=0}^{m-q} \mathcal{P}\{\bar{r}+1,i|q\} &= \sum_{i=0}^{m-q} p^{m-q-i}\sum_{j=0}^{i} \binom{m-q-i+j}{j} \\
&\quad (1-p)^j \mathcal{P}\{\bar{r},i-j|q\} \qquad (40) \\
&= \sum_{l=0}^{m-q}\sum_{j=0}^{m-q-l} p^{m-q-l-j}\binom{m-q-l}{j} \\
&\quad (1-p)^j \mathcal{P}\{\bar{r},l|q\} \qquad (41) \\
&= \sum_{l=0}^{m-q} \mathcal{P}\{\bar{r},l|q\} \sum_{j=0}^{m-q-l} \binom{m-q-l}{j} \\
&\quad p^{m-q-l-j}(1-p)^j \qquad (42) \\
&= \sum_{l=0}^{m-q} \mathcal{P}\{\bar{r},l|q\}(p+1-p)^{m-q-l} \\
&= \sum_{l=0}^{m-q} \mathcal{P}\{\bar{r},l|q\} = 1. \qquad (43)
\end{aligned}$$

The first step (40) follows from the update equation (36) and we have done a change of variable, $l=i-j$ and used the summing property,

$$\sum_{i=0}^{m-q}\sum_{j=0}^{i} g(i,j) = \sum_{l=0}^{m-q}\sum_{j=0}^{m-q-l} g(j+l,j),$$

to arrive at (41).

■

References

[1] R. E. Bellman, *Dynamic Programming*. Princeton, NJ: Princeton University Press, 1957.

[2] G. Gordon, "Approximate solutions to Markov decision processes," Ph.D. dissertation, Carnegie Mellon University, Pittsburg, PA, 1999.

[3] W. Powell, *Approximate Dynamic Programming - Solving the Curses of Dimensionality*. John Wiley & Sons, 2007.

[4] K. Krishnamoorthy, M. Pachter, S. Darbha, and P. Chandler, "Approximate dynamic programming with state aggregation applied to UAV perimeter patrol," *Int. J. Robust and Nonlinear Control*, to appear in special issue on Cooperative Control of Autonomous Systems.

[5] S. Darbha, K. Krishnamoorthy, M. Pachter, and P. Chandler, "State aggregation based approximate linear programming approach to approximate dynamic programming," in *Proc. IEEE Conf. Decision and Control*, Atlanta, GA, Dec 2010, pp. 935–941.

[6] D. Gross, S. Rasmussen, P. Chandler, and G. Feitshans, "Cooperative Operations in UrbaN TERrain (COUNTER)," in *Defense and Security Symposium*. Orlando, FL: SPIE, Apr. 2006.

[7] P. Chandler, J. Hansen, R. Holsapple, S. Darbha, and M. Pachter, "Optimal perimeter patrol alert servicing with Poisson arrival rate," in *AIAA Guidance, Navigation and Control Conf.*, Chicago, IL, August 2009.

[8] R. Howard, *Dynamic Programming and Markov Processes*. Cambridge, MA: MIT Press, 1960.

[9] D. Bertsekas, *Dynamic Programming and Optimal Control*, 3rd ed. Belmont, MA: Athena Scientific, 2005, vol. 1.