

A Class of Tractable Partially Observed Discrete Stochastic Games

William M. McEneaney ¹

Depts. of Mathematics and Mechanical/Aerospace Engineering

University of California at San Diego

La Jolla, CA 92093-0112, USA

<http://math.ucsd.edu/~wmcenean/>
wmceneaney@ucsd.edu

Abstract

Stochastic games under partial information are typically computationally intractable even in the discrete-time/discrete-state case considered here. We consider a problem where one player has perfect information. A chief problem is that the information state for the player with imperfect information is a function over the space of probability distributions (a function over a simplex), and so infinite-dimensional. However, in the problem form here, the payoff is only a function of the terminal state of the system, and the initial information state is either linear or a sum of max-plus delta functions. In this case, the information state and state-feedback value functions belong to finite-dimensional sets. Thus computational tractability is greatly enhanced.

1 Introduction

For a discrete deterministic game, one can apply dynamic programming techniques to obtain a solution (and “optimal” moves) by computing the appropriate value function. This function is defined over the space of possible system states – possibly annexed by the time variable depending on the game definition. For very simple games this is tractable, but for reasonably complex games such as chess, one must typically apply heuristic techniques and receding horizon approximations in order to reduce the computational complexity. For discrete *stochastic* games, the value function is defined over the space of all possible probability distributions over the state space. Consequently, the problem is much more computationally intensive. Alternatively, for partially observed discrete deterministic H_∞ -type games, the value is defined over the information state space which is again a space of functions over the original state space, and so again tremendously more difficult (cf. [1] and [10]). Further, one needs to propagate this information state forward in real-time based on the observations obtained. (Note that we are being vague here about the particular value function being used; we will be using the Elliott–Kalton definition [4], or more exactly, its extension to stochastic games (cf. [5], [6]). More specific definitions will follow of course.) Finally, for discrete stochastic games with partial observations, the problem is much more complex, and even simple games and their information state formats become quite difficult to analyze.

¹Research partially supported by DARPA grant F30602-99-2-0548 and NSF grant DMS-9971546.

We will be concerned here with a specific class of discrete stochastic games under partial 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. Since this is a first paper on the subject, we will concentrate on a simple (sub)class where one player has perfect information; however, the extension to the case where both players have only partial information will be relatively clear. We consider only zero-sum games.

The motivational application here is the military command and control (C^2) problem for air operations (most likely with uninhabited combat air vehicles (UCAV's)). For related information, see [2], [3], [7], [8], [9], [11]. This application has specific characteristics such that we will be able to construct a reasonable problem formulation which is particularly nice from the point of view of analysis and computation. In particular, the controls which affect the dynamics at each step will be the same controls which affect the observation process (cf. [11]), although this last point is not essential to the analysis. For instance, an air vehicle may obtain information on enemy air defenses by flying to a given waypoint. If the air defense radars turn on, both players may obtain information on the overall system state, but this may also lead to potential loss of the air vehicle and/or air defenses through an engagement. More importantly, we will not include running costs in the model, but only have a terminal cost; one could interpret this terminal cost in the application as the outcome of the battle (e.g. weighted cost of vehicles, air defenses, etc... lost to each side). Although the motivational application is specific to military C^2 , one could easily imagine other applications which one would hope to formulate similarly, given the relatively high level of computational tractability.

2 Problem Formulation

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 . Each state x will be associated with a unit basis vector in $\mathbf{R}^{(\#\mathcal{X})}$. 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 .

Although one could also consider an exit time formulation, we will consider a finite time problem with time taking values in $\{0, 1, 2, \dots, T\}$. We will denote the terminal cost as $\mathcal{E} : \mathcal{X} \rightarrow \mathbf{R}$; 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 (where the dynamics are time independent for simplicity of exposition). 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 by both players. Let the observation process be y . with $y_t \in Y$ where the probability that observation $y_t = \bar{y}$ given $X_t = i$ and controls $u_t = u, w_t = w$ is denoted as

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

In a deterministic game under partial observations, the information state for player 1 is a function of the state, and it represents the minimal cost to the opposing player (maximal cost from the point of view of player 1) for the state to be x at current time t given the observations up to the current time. Alternatively, in a stochastic control problem under partial observations, the information state is simply the probability that $X_t = x$ conditioned on the observations up to the current time t . Here however, player 2 can affect the observation process, so one must consider the cost to player 2 to produce a possibly misleading conditional probability distribution. Thus, it is natural to define an information state for player 1 as $\mathcal{I}_t : Q(\mathcal{X}) \rightarrow \mathbf{R}$ where $Q(\mathcal{X})$ is the space of probability distributions over state space \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 $\mathcal{I}_0(\cdot) = \phi(\cdot)$. 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

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

3 Information State Propagation and Value Function

We work first with the information state which is propagated up to the current time. Let the current time be $t_0 \in \{0, 1, 2, \dots, T\}$. Let the conditional probability of the state at time t be denoted 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 for $t \leq t_0$, u_t is known by player 1 while w_t is unknown. If there was only one possibility for w_t , say \bar{w} , then the information state for player 1 would propagate by

$$\mathcal{I}_{t+1}(q) = \mathcal{I}_t(P^{-T}(u_t, \bar{w})q)$$

for all $q \in \mathcal{Q}_{t+1}$ where P^{-T} is the inverse of P^T and \mathcal{Q}_{t+1} is the set of feasible q at time $t + 1$. (Note that $\mathcal{Q}_{t+1} = P^T(u_t, \bar{w})\mathcal{Q}_t$.) Further, for example in the case where $w_r = \bar{w}$ for all $r \leq t$, one sees that

$$\mathcal{Q}_{t+1} = P^T(u_t, \bar{w})P^T(u_{t-1}, \bar{w}) \cdots P^T(u_0, \bar{w})Q(\mathcal{X}),$$

and we note again that $Q(\mathcal{X})$ is the space of probability distributions over \mathcal{X} . In the more general case, given \mathcal{Q}_t ,

$$\mathcal{Q}_{t+1} = \{\tilde{P}^T(u_t, \vec{w})q : q \in \mathcal{Q}_t, \vec{w} \in W^n\}$$

where

$$\tilde{P}_{ij}(u_t, \vec{w}) \doteq P_{ij}(u_t, \vec{w}_i)$$

for all i, j , and where \vec{w} is a vector of length n with elements in W (i.e. $\vec{w} \in W^n$). Note that this definition allows player 2 to have a control which depends on the true current state of the system – desired since this player has full state knowledge. The information state for player 1 is propagated by

$$\begin{aligned} \mathcal{I}_{t+1}(q) &= \max \left\{ \mathcal{I}_t[\tilde{P}^{-T}(u_t, \vec{w})q] : \vec{w} \in W^n \text{ such that } \tilde{P}^{-T}(u_t, \vec{w})q \in \mathcal{Q}_t \right\} \\ \mathcal{I}_0(q) &= \phi(q) \quad \forall q \in \mathcal{Q}_0 = Q(\mathcal{X}). \end{aligned}$$

Note that $\mathcal{I}_t(\cdot)$ is a piecewise linear, concave function over a subset of the simplex $Q(\mathcal{X})$ which has a piecewise linear boundary. One might also note that the maximum here used to compute the player 1 information state allows \vec{w} to be chosen depending on u_t (upper value). For each possible distribution, q , this represents the maximal cost (minimal from player 2's perspective) for the computed conditional probability to be q given the original cost. Again, in the case that ϕ has the max-plus delta function form $\phi(q) = \delta_{q_0}(q)$ for some $q_0 \in Q(cX)$ and $W = \{\vec{w}\}$, $\mathcal{I}_t(\cdot)$ is also a max-plus delta function at $\vec{q}_t = P^T(u_{t-1}, \vec{w}) \cdots P^T(u_0, \vec{w})q_0$ (i.e. $\mathcal{I}_t(q) = \delta_{\vec{q}_t}(q)$).

So far we have ignored the possibility of an observation process. Let us now include this in the propagation. We will assume that the observations may occur at each time step, t . We will now need to distinguish between a priori conditional distributions, denoted as q_t , and a posteriori distributions, denoted as \hat{q}_t . That is, \hat{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 \tilde{R} have components

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

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

$$\hat{q}_t = \left(\frac{1}{\tilde{R}^T(\bar{y}, u_t, \vec{w})q_t} \right) D(\tilde{R}(\bar{y}, u_t, \vec{w}))q_t = \left(\frac{1}{\sum_i [\hat{q}_t]_i} \right) D(\tilde{R}(\bar{y}, u_t, \vec{w}))q_t. \quad (3.1)$$

The possible set of posteriori distributions, $\hat{\mathcal{Q}}_t$ is the set of all \hat{q}_t given by (3.1) for some

$q_t \in \mathcal{Q}_t$. Thus the a posteriori information state would be

$$\widehat{\mathcal{I}}_t(\widehat{q}) = \max \left\{ \mathcal{I}_t \left[\frac{1}{\widehat{R}^T(\bar{y}, u_t, \bar{w})\widehat{q}_t} D^{-1}(\widetilde{R}(\bar{y}, u_t, \bar{w}))\widehat{q}_t \right] : \bar{w} \in W^n \text{ such that } \frac{1}{\widehat{R}^T(\bar{y}, u_t, \bar{w})\widehat{q}_t} D^{-1}(\widetilde{R}(\bar{y}, u_t, \bar{w}))\widehat{q}_t \in \mathcal{Q}_t \right\}$$

where \widehat{R} is the vector of components \widehat{R}_i^{-1} .

A problem is that the normalization in (3.1) induces nonlinearities in the propagation. Consequently, we will work with the unnormalized distribution. The a priori and a posteriori unnormalized distributions at time t will be denoted as \widetilde{q}_t and $\widehat{\widetilde{q}}_t$, respectively. At any time t , one can renormalize by dividing by $\sum_i [\widetilde{q}_t]_i$ for the a priori distribution, and similarly for the a posteriori. The feasible sets of a priori and a posteriori unnormalized distributions will be denoted by $\widetilde{\mathcal{Q}}_t$ and $\widehat{\widetilde{\mathcal{Q}}}_t$, where the propagation formulae are obvious.

Let us suppose that observations occur at each time step. If the control processes, u , and \bar{w} , and the observation process, \bar{y} , are given, then the unnormalized distribution would propagate as

$$\widetilde{q}_{t+1} = \widetilde{P}^T(u_t, \bar{w}_t)\widehat{\widetilde{q}}_t, \quad \widehat{\widetilde{q}}_t = D(\widetilde{R}(\bar{y}_t, u_t, \bar{w}_t))\widetilde{q}_t \quad (3.2)$$

for given initial $\widetilde{q}_0 = q_0$. The information state as a function of the unnormalized distribution, denoted by $\widetilde{\mathcal{I}}_t$, propagates by

$$\widetilde{\mathcal{I}}_{t+1}(\widetilde{q}) = \max \left\{ \widetilde{\mathcal{I}}_t [D^{-1}(\widetilde{R}(\bar{y}_t, u_t, \bar{w}))\widetilde{P}^{-T}(u_t, \bar{w})\widetilde{q}] : \bar{w} \in W^n \text{ such that } D^{-1}(\widetilde{R}(\bar{y}_t, u_t, \bar{w}))\widetilde{P}^{-T}(u_t, \bar{w})\widetilde{q} \in \widetilde{\mathcal{Q}}_t \right\} \quad (3.3)$$

where

$$\widetilde{\mathcal{Q}}_{t+1} = \left\{ q \in Q(\mathcal{X}) : \exists q_t \in \widetilde{\mathcal{Q}}_t, \bar{w} \in W^n \text{ such that } q = \widetilde{P}^T(u_t, \bar{w})D(\widetilde{R}(\bar{y}_t, u_t, \bar{w}))q_t \right\} \quad (3.4)$$

with initial conditions $\widetilde{\mathcal{I}}_0(q) = \phi(q)$ and $\widetilde{\mathcal{Q}}_0 = Q(\mathcal{X})$. Assume that the initial cost, ϕ is linear. We see that, even when including the observation process, the *unnormalized* information state remains a piecewise linear, concave function on a convex subset of the simplex $Q(\mathcal{X})$ with piecewise linear boundary. Although we will not consider the actual computational algorithm here, the propagation of this information state is clearly tractable in real-time for reasonably small problems.

We now turn to the state feedback value function. When the Certainty Equivalence Principle holds, this can be combined with the information state to obtain the “optimal” controls. The full state of the system is now described by the true state taking values $x \in \mathcal{X}$ and the player 1 information state taking values $q \in Q(\mathcal{X})$. As before, we denote the terminal payoff for the game as $\mathcal{E} : \mathcal{X} \rightarrow \mathbf{R}$ (where of course this does not depend on the internal information state of player 1). Thus the state feedback value function at the terminal time is

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

The state feedback value can be propagated backward in time via dynamic programming. 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, one obtains nice robustness properties if it is also assumed to depend on (i.e. know) the conditional distribution, q . This is the form to be assumed here, however, one could certainly investigate the other option as well. Thus the state feedback value propagates backward according to

$$V_t(x, q) = \sum_{j \in \mathcal{X}} \tilde{P}_{xj}(u_t^0, \bar{w}_t^0) V_{t+1}(j, q'(q, u_t^0, \bar{w}_t^0)) \quad (3.5)$$

where

$$q'(q, u_t^0, \bar{w}_t^0) = \tilde{P}^T(u_t^0, \bar{w}_t^0)q \quad (3.6)$$

$$\bar{w}_t^0 = \operatorname{argmax}_{\bar{w} \in W^n} \left\{ \sum_{j \in \mathcal{X}} \tilde{P}_{xj}(u_t^0, \bar{w}) V_{t+1}(j, q'(q, u_t^0, \bar{w})) \right\} \quad (3.7)$$

$$u_t^0 = u_t^0(q) = \operatorname{argmin}_{u \in U} \mathbf{E}_q \left\{ \max_{\bar{w} \in W^n} \left[\sum_{j \in \mathcal{X}} \tilde{P}_{xj}(u, \bar{w}) V_{t+1}(j, q'(q, u, \bar{w})) \right] \right\}. \quad (3.8)$$

where \mathbf{E}_q indicates expectation (over x) with respect to distribution q . Consequently, $V_t(x, \cdot)$ is a piecewise constant function over simplex $Q(\mathcal{X})$.

Due to this piecewise constant nature, propagation is relatively straight-forward (in particular, it is finite-dimensional in contradistinction to the general case). However, this is slightly less critical than the propagation issue for the information state of the unnormalized distribution, $\tilde{\mathcal{I}}_t$, since the state feedback value may be pre-computed, while the information state must be propagated in real-time.

The last step in the computation of the control at each time instant is now discussed. Due to the page limit, and the necessarily long description of the information state, not all the details will be given here. The control computation for such games is typically performed via the use of the Certainty Equivalence Principle (cf. [1], [10]). As with the Separation Principle in stochastic control, the Certainty Equivalence Principle is only proven for a limited class of problems. Using the Certainty Equivalence Principle, the control to be applied by player 1 at time t in the partially observed case is obtained by computing

$$q_t^0 \doteq \operatorname{argmax}_{q \in Q(\mathcal{X})} \{ \mathcal{I}_t(q) + \mathbf{E}_q V_t(x, q) \}. \quad (3.9)$$

Note here that this uses \mathcal{I}_t not $\tilde{\mathcal{I}}$ (the function of unnormalized distribution), and one transforms via the transformation from unnormalized \tilde{q} to normalized q . Alternatively, it may sometimes be computationally more efficient to do the maximization in the unnormalized space since $\tilde{V}_t(x, \tilde{q}) \doteq V_t(x, q(\tilde{q}))$ remains piecewise constant; in that case, one would compute $\tilde{q}_t^0 \doteq \operatorname{argmax}_{\tilde{q}} \{ \tilde{\mathcal{I}}_t(\tilde{q}) + \mathbf{E}_q \tilde{V}_t(x, \tilde{q}) \}$, and then transform to obtain q_t^0 . The optimal control for player 1 is then the value of u_t^0 obtained from (3.8) as

$$u_t^m = u_t^0(q_t^0).$$

Under sufficiently strong certainty equivalence-type conditions, one then has standard robust game inequalities. For instance, the following result which is quite easily proved.

Theorem 3.1. *Suppose u_t^m is a strict minimizer. Then, given any $\tilde{u}_t \neq u_t^m$, there exist q^1, \tilde{w}^1 and $\varepsilon > 0$ such that*

$$\begin{aligned} & \left\{ \mathcal{I}_t(q^1) + \mathbf{E}_{q^1} \left[\sum_{j \in \mathcal{X}} \tilde{P}_{xj}(\tilde{u}_t, \tilde{w}^1) V_{t+1}(j, q'(q^1, \tilde{u}_t, \tilde{w}^1)) \right] \right\} \\ & > \max_{q \in Q(\mathcal{X})} \left\{ \mathcal{I}_t(q) + \mathbf{E}_q \max_{\tilde{w} \in W^n} \left[\sum_{j \in \mathcal{X}} \tilde{P}_{xj}(u_t^m, \tilde{w}) V_{t+1}(j, q'(q, u_t^m, \tilde{w})) \right] \right\} + \varepsilon. \end{aligned}$$

4 Computational Tractability

Although one can obtain results such as Theorem 3.1, a main motivation 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 partially observed problems is the propagation of the information state forward in real-time. A secondary problem is of course the computation of the argmax in (3.9). We briefly discuss computational tractability for two cases: linear ϕ and max-plus delta function ϕ . A key to the tractability is that the costs are only initial and final, and in particular, the cost to the players to affect the observation process is only indirectly felt through the effects those same controls may have on the state process. (For example, in the military application referred to above, this effect might be the loss of aircraft whose controlled trajectories not only lead to observations but also to potential loss of the aircraft.)

Consider the case of linear $\tilde{\mathcal{I}}_0 = \mathcal{I}_0 = \phi$. The propagation of $\tilde{\mathcal{I}}$ is given by (3.3) (with domain propagation (3.4)). In the case where there is only one choice of control for the player 2, this would simply be a linear mapping of the underlying distribution, and so $\tilde{\mathcal{I}}_t(\cdot)$ would remain a linear function. Note that the domain remains a simplex subset of an affine hyperplane, but this may not be the initial simplex $Q(\mathcal{X})$. In the more realistic situation where W is not a single point (but recall that it is still assumed finite), this leads to a piecewise linear $\tilde{\mathcal{I}}_t$ over a simplex subset of an affine hyperplane. Thus, propagation of the information state forward in time is a finite-dimensional process, and consequently reasonably tractable. As noted above, the transformed version of $V_t(x, q)$, $\tilde{V}_t(x, \tilde{q}) \doteq V_t(x, q(\tilde{q}))$, remains piecewise constant. Thus $\tilde{\mathcal{I}}_t(\tilde{q}) + \tilde{V}_t(x, \tilde{q})$ is a discontinuous piecewise linear function. (That is, it consists of a union of linear pieces, and may be discontinuous along the boundaries of the pieces.) Consequently the argmax computation reduces to a comparison among a finite set of maxima of each of the linear pieces.

The case where ϕ is a max-plus delta function, i.e. $\phi(q) = \delta_{q_c}(q)$ for some $q_c \in Q(\mathcal{X})$, leads to a particularly tractable problem. Recall that this case corresponds to a model where the initial distribution for player 1 state information is not subject to disruption by some initial control of player 2. (More specifically, such a control is not considered within the game.) In this case, $\tilde{\mathcal{I}}_t$ is 0 only at a finite number of points, and is $-\infty$ elsewhere. Thus, \mathcal{I}_t retains this property. The propagation of these points proceeds by (3.2) for each possible player 2 control. Thus, the information state is easily propagated. Further, since \mathcal{I}_t is not $-\infty$ at

only a finite number of points, the argmax computation of (3.9) involves only comparison of a finite number of values of $\max_x V_t(x, q)$ for these select values of q .

References

- [1] T. Basar and P. Bernhard, **H_∞ –Optimal Control and Related Minimax Design Problems**, Birkhäuser (1991).
- [2] D.P. Bertsekas, D.A. Castañón, M. Curry and D. Logan, “Adaptive Multi-platform Scheduling in a Risky Environment”, *Advances in Enterprise Control Symp. Proc.*, (1999), DARPA–ISO, 121–128.
- [3] 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.
- [4] R. J. Elliott and N. J. Kalton, “The existence of value in differential games”, *Memoirs of the Amer. Math. Society*, **126** (1972).
- [5] J. Filar and K. Vrieze, **Competitive Markov Decision Processes**, Springer (1997).
- [6] 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.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] J.W. Helton and M.R. James, **Extending H_∞ Control to Nonlinear Systems: Control of Nonlinear Systems to Achieve Performance Objectives**, SIAM 1999.
- [11] W.M. McEneaney and K. Ito, “Stochastic Games and Inverse Lyapunov Methods in Air Operations”, *Proc. 39th IEEE CDC*, Sydney (2000), 2568–2573.