Sequential Decision Making for Intelligent Agents
Papers from the AAAI 2015 Fall Symposium

Complexity of Self-Preserving, Team-Based Competition in Partially Observable Stochastic Games

M. Allen
Computer Science Department
University of Wisconsin-La Crosse
La Crosse, Wisconsin

Abstract

Partially observable stochastic games (POSGs) are a robust and precise model for decentralized decision making under conditions of imperfect information, and extend popular Markov decision problem models. Complexity results for a wide range of such problems are known when agents work cooperatively to pursue common interests. When agents compete, things are less well understood. We show that under one understanding of rational competition, such problems are complete for the class NEXP^{NP}. This result holds for any such problem comprised of two competing teams of agents, where teams may be of any size whatsoever.

Introduction

Markov decision processes (MDPs) are a well known mathematical model of a single agent taking actions with uncertain outcomes, modeled probabilistically, and have over the decades begot numerous variations, including partially observable models (POMDPs), in which the agent is uncertain not only about action outcomes, but also about their environment. Decentralized MDPs and POMDPs extend the models to cases in which multiple agents act cooperatively in order to maximize the utility of the group. Finally, partially observable stochastic games (POSGs) allow that agents may have divergent interests, so that competition may arise where policies benefit one agent, or set of agents, over others. As such, POSGs provide an exact mathematical framework for the analysis of multiagent decision making in a wide range of real-world contexts in which groups of agents must negotiate uncertainty as they seek to maximize their utility. The general POSG model encompasses many others, and understanding that model provides insight into many planning and learning problems.

The computational complexity of many of these various sorts of decision problems has been extensively studied, dating at least to (Papadimitriou and Tsitsiklis 1987), where it was shown that for both finite-horizon cases (where all policies of action must come to an end by some fixed, finite point) and infinite-horizon cases (where policies may continue indefinitely), MDPs are P-complete, while finite-horizon POMDPs are harder, being complete for PSPACE. More details on the finite MDP case is given by (Mundhenk et al. 2000). (Lusena, Mundhenk, and Goldsmith 2001) showed that POMDPs are not generally approximable, and (Madani, Hanks, and Condon 2003) showed that for the infinite-horizon case POMDPs are in fact undecidable.

Interest in the complexity of stochastic games goes back as far as (Condon 1992), where it was shown that simple games where players compete with shared perfect information were in the class NP $\cap$ co-NP. For decentralized POMDPs (Dec-POMDPs), where the agents cooperate, but the information is imperfect and private, (Bernstein et al. 2002) showed the problems to be complete for nondeterministic exponential time (NEXP), representing a significant upgrade in difficulty. Since the Dec-POMDP incorporates a wide range of other formal models of decision making (see (Goldman and Zilberstein 2004) and (Seuken and Zilberstein 2008) for surveys), this indicated that many interesting real-world problems were unlikely to yield to optimal solution. Following on this work, (Becker et al. 2004) showed that the problems became "merely" NP-complete under stringent restrictions on the ways in which agents interacted—namely if they shared a common reward function, and might affect what one another observed, but otherwise acted with complete independence from one another. While a number of other restrictions on the basic model have been suggested, under many of these assumptions they remain NEXP-hard (Allen and Zilberstein 2009).

In the general POSG case, once competition is possible between agents, things become much less clear. In part, this is due to the fact that game theory does not always dictate a particular solution concept. It is well known via such as the Prisoner's Dilemma that equilibria of various sorts are not always best-possible solutions, and other candidates, like zero-regret strategies have their own quirks. (Goldsmith and Mundhenk 2008) considers a particular version of this question, whether one team in a two-team game has a strategy with guaranteed positive expectation, no matter what strategy is followed by the other team, and show that it is complete for the (highly complex) class NEXP^{NP}(so long as each team has at least two members on it).

Our work here follows up on this line of thought, but departs from the "all-out" understanding of competition, in which a team seeks a policy that guarantees good results no matter what their opponents do. Under this notion, the team is only successful if they can expect positive reward even in cases where their opponents do not have any such expectation and may even expect lower reward yet. Instead, we suggest another possible definition of rational competition, under which the first team seeks a policy that provides positive expectation *so long as the other team does also*, preventing for instance self-sabotage by those who wish more to impose costs on others than to gain rewards themselves. We show that this class of problems is also complete for NEXP$^{\text{NP}}$, and that the result holds no matter what size the teams have. This demonstrates that competitive problems remain highly difficult in general under at least two different ways of measuring success, and provides another piece in the framework of results about utility maximization and decision making in sequential and stochastic domains.

Basic Definitions

We begin by defining two main constructs: the decision problems for which we wish to determine complexity, and those used in reduction proofs to follow.

Partially Observable Stochastic Games

A POSG involves two or more agents seeking to maximize utility under conditions of probabilistic uncertainty about their environment and about the outcomes of their actions. Our definition follows the approaches of (Bernstein et al. 2002) and (Hansen, Bernstein, and Zilberstein 2004).

Definition 1. A *partially observable stochastic game* is a tuple $\mathcal{G} = (I, S, s_0, \{A_i\}, \{\Omega_i\}, P, O, \{R_i\})$, where:

- I is a finite, indexed set of n agents, $\{1, \ldots, n\}$.
- S is a finite set of system states, with starting state s_0.
- For each agent i, A_i is a finite set of available actions. A *joint action*, $(a_1, \ldots, a_n) \in \times_{i \in I} A_i$ is a sequence of n actions, one per agent.
- For each agent i, Ω_i is a finite set of observations. *Joint observations* $(o_1, \ldots, o_n)$ are defined like joint actions.
- P is a table of transition probabilities. For each pair of states s, $s' \in S$, and each joint action $(a_1, \ldots, a_n)$, the value $P(s' \mid s, a_1, \ldots, a_n)$ is the (Markovian) probability that the system enters s' from s, following that action.
- O is a table of observation probabilities. For each pair of states $s, s' \in S$, each joint action $(a_1, \ldots, a_n)$, and each joint observation $(o_1, \ldots, o_n)$, the value $O(o_1, \ldots, o_n \mid s, a_1, \ldots, a_n, s')$ is the probability of that observation following the given state-action transition.
- For each agent i, $R_i : S \times_{i \in I} A_i \times S \to \Re$ is a (real-valued) reward function. $R_i(s, a_1, \ldots, a_n, s')$ agent i's accrued reward after the given state-action transition.

As already described, a Dec-POMDP, where agents have common ends and maximize utility via cooperation, is a special case of the general model described here, in which each reward function R_i is identical. A POSG with only a single agent is simply a POMDP.

In such a problem, the system begins in start-state s_0, and then transitions state-by-state according to joint actions taken by the agents, who receive generally imperfect information about the underlying system via their own, private observations. For agent i, a *local history of length* t is a sequence of observations over time, $\mathbf{o}_i^t = (o_{i_1}, \ldots, o_{i_t}) \in \Omega_i^t$. The set of all local histories for agent i, up to some maximum length T, is then $\mathbf{\Omega}_i^T = \cup_{t=1}^T \Omega_i^t$. For all n agents, a sequence of local histories of same length t forms a *joint history*, written $\mathbf{o}_{1,n}^t = (\mathbf{o}_1^t, \ldots, \mathbf{o}_n^t)$.

Each agent i acts based on a *history-based local policy*, i.e. a function from local histories to actions, $\pi_i : \mathbf{\Omega}_i^T \to A_i$. A *joint policy* $\Pi = (\pi_1, \ldots, \pi_n)$ is a sequence of policies, one for each agent. For any joint history of length t, the composite policy Π yields a unique joint action, written $\Pi(\mathbf{o}_{1,n}^t) = (\pi_1(\mathbf{o}_1^t), \ldots, \pi_n(\mathbf{o}_n^t))$.

For any joint policy Π, states s, s', and joint history, the probability of making the transition from s to s' while each agent i observes its own local portion of that history is defined inductively on its length t. In the base case, where $t = 0$ and ϵ is the empty history, a sole deterministic transition is possible: $\mathbf{P}^\Pi(s, \epsilon, \ldots, \epsilon, s) = 1$. For histories of length $t \geq 1$, we define $\mathbf{P}^\Pi(s, \mathbf{o}_{1,n}^t, s')$ as the product of (a) its single last state-observation probability and (b) the probability of the sub-history leading up to that point:

$$\sum_{s'' \in S} \mathbf{P}^\Pi(s, \mathbf{o}_{1,n}^{t-1}, s'') \cdot P(s' \mid s'', \Pi(\mathbf{o}_{1,n}^{t-1})) \cdot P(o_{1_t}, \ldots, o_{n_t} \mid s'', \Pi(\mathbf{o}_{1,n}^{t-1}), s'),$$

where each component-history in $\mathbf{o}_{1,n}^t$ is $\mathbf{o}_i^t = \mathbf{o}_i^{t-1} o_{i_t}$.

For each agent i, the *expected value* of a joint policy Π, starting in state s and proceeding for t steps, is given by the weighted sum of rewards available to the agent under that policy, computed over all possible local histories of length up to and including $t - 1$:

$$EV_i^t(s \mid \Pi) = \sum_{k=1}^{t-1} \sum_{\mathbf{o}_{1,n}^k} \sum_{s'' \in S} \sum_{s' \in S} \mathbf{P}^\Pi(s, \mathbf{o}_{1,n}^k, s'') \cdot P(s' \mid s'', \Pi(\mathbf{o}_{1,n}^k)) \cdot R(s'', \Pi(\mathbf{o}_{1,n}^k), s').$$

We are interested in problem domains with a finite time-horizon, and so we set a limit $T = |G|$, such that the maximum number of time-steps for which agents must act is limited to the size of the problem description. (Infinite-horizon problems are undecidable, since infinite-horizon POMDPs are a sub-case (Madani, Hanks, and Condon 2003).) Further, since a POSG always begins in state s_0, the value of any policy Π for any agent i can be abbreviated as:

$$EV_i(\Pi) = EV_i^T(s_0 \mid \Pi).$$

Tiling Problems

In a tiling problem, the goal is to completely fill a square grid of size $(N \times N)$ with unit-square tiles. Each tile is chosen from a set of *tile-types* L, with no limits on the number of tiles of each type. A tiling is *valid* if the placement of tiles is consistent with each of two sets of *constraints* H and V, describing what types of tiles are allowed to be placed next to one another horizontally or vertically, respectively. Figure 1 shows a simple example, with one possible valid solution.

Tiling problems seem to have been first introduced by (Wang 1961), in connection with systems of logical proof.

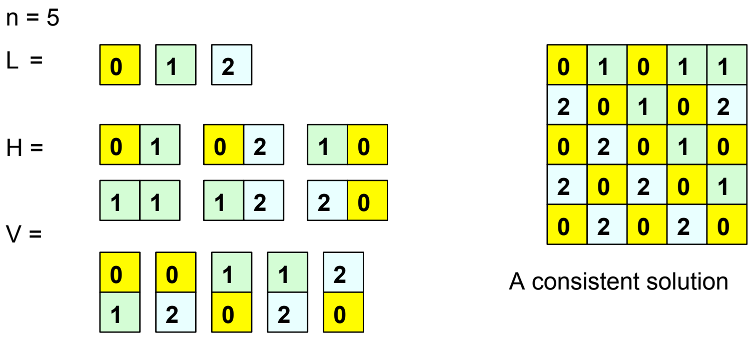

Figure 1: A (5×5) tiling problem instance, with one possible valid solution.

As a decision problem, the question whether a valid tiling exists for a given problem instance has been remarkably useful in computational complexity. As discussed by (Lewis 1978) and (Savelsbergh and van Emde Boas 1984), when the size of the board, N, is given in logarithmic fashion (typically in binary), then the decision question is complete for nondeterministic exponential time (NEXP). Using a unary representation of N, the complexity is reduced to NP-completeness (meaning that tiling is NEXP-complete, but not strongly so). A variety of uses of the problem and its variants can be found in (Papadimitriou 1994).

(Goldsmith and Mundhenk 2008) use a version of tiling called the *exponential square* problem, in which the value N is given in unary, but the board to be tiled is is presumed of size $(2^N \times 2^N)$. This is thus simply a version of the base problem, with larger input sizes, and remains NEXP-complete. Of more interest is the more complex problem they introduce, called the Σ_2 tiling problem which asks whether a valid tiling of the grid exists with a bottom row that never appears as the top row of any valid tiling (the same, or different). Intuitively, this is analogous to asking whether some exponential time computation exists in which the final state of the machine and its computation tape are never the starting state for any other such computation. This latter problem, by Theorem 2.2 of the cited work, is complete for the class NEXP$^{\text{NP}}$. This class—a generally unfamiliar one, as they note—is the set of problems decidable in exponential time by a nondeterministic machine with access to an NP *oracle*. That is, such a machine can, during its computation, ask for and receive answers to a problem in NP “for free” (that is, without any cost to be factored into the overall runtime of the algorithm). Equivalently, such problems are those decidable by a NEXP machine that makes exactly one query to a co-NP oracle.

As noted, while (Goldsmith and Mundhenk 2008) work with an exponential square version of tiling, that detail is not important to the complexity results they generate, and is really a matter of preference. Our work here draws also on that of (Bernstein et al. 2002), in which the input value N is given in logarithmic form; thus, to better unify the results of those papers with our own, we define it as follows:

Definition 2. An instance of the Σ_2 *tiling problem* consists of a tiling problem instance with grid size N given in binary form; the decision problem is whether a valid tiling T exists with bottom row r such that, for any valid tiling T', the top row of T' is not equal to r.

Known Results and Proof Techniques

Our results draw upon and extend two earlier research projects, one of which showed that POSGs in which agents cooperate are complete for nondeterministic exponential time, and one of which showed that teams of competing agents can increase that complexity.

Cooperative POSGs

(Bernstein et al. 2002) showed that Dec-POMDPs (i.e. POSGs with a common reward function, in which agents maximize expected value cooperatively) are NEXP-complete; as usual, the optimization problem of finding a joint policy that maximizes collective reward is re-framed as a decision problem, asking the ***cooperative question***, namely whether there exists a joint policy Π under which every agent has positive expectation:

$$\forall i \in I, EV_i(\Pi) > 0. \tag{1}$$

Here, the upper bound is achieved by showing how to convert any such problem, featuring $n \geq 2$ agents, along with a policy that has been guessed nondeterministically, first into an equivalent single-agent POMDP, and then into an equivalent belief-state MDP. Verifying the value of the guessed policy in that MDP can then be done in polynomial time; however, the size of the final problem version is exponential in the size of the original, yielding nondeterministic exponential time (NEXP) overall.

For the lower bound, and completeness, the basic tiling problem is reduced to a 2-agent Dec-MDP, which is a Dec-POMDP for which the underlying system state can be computed with certainty, if one is given access to observations of all agents. Specifically, any instance of the tiling problem is turned into a Dec-MDP in which each agent is given one randomly chosen location on the board, and responds with a tile to be placed at that location. Rewards are designed so that agents have positive expected value if and only if they know a valid tiling of the entire game board, establishing NEXP-completeness of Dec-POMDPs.

This reduction proof has the following important features:

Logarithmic representation When converting the Dec-POMDP to an instance of tiling, it would be a mistake to have the locations chosen be part of the state space of the new problem. Doing so would result in a Dec-POMDP with state-space of size at least N^2, which would then be exponential in the size of the original problem for sufficiently large N, since the tiling problem encodes the value N using a logarithmic binary representation. Thus, the Dec-POMDP reveals the locations to each agent bit-by-bit, sacrificing time to convey information for space, and ensuring that the size remains polynomial in that of the original tiling instance.

Necessary decentralization It is a key necessary feature of these reductions that there be at least two agents, each of which only knows one location. Should the agents know both locations at any point, then the proof breaks down, since it is possible for them to feign a valid tiling even though none exists. (For instance, if the agents knew the two locations were the same, they could reply with some same pre-arranged identical tiles.)

Proof against cheating Not only is decentralization necessary to prevent gaming the system, agents must also echo back the location they were given when choosing tiles. This allows the system to compute whether or not the locations and tiles chosen are consistent, without requiring that it remember those locations itself (as already described, requiring such a system memory would violate space requirements for the reduction). Thus, a fixed number of bits of each location are retained by the system and used to validate what the agents echo back—meanwhile, agents are unaware of the exact bits recorded, and again cannot dupe the system.

Competitive POSGs

(Goldsmith and Mundhenk 2008) showed that certain forms of competition between teams of agents increased complexity. In particular, they consider POSGs with $n = 2k$, $k \geq 2$ agents, divided into two teams of size k. They then ask the ***all-out competitive question*** (our term for it, not theirs), namely whether there exists some set of policies for the first team under which each agent on that team has positive expectation, no matter what the other team may do:

$$\exists \pi_1, \ldots, \pi_k, \forall \pi_{k+1}, \ldots, \pi_{2k}, \forall i \leq k, EV_i(\Pi) > 0, \quad (2)$$

where joint policy $\Pi = (\pi_1, \ldots, \pi_k, \pi_{k+1}, \ldots, \pi_{2k})$.

It is then determined that this problem is complete for NEXP$^{\text{NP}}$. As already discussed, this is the class decidable in exponential time by a nondeterministic machine that makes a single oracle query for the solution to a problem in co-NP, a fact used in the proof of upper bounds. Prior work showed that under *stationary policies*—i.e., those based on single observations rather than histories—the cooperative problem for a single-agent POMDP is NP-complete (Mundhenk et al. 2000). Similar techniques are used to show that in a POSG, a set of stationary policies can be guessed, and their expected values checked, in polynomial time, placing for example the cooperative problem (Eq. 1) for stationary polices in NP. This result also means that the question of whether all agents have positive expectation under *all possible* stationary policies is in co-NP, since we can answer no by simply guessing and checking a single counter-example policy under which some agent has non-positive expectation.

Finally, based on this fact, the NEXP$^{\text{NP}}$upper bound for the competitive problem (Eq. 2) is shown via a constructive proof: for any POSG $\mathcal{G}$, a set of history-based policies for the first team is guessed, and then a new POSG $\mathcal{G}'$ is built in which the second team alone must act. In $\mathcal{G}'$, the system reproduces joint actions comprised of those chosen by the first team's guessed policies and those now chosen by the second team, via the state-transition and observation functions. (Since $\mathcal{G}'$ is exponentially larger than $\mathcal{G}$, this is a NEXP algorithm so far.) Rewards in $\mathcal{G}$ are "re-routed" so that each member of team two now receives the reward that would have been received in $\mathcal{G}$ by a matching member of team one under the corresponding joint action. Finally, it is shown that the expectation for any stationary policy by a member of the second team in G' is identical to the expectation for the associated, history-dependent first-team policy guessed for $\mathcal{G}$. Thus, all agents in G' have positive expectation under every stationary policy if and only if all agents in team one have positive expectation no matter what team two does in $\mathcal{G}$, which places the original problem in NEXP$^{\text{NP}}$.

Lastly, NEXP$^{\text{NP}}$-completeness is established via reduction from Σ_2 tiling. For a given tiling instance, a composite POSG is created that first establishes whether the first team of $k \geq 2$ agents know a valid tiling, before checking that the second team knows one as well. Rewards are set up so that the first team has a positive reward so long as a valid tiling does exist, unless every such tiling has a bottom row that appears at the top of some valid tiling as well—in the latter case, the second team can mirror that bottom row at the top of their own valid tiling, and deny the first team positive expected reward.

As in the NEXP-completeness proofs for Dec-POMDPs discussed above, the reduction portion of this proof features the need for a logarithmic (i.e., binary) representation of locations on the board, so as not to over-inflate the state-space size upon reduction. As discussed, the use of the "exponential" version of tiling here is non-essential, and the same result could be had for the one in which the board size N is given logarithmically (as in Definition 2). In addition, the POSG again features checks to ensure both that no single agent, nor team of agents, can cheat to achieve positive expectation without actually possessing a proper tiling of the grid. Other important features are:

One-sided rationality We have termed the competitive question in (Eq. 2) a form of "all-out" competition, since the question is simply whether or not the first team of players in a POSG has a policy with positive expectation, *no matter what* the second team does—even if the competition itself is following a policy with non-positive expectation. Thus, while a positive answer means that team one is guaranteed some form of a "win" in the game being played, a negative answer *does not* mean that team two is guaranteed its own "win" in turn.

Minimal team-sizes A key element of the proofs cited is that each team in the POSG contain at least 2 agents. By construction, each team must be separate, to prevent team two from unfairly depriving team one of reward by echoing its bottom row as the top of its own feigned tiling. Furthermore, each team must have at least two members, since any team with only a single member would know all locations to be tiled at one stage of the game, and could give answers that appeared consistent even though they did not in fact know of any such tiling.

This last feature is especially key. As the authors note, their proofs leave open two key questions, namely the complexity for competitive POSGs with less than four agents, where one team, or both, has only a single member. While these versions of the problem may well have the same NEXP$^{\text{NP}}$complexity as the others, it is an open possibility that the complexity is somewhat less, as it does not seem possible to construct a reduction of the form used by (Goldsmith and Mundhenk 2008) that permits single-agent teams.

New Results

As already discussed, the question of all-out competition involves a one-sided view of rational game play, since it only asks if the first team in a two-team POSG has a response with positive expectation to literally any strategy employed by team two, including ones in which team two has no positive expectation of its own—and may even fare worse than team one. This is certainly an interesting question, and can tell us whether the first team is guaranteed some positive value in a game. At the same time, it is not the only way in which to understand rational competition, since it presupposes no necessary self-interest on the part of the second team of players. We thus propose another competitive question, at least as interesting as the first, which we call the ***self-preserving competitive question***: does team one in the POSG have a policy with positive expectation *under any circumstances*, and is that policy a positive response to every policy of the second team that *also* has positive expectation? That is, for a POSG with n agents, divided into one team of $k < n$ agents, $(1, \ldots, k)$ and a second team of $(n - k)$ agents $(k + 1, \ldots, n)$, we ask if is it true that:

$$\begin{aligned}\exists \pi_1, \ldots, \pi_k,\ [\exists \pi_{k+1}, \ldots, \pi_n, \forall i \leq k,\ EV(\Pi) > 0)] \wedge \\ [\forall \pi'_{k+1}, \ldots, \pi'_n, \forall j > k,\ EV_j(\Pi') > 0 \rightarrow \\ \forall i \leq k,\ EV_i(\Pi) > 0], \quad (3)\end{aligned}$$

where joint policy $\Pi = (\pi_1, \ldots, \pi_k, \pi_{k+1}, \ldots, \pi_n)$, and joint policy $\Pi' = (\pi_1, \ldots, \pi_k, \pi'_{k+1}, \ldots, \pi'_n)$ in each case.

For this hybrid question, then, a positive answer means not only that team one can achieve a positive result in the POSG, under some response by team two, but that team one can guarantee such a result so long as their opponent is also trying to achieve positive utility. Conversely, a negative answer means that the first team can not expect a positive result: either the game is simply a no-win situation for them, or their opponents have some self-preserving strategy that guarantees team one non-positive results. Under this understanding of rational competition, we have the following main result, which we then go on to prove in two separate results, as is typically the case:

Theorem 1. For any POSG with $n \geq 2$ agents, and any first team of size k, $1 \leq k < n$, deciding the self-preserving competitive question (Eq. 3) is NEXP$^{\text{NP}}$-complete.

Upper Bounds

We begin by establishing that problem is decidable given the required bounds. To do so, we will use the following:

Claim 1. Let $\mathcal{G}$ be POSG with $n \geq 2$ agents, divided into one team of $k < n$ agents and a second team of $(n - k)$ agents. Let $\mathbf{\Pi_S}$ be the set of all *stationary joint policies* for $\mathcal{G}$; that is, any $(\pi_1, \ldots, \pi_n) \in \mathbf{\Pi_S}$ is such that every individual policy is a function from individual observations to actions: $\forall i \in I,\ \pi_i : \Omega_i \rightarrow A_i$. Then the question whether any such policy with positive performance for team two also has positive performance for team one, namely:

$$(\forall j > k,\ EV_j(\Pi) > 0) \rightarrow (\forall i \leq k,\ EV_i(\Pi) > 0),$$

for every $\Pi \in \mathbf{\Pi_S}$, is in the class co-NP.

Proof. This result is merely a modification of (Goldsmith and Mundhenk 2008), Corollary 3.3, and can be proven in essentially the same way. To show that the problem stated is in co-NP, we must show that its negation is in NP; that is, we can verify the existence of some $\Pi \in \mathbf{\Pi_S}$ such that

$$(\forall j > k,\ EV_j(\Pi) > 0) \wedge (\exists i \leq k,\ EV_i(\Pi) \leq 0),$$

using only nondeterministic polynomial time. This is straightforward, since all we need to do is guess some such stationary joint policy, and then evaluate it. Since the policies under consideration are stationary, writing them out can take no more space than it takes to represent $\mathcal{G}$ itself—in fact, as Theorem 3.1 in the cited paper points out, it takes no more space than the tabular representation of transition-function P. Evaluating the fixed set of policies in the POSG is then straightforward to perform in polynomial time (recall that throughout this work, all policies of concern are of length no more than $|G|$). □

We now use this result in the proof of upper bounds. Since our competitive question is a conjunction of two different claims, this will involve a somewhat more involved proof that prior results, but in any case we show that nondeterministic exponential time, along with a co-NP oracle, is sufficient to decide both conjuncts.

Lemma 1. For any POSG with $n \geq 2$ agents, and any first team of size k, $1 \leq k < n$, the self-preserving competitive question (Eq. 3) is in NEXP$^{\text{NP}}$.

Proof. To establish an answer to the self-preserving competitive question, we must ascertain first whether or not a joint policy with positive expectation for the first team of k agents exists. This phase of the problem is in all essentials identical to the cooperative Dec-POMDP question; to solve it, we guess policies for both teams and proceed in the manner of (Bernstein et al. 2002). That is, we convert the POSG and joint policy for all n agents into a single agent POMDP, and then into a belief-state MDP, before verifying its value, in total time exponential in the size of the original POSG (since the equivalent single-agent models, that incorporate that policy directly, are exponentially larger). The only effective difference between that original proof is that the reward function produces a tuple of rewards, one for each of the k agents in team one, and the verification stage checks not merely a single expected value, but that the expectation for every element of the reward-tuple is positive. If the expectation for any of the k elements is negative, we reject; otherwise, we move on to the next phase.

In the second part of our algorithm, we must verify that the policy guessed for team one has positive expectation under any response under which team two also has positive expectation. Here, we proceed analogously to (Goldsmith and Mundhenk 2008). That is, we construct a new POSG that takes the guessed policy for team one and encodes it into the state-transitions, again using exponential time, as the domain grows exponentially larger. (Full details can be found in Lemma 3.3 of the cited paper.) In our version of the construction however, each of the $(n-k)$ members of team two, who act in the new domain, retain their original reward functions. In turn the reward functions for the k members of team one are shifted to a new team of k agents, each of which has a single available action that has no effect on the state transitions at all. In this fashion, the values accrued by each agent under stationary policies of team two are as in the original POSG. Finally, we can query an oracle whether every such stationary policy that has positive value for team two also has positive value for the agents receiving team one's rewards (this is the role of Claim 1), accepting if the answer is yes and rejecting otherwise. □

It is worth emphasizing that the total time required for the combined algorithm of the prior proof, which checks both conjuncts in Equation 3, is still exponential in the size of the original POSG $\mathcal{G}$. Although it takes longer than either of the original algorithms on which it is based, it is simply a sum of two exponential-time sets of operations.

Lower Bounds and Completeness

We now show that our deciding the question of self-preserving competition actually requires NEXP^{NP}resources, establishing tight bounds on our problem.

Lemma 2. For any POSG with $n \geq 2$ agents, and any first team of size k, $1 \leq k < n$, the self-preserving competitive question (Eq. 3) is NEXP^{NP}-hard.

Proof. We show hardness—and indeed completeness—by reduction from the NEXP^{NP}-complete Σ_2 tiling problem. As already discussed, this problem asks, for a given tiling instance, whether (a) some valid tiling of the square exists, and (b) whether there is such a tiling for which the bottom row never appears as the top row of any valid tiling (same or otherwise). We show how to reduce any such problem instance to a POSG with two teams, with a single agent each, for which the answer to the self-preserving competitive question is yes if and only if a tiling of the desired type exists. Doing so establishes that any POSG with larger teams is similarly hard (in fact, one could simply show hardness for any such case directly, since the reduction could always add members to each team whose actions did nothing, and whose rewards were identical to their active team-mates).

The full details of such a reduction are quite complex, especially the binary encoding of tile locations into the POSG problem domain, and the precise specification of the transition and observation functions; for examples of proofs worked out in all the gory details, see (Bernstein et al. 2002; Bernstein 2005; Allen 2009; Allen and Zilberstein 2009). To aid in exposition, Figure 2 gives an overview of the problem instance produced in the reduction. The POSG begins at the position marked ***Start*** and then proceeds through a number of stages as follows:

Query (01). A single (x, y) location in the tiling grid is revealed to the player on each team. As is usual in these reductions, each player receives the location bit-by-bit (this ensures that the size of the overall state-space of the reduced problem is polynomial in that of the original). Also as usual, neither player observes the location given to the other, and locations are chosen stochastically from all possible grid-squares, so no player can cheat the process.

Choose (01). Each player chooses a tile type to be placed at its location. Again, the system ensures that players do not cheat, while maintaining a state-space of permissible size, by ensuring that each player repeats back the location at which they will place the tile.

Consistency Check. The system checks whether the chosen pair of tiles is consistent or not. That is, if the locations given to each team were the same, the tiles must be the same; if they are adjacent horizontally or vertically, then they must be valid according to the relevant tiling constraints; if they are non-identical and non-adjacent, then any choice at all is considered consistent. Again, the system can verify these properties in the requisite amount of space, and collusion and cheating are prevented by the requirement that agents truthfully report the locations they were first given (a fixed number of bits of each location are recorded by the system ahead of time for verification purposes).

Reward Phase. If the tiles chosen by the agents do not pass the consistency check, then each team receives reward of -1 (all actions up to this point have reward of 0), and the POSG terminates in the absorbing ***End*** state. If the tiles are consistent, each team receives reward of $+1$, and continues.

Query (02). A second set of grid locations are revealed to each team, separately as before.

Choose (02). Each team again chooses a tile type for its location, and repeats its location back.

Consistency Check. The chosen pair of tiles is subjected to the same consistency check as before.

(Possible) Reward Phase. As before, if the agents fail the consistency check, then each team receives -1 reward (for a net of 0 units accumulated each), and the process terminates; if they pass the consistency check, the process continues without reward for the moment.

Check for Top-Bottom Matches. A second check is made for the two most recently chosen tiles. These tiles are said to comprise a *top-bottom match* if (a) one is in the top row of the tiling grid, and the other is in the bottom row, (b) the

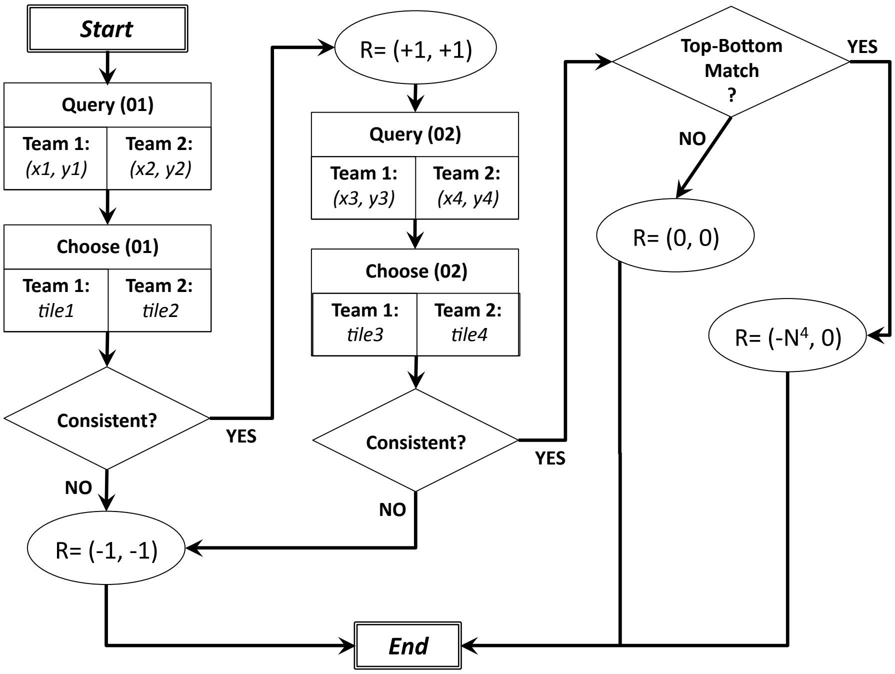

Figure 2: A two-team POSG corresponding to an instance of the Σ_2 tiling problem.

columns of each is the same, and (c) the tiles are identical. If any of these conditions fail to hold (including when both tiles are in the same row), then no such match exists.

Reward Phase. If there was no top-bottom match, then each team receives 0 reward (for a net of $+1$ units accumulated each). If such a match is found, then the first team receives a penalty of $-N^4$ (where N is the size of the tiling grid), and the second team receives 0 reward. In either case, the POSG terminates and the process is over.

We can now argue that the resulting POSG satisfies the self-interested competition condition if and only if the tiling problem with which we begin has a valid tiling with no bottom row that appears as a top row in any such tiling. Suppose that such a tiling exists. Then a policy exists under which team one chooses tiles according to that tiling when queried, and this policy has positive expected reward in those cases in which team two does the same. Furthermore, the only possible responses for which team two can expect positive rewards also choose tiles validly, and in the resulting joint policy, chosen tiles will pass both consistency checks, yielding positive guaranteed reward, and the fact that the bottom and top rows of the tiling must be distinct from one another means that the top-bottom matching penalty will not reduce expectation below 0.

If such a tiling does not exist, then there are two possibilities: either no valid tiling of the grid is possible at all, or any valid tiling has a bottom row that is identical to the top row of some valid tiling. In the first case, any joint policy in the resulting POSG will have negative expectation, since there is no way for agents to game the system by simply guessing tile types. This means that there is no policy under which the first team has positive expectation at all, no matter what team two does, and the first conjunct of Equation 3 fails. In the second case, since valid tilings do exist, but the bottom rows are repeated elsewhere as top rows. Thus, if team one chooses tiles validly, there are joint policies for which the second team has positive expectation, but the first team does not, and the second conjunct fails. (In these policies, team two will again tell the truth about valid tilings, but these tilings will have the same tiles in the bottom and top rows.) Too, if team one ignores the valid tiling, then they cannot expect positive reward at all. In either case, then, one conjunct of Equation 3 fails, showing that Σ_2 tiling reduces to self-preserving competition for POSGs. □

Conclusions and Future Work

We have shown that self-preserving competitive solutions to team-based partially observable stochastic games are NEXP$^{\text{NP}}$-complete. The problem of whether agents can expect positive value, when faced with opponents who also seek positive expected value is significantly more complex than the similar problem, in which agents actually work together. While previous work has clarified that under all but the most restrictive assumptions the cooperative (Dec-POMDP) version of the problem remains $NEXP$-hard, we see now that under common understandings of rational competition, the full POSG problem is harder yet, requiring not only nondeterministic exponential time, but access to NP-class oracles as well. As for the cooperative problem, of course, this is not the end of the story. While these results generally mean that optimal solution algorithms will be simply infeasible, much work has already gone into studying approximation techniques. Given that POSGs represent many real-world scenarios in which human and automated problem-solving is applied, there is still much to be gained from such studies.

In truth, this research began as an attempt to answer two open questions found in (Goldsmith and Mundhenk 2008): the complexity of "all-out" competition (Eq. 2) when (a) each team has only a single player (1 versus 1 play), and (b) the first team has a single player, but the second team has more than one (1 versus many play). While that work was able to show both problems to be in the class $\mathrm{NEXP}^{\mathrm{NP}}$, via a completely general upper-bound result, lower bounds (and completeness) are left open, since the existing reductions make intrinsic use of multiple players on each team, in order to provide proof against cheating via true decentralization. As is often the case, trying to prove one thing often leads another, as we discover what additional assumptions need to be made for a given form of proof to go through. In this case, we discovered that the additional requirements of self-preservation allowed a fully general complexity result; fortunately, this is interesting enough in its own right.

Still, the open questions remain. We are currently engaged in other ways of approaching the still open questions. If answered, these promise to fill in one remaining blank spot in what is now a rather complete framework of complexity results for stochastic decision problems, both single- and multi-agent, both cooperative and competitive. In this connection, we do state one conjecture, based on preliminary results: in stochastic games in which all agents share observations in common (whether fully observable or partially so), the 1 versus 1 and 1 versus many problems are in fact NEXP-complete. Whether this reduction in complexity (relative to $\mathrm{NEXP}^{\mathrm{NP}}$, anyhow) holds for those problems without the restriction on observations, or holds for the many versus many problem under the same restriction, is less certain.

References

Allen, M., and Zilberstein, S. 2009. Complexity of decentralized control: Special cases. In Bengio, Y.; Schuurmans, D.; Lafferty, J.; Williams, C. K. I.; and Culotta, A., eds., *Advances in Neural Information Processing Systems 22*, 19–27.

Allen, M. 2009. *Agent Interactions in Decentralized Environments*. Ph.D. Dissertation, University of Massachusetts, Amherst, Massachusetts.

Becker, R.; Zilberstein, S.; Lesser, V.; and Goldman, C. V. 2004. Solving transition independent decentralized MDPs. *Journal of Artificial Intelligence Research* 22:423–455.

Bernstein, D. S.; Givan, R.; Immerman, N.; and Zilberstein, S. 2002. The complexity of decentralized control of Markov decision processes. *Mathematics of Operations Research* 27(4):819–840.

Bernstein, D. S. 2005. *Complexity Analysis and Optimal Algorithms for Decentralized Decision Making*. Ph.D. Dissertation, University of Massachusetts, Amherst.

Condon, A. 1992. The complexity of stochastic games. *Information and Computation* 96(2):204–224.

Goldman, C. V., and Zilberstein, S. 2004. Decentralized control of cooperative systems: Categorization and complexity analysis. *Journal of Artificial Intelligence Research* 22:143–174.

Goldsmith, J., and Mundhenk, M. 2008. Competition adds complexity. In Platt, J.; Koller, D.; Singer, Y.; and Roweis, S., eds., *Advances in Neural Information Processing Systems 20*. Cambridge, MA: MIT Press. 561–568.

Hansen, E. A.; Bernstein, D. S.; and Zilberstein, S. 2004. Dynamic programming for partially observable stochastic games. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence*, 709–715.

Lewis, H. R. 1978. Complexity of solvable cases of the decision problem for predicate calculus. In *Proceedings of the Nineteenth Symposium on the Foundations of Computer Science*, 35–47.

Lusena, C.; Mundhenk, M.; and Goldsmith, J. 2001. Nonapproximability results for partially observable Markov decision processes. *Journal of Artificial Intelligence Research* 14:83–103.

Madani, O.; Hanks, S.; and Condon, A. 2003. On the undecidability of probabilistic planning and related stochastic optimization problems. *Artificial Intelligence* 147:5–34.

Mundhenk, M.; Goldsmith, J.; Lusena, C.; and Allender, E. 2000. Complexity results for finite-horizon Markov decision process problems. *JACM* 47(4):681–720.

Papadimitriou, C. H., and Tsitsiklis, J. 1987. The complexity of Markov decision processes. *Mathematics of Operations Research* 12(3):441–450.

Papadimitriou, C. H. 1994. *Computational Complexity*. Reading, Massachusetts: Addison-Wesley.

Savelsbergh, M. W., and van Emde Boas, P. 1984. Bounded tiling, an alternative to satisfiability? In Wechsung, G., ed., *Proceedings of the Frege Conference (1984, Schwerin)*, volume 20 of *Mathematical Research*, 354–363. Berlin: Akademie-Verlag.

Seuken, S., and Zilberstein, S. 2008. Formal models and algorithms for decentralized decision making under uncertainty. *Autonomous Agents and Multi-Agent Systems* 17(2):190–250.

Wang, H. 1961. Proving theorems by pattern recognition—II. *Bell System Technical Journal* 40(1):1–41.