Unmanned Vehicle Operations under Imperfect Information in an Adversarial Environment

William M. McEneaney [*] Rajdeep Singh [†]

February 5, 2004

Abstract

Command and Control (C^2) decisions must of necessity be based on imperfect knowledge of the battlespace. With the advent of unmanned air vehicles (UAVs), the pace at which decision points arise will increase. This necessitates the development of automated C^2 tools for use in the decision-making process. Estimation and control in the presence of purely random input noise is well-understood, and produces excellent results. In the context of C^2 decisions, one most consider observations contaminated by both random noise and adversarially induced "noise". Consequently, zero-sum, discrete, stochastic games under imperfect observations are considered here. Machinery has recently been developed which allows one to solve such problems. The theory is summarized. For problems in the class considered here, the resulting algorithms are computationally feasible. The method is applied on a small game testbed. The behavior of the resulting controls are discussed. An alternate (naive) approach is to apply the optimal state feedback game controls to the maximum likelihood state. This alternate approach is susceptible to deception by the opponent. It is shown that the improvements in using the robust approach range from small to tremendous depending on certain factors.

1 Introduction

Command and Control (C^2) decisions must of necessity be based on imperfect knowledge of the battlespace. With the advent of unmanned air vehicles (UAVs), the size of vehicle

[*] Depts. of Mechanical/Aerospace Engineering and Mathematics, University of California at San Diego, La Jolla, CA 92093-0112, USA, http://math.ucsd.edu/~wmcenean/, wmceneaney@ucsd.edu, Research partially supported by DARPA/AFRL contract F30602-01-C-0201 and NSF grant DMS-0307229.

[†] Dept. of Mechanical/Aerospace Engineering, University of California at San Diego, La Jolla, CA 92093-0112, USA, rsingh@mechanics.ucsd.edu, Research partially supported by DARPA/AFRL contract F30602-01-C-0201.

fleets and operation speed will undoubtedly increase. Consequently, the pace at which decision points arise will increase. This necessitates the development of automated C^2 tools for the decision-making process (and/or advisement to the decision-making process).

The availability of automated tools for decision making in the presence of only random noise is excellent, surpassing that of human operators in most cases. Notable tools are the Kalman filter, Bayesian estimators, and their many progeny. However, battlepsace observations are contaminated not only by random noise, but also by purposefully deceptive information. Deception and diversion have been the basis of many of the most significant military successes in history. In the current environment of asymmetrical warfare, deception has continued to play a tremendously important role. In particular, inadequate reasoning about deception has led to serious failures.

In stochastic estimation and control applications, one carries forward conditional probability distributions concerning the current state of the system, and applies state feedback controls at state estimates based on these distributions. Typically, the maximum likelihood state is used as the basis for the state feedback control computation. (It should be noted that the probability distributions are "information states" in that they contain sufficient information to compute the optimal control [9], [17].) However, the reduction of the distribution to the maximum likelihood state, and computation of the corresponding state feedback control is typically suboptimal in the nonlinear case. Computational limits generally preclude any other approach than reduction to the maximum likelihood state.

An alternate approach that has grown in influence in the last two decades has been the H_∞ approach. Here, a form of worst-case analysis is used where the disturbances are essentially modeled as antagonistic. The conditional probability distribution is replaced by an information state based on past costs. The state feedback control is computed at the argmax of the sum of this information state and the value function (a measure of future cost). This is only provably optimal on a thin slice of the problem space [3], [15].

In the context of C^2 decisions, one most consider both random noise and adversarially induced "noise". We will be concerned here with a specific class of discrete stochastic games under imperfect observations. The choice of this class will be affected by both application considerations and computational feasibility considerations. We will describe a reasonable class of problems whose solutions are much more tractable than one would expect. We consider only zero-sum games (see, for example, [3], [4], [8], [10] [11]). In particular, we concentrate on a simple (sub)class of discrete-time/discrete-space stochastic games where one player (the opponent) has perfect information.

This paper has two components. The first is the theory for computation of a robust controller for such a game. This theory is fully developed in [18], [19], [20], and will only be briefly described here. Although this theory is a new development, some earlier work in related directions can be found in [5], [24]. The second component is the study of an application of this theory to a simple C^2 problem for UAV operations. (For related information, see [1], [6], [7], [12], [13], [14], [16], [21], [22], [23].) The application example will indicate the importance of reasoning about deception in this context, and provide some understanding of the behaviors induced by such a tool.

The application study problem will be one where player 1 (with imperfect information) is attempting to prevent player 2 from taking certain fixed-location, ground assets. In the model problem, player 2 is moving entities (say general ground vehicles and air defenses) toward one or both of these player 1 fixed-location strategic assets. Player 2 "takes" the assets by successfully moving the entities to their locations. Player 1 has UAVs which can move to intercept the player 2 entities. Player 2 may choose to move the entities in one of two modes: 'stealthy' or 'non-stealthy'. In the former case, the entities are difficult to detect, and in the latter they are detected more easily. We consider both the case where player 1's UAVs are also more effective against non-stealthy player 2 entities, and the case where they are not. (Perhaps this second case should be clarified. In this latter case, the probability of observing a stealthy entity is lower than that of a non-stealthy entity, but the probability of destruction once player 1 attacks the entity is independent of its stealthiness.) We also allow player 2 to use inexpensive decoys, and we assign probabilities by which these decoys are mistaken for player 2 vehicles. This problem has sufficient complexity in order for diversionary tactics on the part of player 2 to be effective. Specifically, it can be effective for player 2 to make one group of entities highly visible by making them non-stealthy and accompanying them with decoys, while attempting to keep a second group of entities stealthy.

We compare two player 1 approaches in this simple game problem. The most naive is for player 1 to simply take the maximum likelihood estimate of the player 2 state, and to apply a feedback control for this system state. As one can easily imagine, this approach is open to exploitation by player 2 diversionary tactics. In that case, the maximum likelihood estimate made by player 1 can sometimes be that all or almost all of the player 2 entities are heading for the first player 1 asset. If player 1 does not react quickly enough to the potentiality of player 2 assets moving toward the other asset, then the outcome may often be far from optimal. The second player 1 approach is through the use of the imperfect information stochastic game machinery described below. Both approaches are compared using standard Monte Carlo techniques. As one would expect, there is an improvement in outcome with the approach described herein when compared with the standard maximum likelihood/certainty equivalence approach. On the other hand, there are significant computational requirements when using this new approach. Various parameters will be varied to obtain some heuristic understanding of how much improvement the new approach yields as a function of the specific game model.

Section 2 describes the general imperfect information stochastic game model. The theory/machinery is outlined in Section 3. The theory has three components: the information state definition and propagation, the value function definition and computation, and the combination of these two elements to produce a deception-robust controller. These are discussed briefly in the Subsections 3.1 – 3.3. As indicated above, more detail on the theory can be found in [18], [19], [20]. Some computational complexity issues are very briefly indicated in Section 4. Section 5 contains the application of this technology, and the discussion of the results obtained.

2 System Definition

Potential states of the system will be represented by $x \in \mathcal{X}$ where $\mathcal{X}$ is some finite set. Time will be discrete, and the state of the system at time t will be denoted by X_t. The control for player 1, the minimizing player, will take values $u \in U$ where U is finite. The corresponding controls for player 2 (maximizing) will be $w \in W$ which is also a finite set. Controls for each player at time t will be denoted as u_t and w_t.

We will consider a finite time problem with time taking values in $\{0, 1, 2, \ldots, T\}$. We will denote the terminal cost as $\mathcal{E} : \mathcal{X} \to \mathbf{R}$; the cost of terminal state X_T is $\mathcal{E}(X_T)$. There is no running cost.

We suppose that the state evolves as a controlled Markov chain. Let the probability that $X_{t+1} = j$ given $X_t = i$ with controls $u_t = u \in U$ and $w_t = w \in W$ be

$$p_{ij}(u_t, w_t) = \Pr(X_{t+1} = j | X_t = i, u_t = u, w_t = w),$$

and let the $n \times n$ matrix of the elements p_{ij} be denoted as $P(u, w)$ where $n \doteq \#\mathcal{X}$. We will assume that there is an observation process for player 1 (recall that here player 2 will know the state perfectly) which can be controlled somewhat by both players. Let the observation process be $y_\cdot$ with $y_t \in Y$, where the probability that observation $y_t = \overline{y}$ given $X_t = i$ and controls $u_t = u, w_t = w$ is denoted as

$$R_i \doteq \Pr(y_t = \overline{y} | X_t = i, u_t = u, w_t = w).$$

3 Review of the Theory

As indicated in the introduction, the theory has three components: the information state definition and propagation, the value function definition and computation, and the combination of these two elements to produce a deception-robust controller. These are discussed briefly in the following subsections. (See [18], [19], [20] for further detail on the theory.) Readers interested mainly in the application might prefer to proceed to Sections 4 and 5.

3.1 Information State Propagation

As indicated above, we suppose that both the dynamic and observation processes might be affected by player 2 controls. Suppose player 1 (with imperfect/observation-based information) had an initial probability distribution for the system state at time $t = 0$, say q_0. If player 1 knew all the player 2 inputs prior to the current time t, then player 1 could propagate the observation-conditioned probability distribution forward in time to produce q_t. However, the player 2 controls are unknown. The information state at any time, t, will be a function of current conditional probability distributions, and we denote it as $I_t(q)$. For any probability distribution, q, consistent with the observations, $I_t(q)$ will be the worst-case (from player 1 perspective) cost to player 2 for application of controls that might yield q at time t given the observations so far obtained. The formal definition of $I_t(\cdot)$ and propagation algorithm will now be discussed. Recall that the size of $\mathcal{X}$ is n. The domain of $I_t(\cdot)$ is $Q(\mathcal{X}) \subset \mathbf{R}^n$ – the simplex of probability distributions over $\mathcal{X}$. ($Q(\mathcal{X})$ is the simplex in the first octant of $\mathbf{R}^n$ defined by the unit basis vectors.)

We let the initial information state be $I_0(\cdot) = \phi(\cdot)$. Here, ϕ represents the initial cost to obtain and/or obfuscate initial state information. The case where this information cannot be affected by the players may be represented by a max-plus delta function. That is, ϕ takes the form

$$\phi(q) = \delta_{q_c}(q) = \begin{cases} 0 & \text{if } q = q_c \\ -\infty & \text{otherwise.} \end{cases} \tag{1}$$

This will be the case we will concentrate on here.

Denote a conditional probability of the state at time t be by $q_t \in Q(\mathcal{X})$. In the absence of observations, and for given controls u_t, w_t, this propagates according to $q_{t+1} = P^T(u_t, w_t)q_t$. Note that although this mapping is into $Q(\mathcal{X})$, it is not necessarily onto. Consequently, it will be necessary to keep track of the set of feasible conditional probabilities at time t, Q_t. Note that for t prior to the current time, u_t is known by player 1 while w_t is unknown. Player 2 has full state knowledge, and consequently, it's control must be allowed to depend on the actual state. We encapsulate this by letting the player 2 feedback control at time t be $\vec{w}_t$ where the i^{th} component of $\vec{w}_t$ represents the player 2 state feedback control corresponding to state i. Consequently, for $u \in U$ and any vector $\vec{w} \in W^n$, define the matrix $\widetilde{P}$ by

$$\widetilde{P}_{ij}(u_t, \vec{w}) \doteq P_{ij}(u_t, \vec{w}_i) \qquad \forall\, i, j \in \{1, 2, \ldots, n\}. \tag{2}$$

Now let $\vec{w}_{[0,t-1]} = \{\vec{w}_0, \vec{w}_1, \ldots, \vec{w}_{t-1}\}$ where each $\vec{w}_r \in W^n$ denote a sequence of state-dependent controls for player 2. Forward propagation is via

$$q_{t+1} = \widetilde{P}^T(u_t, \vec{w}_t)q_t. \tag{3}$$

We assume throughout that $\widetilde{P}^{-1}$ exists in the standard sense.

Let us now consider the observation process. We will assume that the observations may occur at each time step, t. We will distinguish between a priori conditional distributions, denoted as q_t, and a posteriori distributions, denoted as $\widehat{q}_t$. That is, $\widehat{q}_t$ incorporates the possible new information in an observation at time t. Recalling the observation discussion of Section 2, and the fact that we are allowing the player 2 control to depend on the true state, we let the vector $\widetilde{R}$ have components

$$\widetilde{R}_i \doteq \Pr(y_t = \overline{y} | X_t = i, u, \vec{w}_i)$$

for each $i \leq n$ where again $\vec{w}$ indicates the possibly state-dependent choice of player 2 control. Let $D(\widetilde{R})$ be the matrix whose i^{th} diagonal element is $\widetilde{R}_i$ for each i, and whose other elements are zero. Then, given any control u and $\vec{w}$ and any observation $\overline{y}$, the a posteriori distribution would be given by

$$\widehat{q}_t = \left(\tfrac{1}{\widetilde{R}^T(\overline{y}, u_t, \vec{w}) q_t} \right) D(\widetilde{R}(\overline{y}, u_t, \vec{w})) q_t. \tag{4}$$

We suppose that D is full rank; i.e. that $\widetilde{R}_i \neq 0$ for all i. Otherwise there is some additional technical analysis which we do not include here. Note that with a little work, one can show that the inversion of this propagation takes the form $q_{ti} = [1/(\sum_i \widetilde{R}_i^{-1} \hat{q}_{ti})] \widetilde{R}_i^{-1} \hat{q}_{ti}$.

We see that the set of feasible distributions at time t is

$$\begin{aligned} Q_t(u_{[0,t-1]}, y_{[0,t-1]}) = \{ q \in Q(\mathcal{X}) : \exists \vec{w}_{[0,t-1]} \in [W^n]^t \text{ such that } q_0 \in Q(\mathcal{X}) \text{ where } q_0 \text{ is} \\ \text{given by backward propagation (6) with } q_t = q \} \end{aligned} \tag{5}$$

where

$$\begin{aligned} q_{r-1} &= G^{-1}(q_r, u_{r-1}, \vec{w}_{r-1}, y_{r-1}) \\ &\doteq \tfrac{1}{\widehat{R}^T(y_{r-1}, u_{r-1}, \vec{w}_{r-1}) q_r} D^{-1}(\widetilde{R}(y_{r-1}, u_{r-1}, \vec{w}_{r-1})) \widetilde{P}^{-T}(u_{r-1}, \vec{w}_{r-1}) q_r \end{aligned} \tag{6}$$

where $\widehat{R}_i(y_{r-1}, u_{r-1}, \vec{w}_{r-1}) \doteq 1/[\widetilde{R}_i(y_{r-1}, u_{r-1}, \vec{w}_{r-1})]$. The information state is

$$I_t(q; u_{[0,t-1]}, y_{[0,t-1]}) \doteq \begin{cases} \sup_{q_0 \in Q_0^{q, u_{[0,t-1]}}} \sup_{\vec{w}_{[0,t-1]} \in [W^n]^t} I_0(q_0) & \text{if } q \in Q_t(u_{[0,t-1]}, y_{[0,t-1]}) \\ -\infty & \text{otherwise} \end{cases}$$

where

$$\begin{aligned} Q_0^{q, u_{[0,t-1]}} \doteq \{ & \widetilde{q} \in Q(\mathcal{X}) : \exists \vec{w}_{[0,t-1]} \in [W^n]^t \text{ such that } q_t = q \text{ given} \\ & q_0 = \widetilde{q} \text{ and propagation (6)} \}. \end{aligned}$$

Note that the information state at time t maps conditional probability distributions (conditioned on the observation process) to costs ($\in \mathbf{R} \cup \{-\infty\}$). It indicates the maximal cost (optimal from player 2 perspective) to generate conditional distribution q in a Bayesian estimator given the player 1 observations up to time t. A cost of $-\infty$ indicates that that value of q is not feasible. For the case we concentrate on here, the initial information state, $I_0 = \phi$ takes the form of a max-plus delta function (1). (The general case appears in [18], [20].) This corresponds to the situation where player 2 controls do not affect the initialization. This can be generalized by taking a sum of max-plus delta functions as the initialization. For each (known) u_0 and (unknown) $\vec{w}_0$, the dynamics and observation propagation discussed above takes q_0 into some q_1. The set of all possible q_1's which may be generated by feasible $\vec{w}_0$'s is Q_1 (as indicated mathematically above). Note that the size of Q_1 is no larger than the size of W^n. Further,

$$I_1(q) = \begin{cases} 0 & \text{if } q \in Q_1 \\ -\infty & \text{otherwise.} \end{cases}$$

This defines the propagation of the information state forward in time by one time-step for this particular class of initial information states.

Theorem 3.1 *If ϕ is a max-plus delta function, then $I_t(q) : Q(\mathcal{X}) \to \{-\infty, 0\}$ takes the form*

$$I_t(q) = \begin{cases} 0 & \textit{if } q \in Q_t \\ -\infty & \textit{otherwise} \end{cases}$$

where Q_t contains at most $(\#W^n)^t$ points. Further,

$$Q_t = \{q \in Q(\mathcal{X}) |\ q = \widetilde{P}(u_t, \vec{w})q_{t-1} \textit{ for some } q_{t-1} \in Q_{t-1} \textit{ and } \vec{w} \in W^n\}.$$

The proof is quite trivial [18].

Various methods are used to reduce the potentially exponential growth in the size of Q_t including pruning and assumptions of consistency of player 2 strategies (which may include consistently randomized strategies).

The more general theory for propagation (when $I_0 = \phi$ is not necessarily a max-plus delta function) can be found in [18]. The only additional remark we make here is that the propagation of the information state remains finite-dimensional in the cases where ϕ is either piecewise linear or a sum of max-plus delta functions.

3.2 State Feedback Value Function

We now turn to the state feedback value function. The full state of the system is now described by the true state taking values $x \in \mathcal{X}$ and the player 1 conditional probability process taking values $q \in Q(\mathcal{X})$. We denote the terminal payoff for the game as $\mathcal{E} : \mathcal{X} \to \mathbf{R}$ (where of course this does not depend on the internal conditional probability process of player 1). Thus the state feedback value function at the terminal time is

$$V_T(x, q) = \mathcal{E}(x). \tag{7}$$

One issue that arises is the information available to player 2. One option would be to assume that it knows only the actual true state, x. However, with full knowledge of the state and observations, player 2 could also construct the conditional probability, q. This second model is more conservative in terms of construction of the player 1 control, and this model will be used here.

The state of the state feedback game at time t is (X_t, q_t). Player 1 will have access only to the probability distributions up to the current time, while player 2 will have access to the true state as well.

The allowable strategies for player 1 are mappings from the space of distributions up to the current time to player 1 controls in U, and are non-anticipative. Player 2 strategies may depend on both the probability distributions up to the current time and the true states X_t up to the current time. Since player 1 knows only q_t in this nonstandard state feedback game, the value function from player 1 perspective is a function of q only, and we denote this as $V_t^1(q)$ It is also necessary to compute a related function, $V_t(x, q)$ which might be thought of as roughly corresponding to the value from the player 2 perspective since that player has access to the true state, X_t. Rigorous definitions appear in [18]. It is proven there that the value function can be obtained by the following backward dynamic programming algorithm.

We suppose that one has $V^1_{t+1}(\cdot), V_{t+1}(\cdot,\cdot)$, and indicate how one obtains $V^1_t(\cdot), V_t(\cdot,\cdot)$.

1. First, let the vector-valued function $\vec{M}_t$ be given component-wise by

$$[\vec{M}_t]_x(q,u) = \max_{\vec{w}\in W^n} \Big[\sum_{j\in\mathcal{X}} \widetilde{P}_{xj}(u,\vec{w}) V_{t+1}(j, q'(q,u,\vec{w}))\Big] \tag{8}$$

where

$$q'(q,u,\vec{w}) = \widetilde{P}^T(u,\vec{w})q \tag{9}$$

and the optimal $\vec{w}$ is

$$\vec{w}^0_t = \vec{w}^0_t(x,q,u) = \operatorname*{argmax}_{\vec{w}\in W^n} \Big\{\sum_{j\in\mathcal{X}} \widetilde{P}_{xj}(u,\vec{w}) V_{t+1}(j, q'(q,u,\vec{w}))\Big\}. \tag{10}$$

2. Then define L_t as

$$L_t(q,u) = q^T \vec{M}_t(q,u), \tag{11}$$

and note that the optimal u is

$$u^0_t = u^0_t(q) = \operatorname*{argmin}_{u\in U} L_t(q,u) = \operatorname*{argmin}_{u\in U} q^T \vec{M}_t(q,u). \tag{12}$$

3. With this, one obtains the next iterate from

$$V_t(x,q) = \sum_{j\in\mathcal{X}} \widetilde{P}_{xj}(u^0_t,\vec{w}^0_t) V_{t+1}(j, q'(q,u^0_t,\vec{w}^0_t)) = [\vec{M}_t]_x(q,u^0_t) \tag{13}$$

and the corresponding best achievable expected result from the player 1 perspective is

$$V^1_t(q) = q^T \vec{M}_t(q,u^0_t). \tag{14}$$

Consequently, for each $t \in \{0,1,\ldots,T\}$ and each $x\in\mathcal{X}$, $V_t(x,\cdot)$ is a piecewise constant function over simplex $Q(\mathcal{X})$. Due to this piecewise constant nature, propagation is relatively straight-forward (more specifically, it is finite-dimensional in contradistinction to the general case). However, this is slightly less critical than the propagation issue for the information state, since the state feedback value may be pre-computed, while the information state must be propagated in real-time.

3.3 Robustness

The remaining component of the computation of the control at each time instant is now discussed. One seeks a way to combine the information state at time t with the state feedback value function at time t, in order to produce a controller which is robust to adversarial inputs (the player 2 controls). Note that this allows one to compute a state feedback controller ahead of time, and then only propagate the information state forward in time.

To simplify notation, note that by (11) and (8)

for any u,

$$L_t(q,u) = \mathbf{E}_q\Big[\max_{\vec{w}\in W^n} \sum_{j\in\mathcal{X}} \widetilde{P}_{Xj}(u,\vec{w})V_{t+1}(j,q'(q,u,\vec{w}))\Big]$$

where the notation $q'(q,u,\vec{w})$ is defined in (9). Let us hypothesize that the optimal control for player 1 is

$$u_t^m \doteq \operatorname*{argmin}_{u\in U}\left[\max_{q\in Q(\mathcal{X})}\{I_t(q)+L_t(q,u)\}\right]. \tag{15}$$

We will assume that u^m is a strict minimizer. Note that u^m is a *strict minimizer* of a function $f(u)$ if $f(u^m) < f(u)$ for all $u \neq u^m$.

In order to prove robustness, one must first define an imperfect observation value function in terms of the worst-case expected cost (from the player 1 point of view). In order to make this section more readable, we will begin by writing down this value function, and then describe the terms within it rather than vice-versa. For technical reasons, it appears best to work with the following value function. This value at any time $\bar{t}$ is

$$Z_{\bar{t}} \doteq \sup_{q_{\bar{t}}\in Q_t} \;\inf_{\lambda_{[\bar{t},T-1]}\in\Lambda_{[\bar{t},T-1]}} \;\sup_{\theta_{[\bar{t},T-1]}\in\theta_{[\bar{t},T-1]}} \Big[I_{\bar{t}}(q_{\bar{t}}) + \mathbf{E}_{q_{\bar{t}}}\left\{\mathbf{E}[\mathcal{E}(X_T)|\, X_{\bar{t}}=X]\right\}\Big]. \tag{16}$$

The expectation uses the (player 1) assumption that the distribution of $X_{\bar{t}}$ is $q_{\bar{t}}$ for each $q_{\bar{t}} \in Q_{\bar{t}}$ and is taken not only over $X_{\bar{t}}$ but also over all observation and dynamic noise from time $\bar{t}$ to terminal time T. Note that this is not a full upper value in that the supremum over $q_{\bar{t}}$ occurs outside the infimum over player 1 controls $\lambda_{[\bar{t},T-1]}$. The strategy set for player 1 is

$$\Lambda_{[\bar{t},T-1]} = \left\{\lambda_{[\bar{t},T-1]} : \, Y^{T-\bar{t}} \to U^{T-\bar{t}}, \text{ nonanticipative in } y_{\cdot-1}\right\}$$

where "nonanticipative in $y_{\cdot-1}$" is defined as follows. A strategy, $\lambda_{[\bar{t},T-1]}$ is nonanticipative in $y_{\cdot-1}$ if given any $t \in (\bar{t}, T-1]$ and any sequences $y_\cdot, \tilde{y}_\cdot$ such that $y_r = \tilde{y}_r$ for all $r \in [\bar{t}, t-1]$, one has $\lambda_t[y] = \lambda_t[\tilde{y}]$. The strategy set for player 2 (neglecting $q_{\bar{t}}$ as a player 2 control) is naturally

$$\Theta_{[\bar{t},T-1]} = \left\{\theta_{[\bar{t},T-1]} : Y^{T-\bar{t}} \to W^{n(T-\bar{t})}, \text{ nonanticipative in } y_{\cdot-1}\right\}. \tag{17}$$

The first step in obtaining the robustness result is to show that the value, $Z_{\bar{t}}$ has the following representation [18].

Theorem 3.2

$$Z_{\bar{t}} = \sup_{q_{\bar{t}}\in Q_{\bar{t}}}\Big[I_{\bar{t}}(q_{\bar{t}}) + V_{\bar{t}}^1(q_{\bar{t}})\Big] \qquad \forall\, \bar{t}\in[0,T]. \tag{18}$$

In other words, the game value $Z_{\bar t}$ is the supremum of the sum of the information state, $I_{\bar t}$, and the optimal expected state feedback value, $V_{\bar t}^1$, from player 1's perspective. Using Theorem 3.2, one can show [18] that

$$Z_{\bar t} = \sup_{q_{\bar t} \in Q_t} \Big[I_{\bar t}(q_{\bar t}) + \min_{u \in U} L_{\bar t}(q_{\bar t}, u) \Big] = \sup_{q_{\bar t} \in Q_t} \min_{u \in U} \Big[I_{\bar t}(q_{\bar t}) + L_{\bar t}(q_{\bar t}, u) \Big]. \tag{19}$$

In order to obtain the Robustness/Certainty Equivalence result to follow, it is sufficient to make the following Saddle Point Assumption. We assume that

$$\sup_{q_{\bar t} \in Q_t} \min_{u \in U} \Big[I_{\bar t}(q_{\bar t}) + L_{\bar t}(q_{\bar t}, u) \Big] = \min_{u \in U} \sup_{q_{\bar t} \in Q_t} \Big[I_{\bar t}(q_{\bar t}) + L_{\bar t}(q_{\bar t}, u) \Big] \qquad \forall\, \bar t \in [0, T]. \tag*{$(A5.1)$}$$

With Assumption (A5.1), (19) becomes

$$Z_{\bar t} = \min_{u \in U} \sup_{q_{\bar t} \in Q_t} \Big[I_{\bar t}(q_{\bar t}) + L_{\bar t}(q_{\bar t}, u) \Big]. \tag{20}$$

Finally, after some work [18], one obtains the robustness result:

Theorem 3.3 *Let $\bar t \in \{0, T-1\}$. Let I_0, $u_{[0,\bar t-1]}$ and $y_{[0,\bar t-1]}$ be given. Let the player 1 control choice, $u_{\bar t}^m$, given by (15) be a strict minimizer. Suppose Saddle Point Assumption (A5.1) holds. Then, given any player 1 strategy, $\lambda_{[\bar t, T-1]}$ such that $\lambda_{\bar t}[y.] \neq u_{\bar t}^m$, there exists $\varepsilon > 0$, $q_{\bar t}^\varepsilon$ and $\vec w^\varepsilon_{[\bar t, T-1]}$ such that*

$$\sup_{q \in Q_{\bar t}} \{ I_{\bar t}(q) + L_{\bar t}(q, u_{\bar t}^m) \} = Z_{\bar t} \le I_{\bar t}(q_{\bar t}^\varepsilon) + \mathbf{E}_{q_{\bar t}^\varepsilon} \Big\{ \mathbf{E}[\mathcal{E}(X_T^\varepsilon) \mid X_{\bar t}^\varepsilon = X] \Big\} - \varepsilon \tag{21}$$

where X^ε denotes the process propagated with control strategies $\lambda_{[\bar t, T-1]}$ and $\vec w^\varepsilon_{[\bar t, T-1]}$.

In summary, the machinery for computation of the deception-robust control is as follows. One precomputes (piecewise constant) $V_t(x, q)$ and $L_t(q, u)$ (possibly by the dynamic programming algorithm of the previous subsection). One also propagates the information state, $I_t(q)$ in real-time. In the case considered here (I_0 being a max-plus delta function), this amounts to propagating the finite set of distributions, Q_t. One then computes the control from (15). In the case considered here, (15) reduces to

$$u_t^m = \operatorname*{argmin}_{u \in U} \left[\max_{q \in Q_t} \{ L_t(q, u) \} \right]. \tag{22}$$

4 Computational Tractability

One of the main motivations for consideration of games of this form is the claim that they can represent useful applications and, at the same time, lead to reasonably tractable algorithms. The largest problem with tractability for imperfectly observed problems is the propagation of the information state forward in real-time. Recall that the case we concentrate on here is where ϕ is a max-plus delta function. This leads to a particularly tractable problem. In this case, I_t is 0 only at a finite number of points, and is $-\infty$ elsewhere (see Theorem 3.1). Thus, the information state is easily propagated. Further, since I_t is not $-\infty$ at only a finite number of points, the computation of the maximum in (15) reduces to the maximum over a finite set as indicated in (22). The size of this finite set is the number of distributions in Q_t. Consequently, an important issue is the growth in the number of probability distributions due to unknown player 2 control options. If there are N_w possible player 2 controls at every time step, and no assumption is made about the continuity of player 2 controls over time, then the set of feasible distributions would grow as $(N_w)^t$. In general, there are some (suboptimal) techniques or approximations that can be employed to prune the growth. An example of a pruning algorithm currently in use is where pairs of distributions whose distance from one another is below a certain tolerance are merged into a single distribution. Another pruning technique which needs to be tested is where distributions for which the value is extremely low are neglected; these are distributions which may be unlikely to become active in the maximization computation. Lastly, an assumption of continuity of strategy by the opponent (or an assumption of a limited number strategy changes by the opponent) greatly reduces the computational burden, and this will be studied.

One should also note that V_t is a piecewise constant function of q. It is easy to observe that the number of pieces is proportional to the product of future player 1 and future player 2 control options. One could prevent exponential growth by restricting the player 1 and/or player 2 controls to predetermined sets of feedback policies (possibly randomized).

5 UAV Operations Example Test-bed

We now apply the above approach to a relatively simple problem. This problem is based on UAV Command and Control applications. Its simplicity will allow us to obtain insight into the potential benefits and behaviors of the imperfect information stochastic game approach. The general motivational application is described more fully elsewhere – such as in [1] and [14], and to a lesser extent in [21] and [22].

We will consider a problem where player 1 (with imperfect information) is attempting to prevent player 2 from taking certain assets. Refer to Figure 1 which is a still shot from the graphic for the MATLAB simulation that runs the example game. Player 1 is depicted with a base (in blue) at the bottom of the figure. Player 2 is depicted with a base (in red) at the top of the figure. The two blue rectangular shapes on either side of the player 1 base represent the positions of two strategic assets belonging to player 1. The two blue lines are meant to indicate routes that player 2 entities (depicted as tanks) could take toward each of the player 1 assets. Each segment of the road constitutes the distance covered by player 2 assets in a single time step. Player 2 may move entities (including say ground vehicles and air defenses, although only a single entity type is used in this example) toward one or

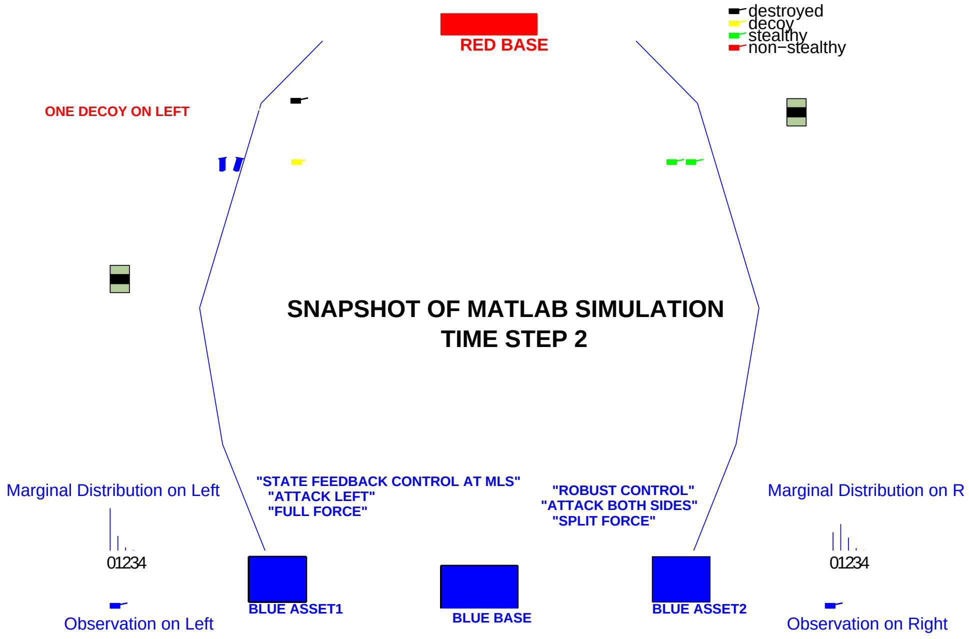


Figure 1:

both of the player 1 assets. The simulation snapshot in Figure 1, is taken between time steps 1 and 2. Player 2 was in the process of moving one tank and one decoy (yellow tank icon) toward the left player 1 asset, and two tanks (their green color indicating stealthy) toward the right player 1 asset. The black tank icon along the left road indicates that at time 1 the tank on the left (which had been operating nonstealthily) was destroyed. As may be intuited from this discussion, modeling of this problem allows player 2 to move the entities in one of two modes: 'stealthy' and 'nonstealthy'. Of course they are detected more easily when they are nonstealthy. Player 1's UAVs are typically expected to be more effective against the player 2 entities when they are moving in the nonstealthy mode (although for study purposes, we include results where the effectiveness against stealthy and nonstealthy entities are identical). We should note that player 2 "takes" the assets by successfully moving (non-decoy) entities to their targets. Player 1 has UAVs which can move to intercept the player 2 entities. In the figure, the missile shaped blue icons moving along the left road indicate player 1 UAVs which are currently attempting to intercept player 2 entities moving along that route.

The current player 1 marginal distributions and observations are also shown at the lower left and lower right of Figure 1. In this snapshot, the annotations indicate that it so happens that player 1 has detected the player 2 entity (which is a decoy) on the left and one of the player 2 entities on the right. The bar graphs indicate the marginal distributions (marginals of q_t) of the numbers of player 2 entities on the left and right as computed by player 1. (These appear due to the use of a maximum likelihood approach being used in this example in contrast with the imperfect information stochastic game approach discussed above. This alternate approach is discussed very briefly below.)

This problem has sufficient complexity in order for diversionary tactics on the part of player 2 to be effective. Specifically, it can be effective for player 2 to move one group of entities toward one player 1 asset in a non-stealthy manner, while moving some other entities toward a different asset in as stealthy a manner as possible. Player 2 may also exaggerate one of the forces (typically the non-stealthy force) by the use of decoys.

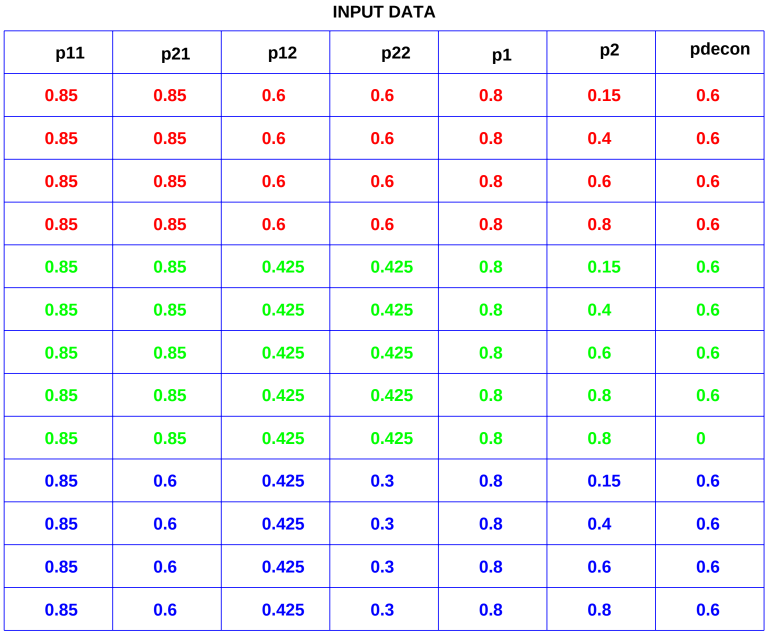
We test two approaches for player 1 in this simple game problem. The standard (naive) approach is for player 1 to simply take the maximum likelihood state estimate of the player 2 state, and to apply a state-feedback control corresponding to this maximum likelihood state. This state feedback control is computed via dynamic programming as applied to the stochastic game described in Section 2 with full state knowledge for both players. We will refer to this approach as the MLS/SF controller. Note that in order to generate a single probability distribution using the above observation model, player 1 needs to assume that the player 2 choice of nonstealthy versus stealthy control is random according to some distribution. In the second approach a robust controller is computed based on the imperfect information stochastic game theoretic method discussed above. As a shorthand, we will refer to this as the *robust* controller.

As one can easily imagine, the MLS/SF control approach for player 1 is open to exploitation by diversionary tactics for player 2. In that case, the maximum likelihood estimate made by player 1 can sometimes be that all or almost all of the player 2 entities are heading for only one of the player 1 assets. If player 1 does not react to the potentiality of player 2 entities moving toward the other asset quickly enough, then the outcome may often be far from optimal. We will see that the diversionary tactic for player 2 can gain quite an advantage against the MLS/SF controller. Further, we test the robust controller, and find that it achieves better results against an intelligent, deceptive opponent than the (naive) MLS/SF approach.

Both these approaches were compared using standard Monte Carlo techniques. Two different models were tested. The major difference between the two being the speed with which player 1 can cross from one side to another. We keep the same diversionary tactic for player 2 in all cases, the player 2 entities on left side are non-stealthy with a decoy whereas the player 2 entities on right are all stealthy. Each model was tested for two different game durations – one where it take player 1 four time steps to reach the assets, and one where it takes five. The details of some parameters and the results in presented in the tables. We do not include all the model parameters nor the values assigned to each entity/asset, but reiterate that player 1 is trying to minimize the cost. The best outcome player 1 could achieve (no assets lost and all player 2 entities destroyed) yields a cost of zero. The worst outcome (both assets lost and no player 2 entities destroyed) yields a cost of 23. Although the actual numbers are irrelevant here, the relative values will be of interest.

We introduce some notation here for the parameters used in the input data table.

- p11: Probability of attrition/kill of non-stealthy player 2 entity, Full force.



INPUT DATA

p11	p21	p12	p22	p1	p2	pdecon
0.85	0.85	0.6	0.6	0.8	0.15	0.6
0.85	0.85	0.6	0.6	0.8	0.4	0.6
0.85	0.85	0.6	0.6	0.8	0.6	0.6
0.85	0.85	0.6	0.6	0.8	0.8	0.6
0.85	0.85	0.425	0.425	0.8	0.15	0.6
0.85	0.85	0.425	0.425	0.8	0.4	0.6
0.85	0.85	0.425	0.425	0.8	0.6	0.6
0.85	0.85	0.425	0.425	0.8	0.8	0.6
0.85	0.85	0.425	0.425	0.8	0.8	0
0.85	0.6	0.425	0.3	0.8	0.15	0.6
0.85	0.6	0.425	0.3	0.8	0.4	0.6
0.85	0.6	0.425	0.3	0.8	0.6	0.6
0.85	0.6	0.425	0.3	0.8	0.8	0.6

Figure 2:

- p12: Probability of attrition/kill of non-stealthy player 2 entity, Split force.
- p21: Probability of attrition/kill of stealthy player 2 entity, Full force.
- p22: Probability of attrition/kill of stealthy player 2 entity, Split force.
- p1: Probability of detecting a non-stealthy player 2 entity.
- p2: Probability of detecting a stealthy player 2 entity.
- pdecon: Probability of detecting a decoy when it is on.

The input parameters are listed in the input data table: Figure 2. These are divided into three sets – distinguished by color (red, green and blue) in the table. Sets 1 and 2 (the first 9 rows of the table – the red and green rows) give player 1 equal kill/strike capability against both stealthy and non-stealthy objects. (Of course player 1 needs to determine some nonzero probability of existence of these entities and decide to attack them first.)

Note that Set 1 gives player 1 somewhat less strike capability if it uses a 'split force' tactic (i.e. uses half of its forces on each side), whereas Set 2 drops the strike capability to half in the 'split force' case. Set 3 (corresponding to the last 4 blue rows) sets the strike capability of player 1 to be lower against stealthy player 2 entities than against non-stealthy entities, and allows only half the strike capability in the 'split force' case.

Intuitively Set 1 should be the best for player 1 and Set 3 should give player 2 the advantage in absolute terms. The results appear in the bottom four tables (discussed below). Also note that within the same set the variation in parameters in different rows is owing to the change in the value of p2 (the observation probability for stealthy player 2 entities).

There are four tables of results: refer to Figures 3 - 6. Game durations of 4 and 5 were allowed, and two models were tested for each duration: thus yielding four variations. As mentioned above, the difference in the two models is through the player 1 control dynamics. Specifically, player 1 vehicles cannot move from one side to the other instantaneously. Player 1 requires two time steps to move all of its vehicles from one side to another in Model 1, and three time steps in Model 2. In each of these results tables, the first two columns correspond to the average outcomes of the MLS/SF and robust controllers – obtained through Monte Carlo iteration. The third and fourth columns indicate the difference and percentage improvement between the first two columns. The rows correspond to the input data sets indicated in the top table.

First we consider the advantage for player 1. It is very clear from the data that the robust approach works better for player 1 than the naive MLS/SF approach (by comparing the first two columns of any Results Table). As one can see from the data corresponding to Set 3, the advantage of robust controller is somewhat reduced if player 1 forces are less effective in attrition potential against stealthy player 2 entities. Thus the robust controller advantage is attenuated if the kill probability for stealthy player 2 entities is lower compared to the non-stealthy kill probability.

It is worth noting that increasing the game duration (T in the above theory) has a natural relationship to the value. It gives player 1 more time to strike and hence gives better results for player 1 in both approaches. Moreover, the robust approach gives comparatively better results as measured by percent improvement in the longer duration case. We provide one such illustration from Figure 5 and 6. Specifically, we are looking at Model 2 Results Table for the Set 1 parameter corresponding to row 1.

For Duration/Terminal Time: $T = 4$

- MLS/SF: Mean payoff = 13.35
- Robust: Mean payoff = 2.344
- Percent advantage: 82.44

For Duration/Terminal Time: $T = 5$

- MLS/SF: Mean payoff = 11.80
- Robust: Mean payoff = 1.134

- Percent advantage: 90.42

Note also that the percent advantage increase in going from Terminal time $T = 4$ to $T = 5$ is higher for Set 3 than for Sets 1 and 2.

It is worth mentioning that in all these results the robust controller happens to be the same for all values of p2; it was not affected by the probability of observing a stealthy entity – **in this particular example**. Consequently, the same (merged) Monte Carlo data appears in the robust control output columns for different values of p2. However, since the naive MLS/SF approach performs more poorly with increasing deception effectiveness (decreasing values of p2), the percent advantage is more for the cases where player 2 deception is more effective. It is useful to note that the advantages of the solution of the imperfect information game are dependent on the effectiveness (or usage) of deception by the opponent. Refer to Figure 6 for this illustration. In particular we choose Set 1 for Model 2 and terminal time $T = 5$.

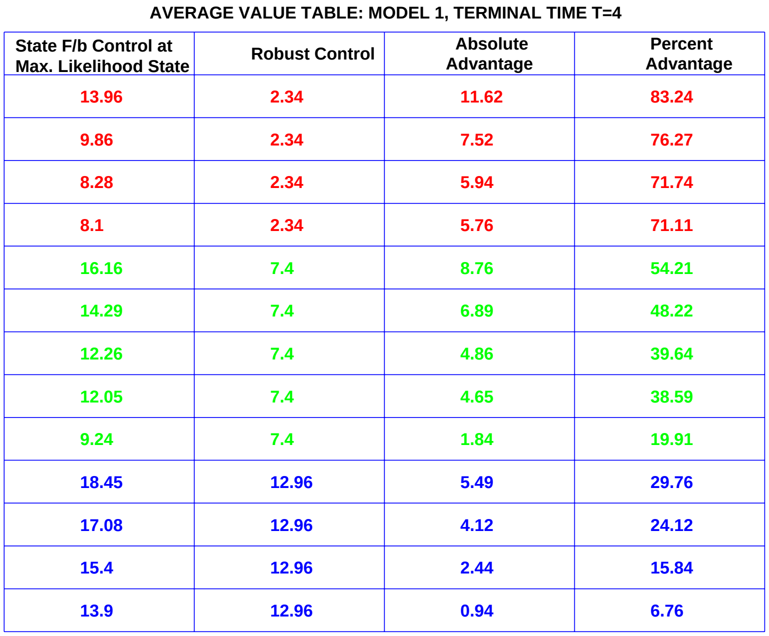
- p2= 0.15, Percent advantage: 90.42
- p2= 0.4, Percent advantage: 86.22
- p2= 0.6, Percent advantage: 79.71
- p2= 0.8, Percent advantage: 78.27

Note that the percent improvement of the robust approach drops as the probability of observing a stealthy entity increases.

Also as one would naturally expect, there is an improvement in outcome (from the point of view of player 2) when player 2 uses a diversionary/deceptive tactic. This is clear by the decrease in average values as p2 increases (observation probability for stealthy player 2 entities increases or deception decreases).

It is worth noting that it is being found that the state feedback control for player 2 is not optimal – even when player 2 has perfect state knowledge. This will be explored in further simulations.

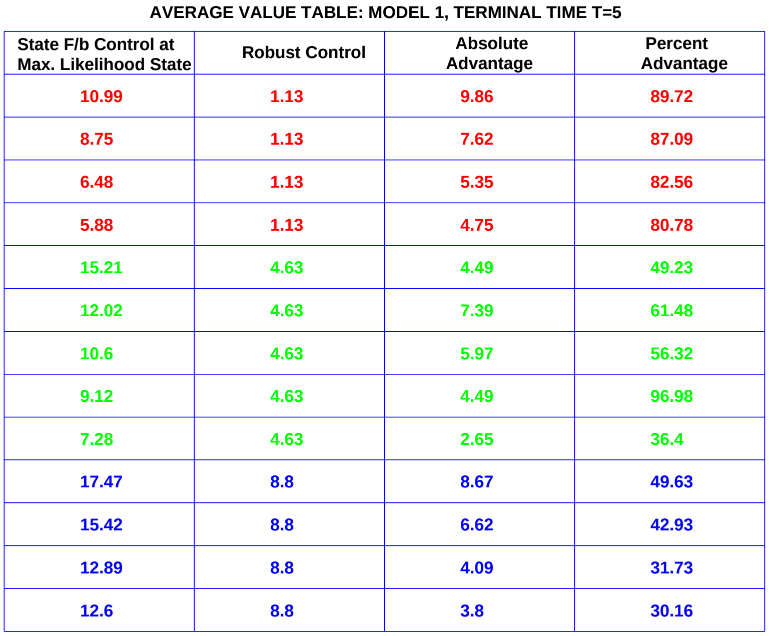
The downside for the imperfect information game (robust) approach is that the computational costs are much greater than for the MLS/SF method. This implies that the fidelity of the model might need to be less for this approach than what one could feasibly compute with when using the MLS/SF method. Trade-offs would need to be studied for any particular problem class that had a relatively complex model, in order to weigh the advantages.



AVERAGE VALUE TABLE: MODEL 1, TERMINAL TIME T=4

State F/b Control at Max. Likelihood State	Robust Control	Absolute Advantage	Percent Advantage
13.96	2.34	11.62	83.24
9.86	2.34	7.52	76.27
8.28	2.34	5.94	71.74
8.1	2.34	5.76	71.11
16.16	7.4	8.76	54.21
14.29	7.4	6.89	48.22
12.26	7.4	4.86	39.64
12.05	7.4	4.65	38.59
9.24	7.4	1.84	19.91
18.45	12.96	5.49	29.76
17.08	12.96	4.12	24.12
15.4	12.96	2.44	15.84
13.9	12.96	0.94	6.76

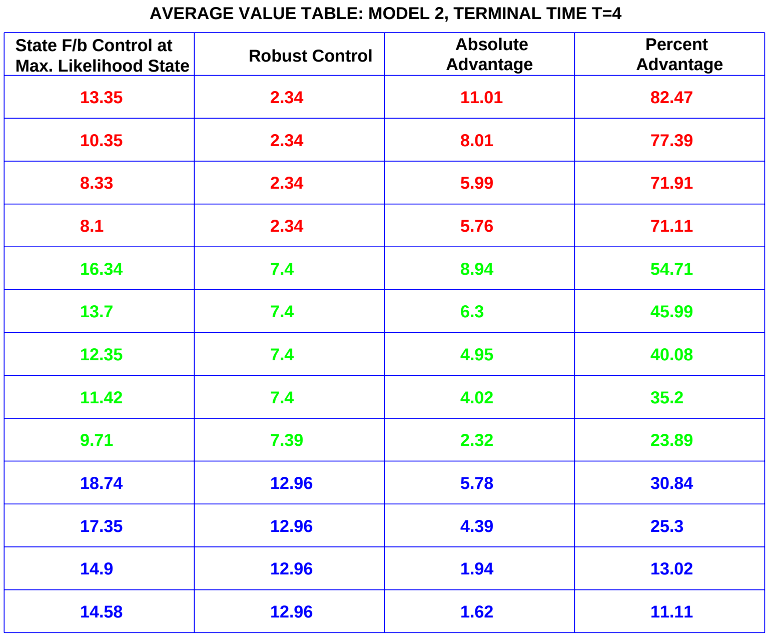
Figure 3:



AVERAGE VALUE TABLE: MODEL 1, TERMINAL TIME T=5

State F/b Control at Max. Likelihood State	Robust Control	Absolute Advantage	Percent Advantage
10.99	1.13	9.86	89.72
8.75	1.13	7.62	87.09
6.48	1.13	5.35	82.56
5.88	1.13	4.75	80.78
15.21	4.63	4.49	49.23
12.02	4.63	7.39	61.48
10.6	4.63	5.97	56.32
9.12	4.63	4.49	96.98
7.28	4.63	2.65	36.4
17.47	8.8	8.67	49.63
15.42	8.8	6.62	42.93
12.89	8.8	4.09	31.73
12.6	8.8	3.8	30.16

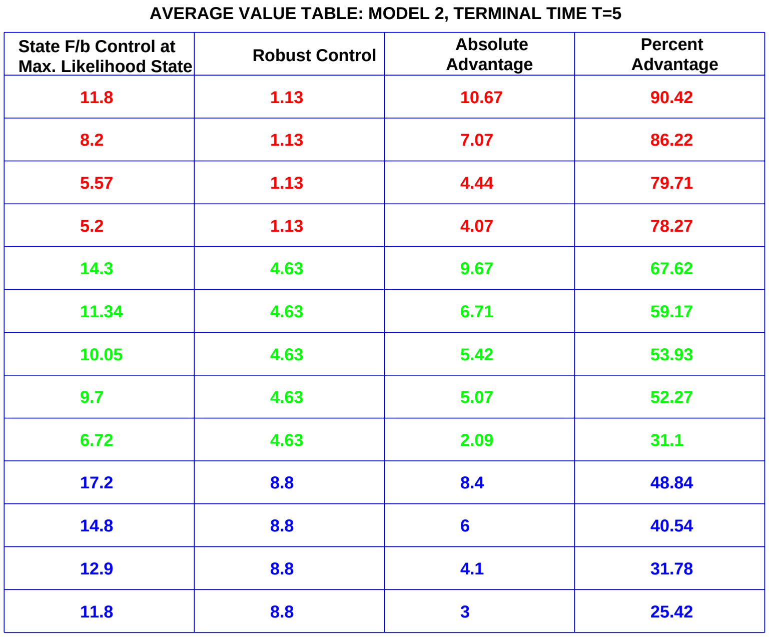
Figure 4:



AVERAGE VALUE TABLE: MODEL 2, TERMINAL TIME T=4

State F/b Control at Max. Likelihood State	Robust Control	Absolute Advantage	Percent Advantage
13.35	2.34	11.01	82.47
10.35	2.34	8.01	77.39
8.33	2.34	5.99	71.91
8.1	2.34	5.76	71.11
16.34	7.4	8.94	54.71
13.7	7.4	6.3	45.99
12.35	7.4	4.95	40.08
11.42	7.4	4.02	35.2
9.71	7.39	2.32	23.89
18.74	12.96	5.78	30.84
17.35	12.96	4.39	25.3
14.9	12.96	1.94	13.02
14.58	12.96	1.62	11.11

Figure 5:



AVERAGE VALUE TABLE: MODEL 2, TERMINAL TIME T=5

State F/b Control at Max. Likelihood State	Robust Control	Absolute Advantage	Percent Advantage
11.8	1.13	10.67	90.42
8.2	1.13	7.07	86.22
5.57	1.13	4.44	79.71
5.2	1.13	4.07	78.27
14.3	4.63	9.67	67.62
11.34	4.63	6.71	59.17
10.05	4.63	5.42	53.93
9.7	4.63	5.07	52.27
6.72	4.63	2.09	31.1
17.2	8.8	8.4	48.84
14.8	8.8	6	40.54
12.9	8.8	4.1	31.78
11.8	8.8	3	25.42

Figure 6:

References

[1] M. Adams, W.M. McEneaney et al., "Mixed Initiative Planning and Control under Uncertainty", Proceedings First AIAA UAV Symposium, Portsmouth, VA, May 22-25, (2002), AIAA-2002-3452.

[2] F.L. Baccelli, G. Cohen, G.J. Olsder and J.-P. Quadrat, *Synchronization and Linearity*, John Wiley (1992).

[3] T. Basar and P. Bernhard, **$\mathbf{H}_\infty$–Optimal Control and Related Minimax Design Problems**, Birkhäuser (1991).

[4] T. Basar and G.J. Olsder, *Dynamic Noncooperative Game Theory*, Classics in Applied Mathematics Series, SIAM (1999), Originally pub. Academic Press (1982).

[5] P. Bernhard, A.-L. Colomb, G.P. Papavassilopoulos, "Rabbit and Hunter Game: Two Discrete Stochastic Formulations", Comput. Math. Applic., Vol. 13 (1987), 205–225.

[6] D.P. Bertsekas, D.A. Castañon, M. Curry and D. Logan, "Adaptive Multi-platform Scheduling in a Risky Environment", Advances in Enterprise Control Symp. Proc., (1999), DARPA–ISO, 121–128.

[7] J.B. Cruz, M.A. Simaan, et al., "Modeling and Control of Military Operations Against Adversarial Control", Proc. 39th IEEE CDC, Sydney (2000), 2581–2586.

[8] R. J. Elliott and N. J. Kalton, "The existence of value in differential games", Memoirs of the Amer. Math. Society, **126** (1972).

[9] W. H. Fleming and E. Pardoux, "Optimal control of partially–observed diffusions", SIAM J. Control and Optim., **20** (1982), 261–285.

[10] W.H. Fleming and P.E. Souganidis, "On the existence of value functions of two–player, zero–sum stochastic differential games", Indiana Univ. Math. Journal, **38** (1989) 293–314.

[11] J. Filar and K. Vrieze, **Competitive Markov Decision Processes**, Springer (1997).

[12] D. Ghose, M. Krichman, J.L. Speyer and J.S. Shamma, "Game Theoretic Campaign Modeling and Analysis", Proc. 39th IEEE CDC, Sydney (2000), 2556–2561.

[13] W.D. Hall and M.B. Adams, "Closed-loop, Hierarchical Control of Military Air Operations", Advances in Enterprise Control Symposium Proc., (1999), DARPA–ISO, 245–250.

[14] S.A. Heise and H.S. Morse, "The DARPA JFACC Program: Modeling and Control of Military Operations", Proc. 39th IEEE CDC, Sydney (2000), 2551–2555.

[15] J.W. Helton and M.R. James, **Extending H_∞ Control to Nonlinear Systems: Control of Nonlinear Systems to Achieve Performance Objectives**, SIAM 1999.

[16] J. Jelinek and D. Godbole, "Model Predictive Control of Military Operations", Proc. 39th IEEE CDC, Sydney (2000), 2562–2567.

[17] P.R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification and Adaptive Control*, Prentice-Hall, 1986.

[18] W.M. McEneaney, "Some Classes of Imperfect Information Finite State-Space Stochastic Games with Finite-Dimensional Solutions", Applied Math. and Optim., (to appear).

[19] W.M. McEneaney, "A Class of Tractable Partially Observed Discrete Stochastic Games", Proc. MTNS 2002.

[20] W.M. McEneaney, "A Class of Reasonably Tractable Partially Observed Discrete Stochastic Games", Proc. 41st IEEE CDC, Las Vegas (2002).

[21] W.M. McEneaney and B.G. Fitzpatrick, "Control for UAV Operations under Imperfect Information", Proceedings First AIAA UAV Symposium, Portsmouth, VA, May 22-25, (2002), AIAA-2002-3418.

[22] W.M. McEneaney, B.G. Fitzpatrick and I.G. Lauko, "Stochastic Game Approach to Air Operations", Submitted to IEEE Trans. on Aerospace and Electronic Sys.

[23] W.M. McEneaney and K. Ito, "Stochastic Games and Inverse Lyapunov Methods in Air Operations", Proc. 39th IEEE CDC, Sydney (2000), 2568–2573.

[24] G.J. Olsder and G.P. Papavassilopoulos, "About When to Use a Searchlight", J. of Math. Analysis and Applics., Vol. 136 (1988), 466–478.