Computing Optimal Strategies to Commit to in Stochastic Games

Joshua Letchford[1] and **Liam MacDermed**[2] and **Vincent Conitzer**[1] and **Ronald Parr**[1] and **Charles L. Isbell**[2]
[1]Duke University, Department of Computer Science, Durham, NC 27708, USA
{jcl,conitzer,parr}@cs.duke.edu
[2]Georgia Institute of Technology, Robotics and Intelligent Machines Laboratory, Atlanta, Georgia 30332
{liam,isbell}@cc.gatech.edu

Abstract

Significant progress has been made recently in the following two lines of research in the intersection of AI and game theory: (1) the computation of optimal strategies to commit to (Stackelberg strategies), and (2) the computation of correlated equilibria of stochastic games. In this paper, we unite these two lines of research by studying the computation of Stackelberg strategies in stochastic games. We provide theoretical results on the value of being able to commit and the value of being able to correlate, as well as complexity results about computing Stackelberg strategies in stochastic games. We then modify the QPACE algorithm (MacDermed et al. 2011) to compute Stackelberg strategies, and provide experimental results.

1 Introduction

Computing game-theoretic solutions is a topic that has long been of interest to AI researchers. A recent line of research focuses on two-player games in which player 1 is able to *commit* to a strategy before the other player moves. The following standard example illustrates the potential benefit of such commitment.

	L	R
U	(1,1)	(3,0)
D	(0,0)	(2,1)

Figure 1: Normal-form game used in Example 1.

Example 1 (known)**.** *Consider the normal-form game in Figure 1. For the case where the players move simultaneously (no ability to commit), the unique Nash equilibrium is* (U, L)*: U strictly dominates D, so that the game is solvable by iterated strict dominance. Player 1 (the row player) receives utility* 1*. However, now suppose that player 1 has the ability to commit. Then, she is better off committing to play D, which will incentivize player 2 to play R, resulting in a utility of* 2 *for player 1. The situation gets even better for player 1 if she can commit to a mixed strategy: in this case, she can commit to the mixed strategy* $(.5 - \epsilon, .5 + \epsilon)$*, which still incentivizes player 2 to play R, but now player 1 receives an expected utility of* $2.5 - \epsilon$*. To ensure the existence of optimal strategies, we assume (as is commonly done (Conitzer and Sandholm 2006; Paruchuri et al. 2008)) that player 2 breaks ties in player 1's favor, so that the optimal strategy for player 1 to commit to is* $(.5, .5)$*, resulting in a utility of* 2.5*. (Note that there is never a reason for player 2 to randomize, since he effectively faces a single-agent decision problem.)*

Besides potentially increasing the utility of player 1 (and never decreasing it (von Stengel and Zamir 2010)), the use of mixed Stackelberg strategies has several other technical advantages. First, it avoids the dreaded *equilibrium selection* problem: in simultaneous-move games, if the players choose their strategies from different equilibria, the result is not necessarily an equilibrium. Second, in two-player normal-form games, an optimal mixed Stackelberg strategy can be computed in polynomial time using linear programming (Conitzer and Sandholm 2006; von Stengel and Zamir 2010), whereas computing a Nash equilibrium is PPAD-complete (Daskalakis, Goldberg, and Papadimitriou 2006; Chen and Deng 2006), and if the goal is to find an *optimal* Nash equilibrium, typical objective functions (such as the sum of the players' utilities or even just player 1's utility) are NP-hard even to approximate (Gilboa and Zemel 1989; Conitzer and Sandholm 2008). (However, an optimal *correlated* equilibrium can be computed in polynomial time using linear programming.) Perhaps in part due to some of these advantages, the computation of mixed Stackelberg strategies has recently found application in various real security problems, including airport security (Jain et al. 2008; Pita et al. 2009), assigning Federal Air Marshals to flights (Tsai et al. 2009), and Coast Guard patrols (Shieh et al. 2012).

Most of the work on computing mixed Stackelberg strategies has focused on games where neither player learns anything about the other's actions until the end of the game, with the exception of player 2 learning player 1's mixed strategy before acting. (An exception is work on computing Stackelberg strategies in extensive-form games (Letchford and Conitzer 2010), on which we will draw later in the paper.) A useful language for describing games that play out over time is that of *stochastic games*, a generalization of MDPs to multiple players. Computing equilibria of stochastic games

S

	L	R
N	(0,1);E	(0,0);C

E

	N
N	(0,0);E

C

	L	R
U	(2,ϵ);E	(0,0);E
D	(0,0);E	(0,2);E

Figure 2: Example stochastic game.

S

	L	R
N	(0,2);E	(0,0);F

C

	N
U	(0,3);E
D	(0,0);E

F

	L	R
N	(2,1);E	(0,0);C

E

	N
N	(0,0);E

Figure 3: Example game where signaling must occur early, but not too early.

presents a number of challenges, but a recent sequence of papers makes significant progress on the problem of computing correlated equilibria that are not necessarily stationary (i.e., the players' actions may depend on the history of play, not just the current state). This line of research replaces the notion of value (the maximum utility) in traditional value iteration with an achievable set (the set of Pareto efficient maximal utilities for each player). Achievable sets represent all possible utility vectors yielded by correlated policies in equilibrium. Murray and Gordon (2007) presented the first exact algorithm for computing these achievable sets. However, the complexity of maintaining these achievable sets increased exponentially with each iteration, leading to an algorithm that where both time and space requirements scaled exponentially. MacDermed et al. (2011) showed how to epsilon-approximate achievable sets efficiently while simultaneously lowering computational complexity, leading to the Quick Polytope Approximation of all Correlated Equilibria algorithm (QPACE).

In this paper, we unite these two lines of work and focus on the problem of computing optimal strategies to commit to in stochastic games. The recent methods for computing correlated equilibria of stochastic games turn out to combine well with recent observations about the relationship between Stackelberg strategies and correlated equilibrium in normal-form games (Conitzer and Korzhyk 2011), although there are some additional subtleties in stochastic games, as we will see. The idea of commitment also combines well with the notion of "grim-trigger" strategies (strategies that aim to forever minimize another player's utility) in stochastic games, because player 1's commitment power will make it credible for her to play such a strategy.

2 Stochastic games

A two-player stochastic game is defined as follows: We have a two players, 1 and 2, a set of states T and in each state t, we have a set of actions for each player A_t. For each state t and each action pair in $A_t^1 \times A_t^2$, we have an outcome that consists of two elements, the utilities that each of the players will achieve in that round and information on what state the game will transition to next (possibly stochastically). Finally, we have a discount factor γ which is used to discount the value of future payoffs.

Consider the game in Figure 2. This game has three states: S, C and E. We assume state S is our initial state, meaning that play will begin with the possible actions A_S^1 for player 1 and A_S^2 for player 2 (throughout this paper play will always begin in the state labeled S). In this state player 1 has only 1 possible action; thus the outcome depends entirely on player 2. If player two chooses to play action L here (which we will denote L_S), then the outcome is (0,1);E. This means that player 1 receives a utility of 0, player 2 receives a utility of 1, and the game transitions to state E, meaning in the next round the two players will be playing the normal-form game depicted as state E. Alternatively, if player 2 chooses R_S, then both players receive a utility of 0 for this round, and play transitions to state C.

3 Commitment and signals

In a two-player Stackelberg game, player 1 may be able to do more than just committing to play a specific mixed strategy; she may also be able to send signals to the other player. This idea has previously been explored for normal-form games (Conitzer and Korzhyk 2011). Specifically, they consider the case where she can commit to drawing from a joint distribution over signals and her own actions, and then send the signal to player 2 before he moves, while playing her drawn action. (Without loss of generality, we can assume that the signal player 1 sends to player 2 is simply the action that he should take.) They show that in a two-player normal-form game, player 1 gains nothing from the ability to commit to such a correlated strategy rather than just a mixed strategy (that does not signal anything to the other player). The reason is that for each signal, there will be a distribution over player 1's actions conditional on that signal—and player 1 may as well just commit to playing the best of those distributions from her perspective.

It turns out that in two-player stochastic games, signaling becomes more meaningful. First, let us return to the game in Figure 2. If we assume $\gamma = 1$, if player 1 commits to with probability $(.5 - \epsilon)$ send signal L_C and play U_C and commits to with probability $(.5 + \epsilon)$ send signal R_C and play D_C, then player 2 can expect a utility of 1+2.5ϵ - ϵ^2 for taking the action R_S. Without signaling, if player 1 commits to U_C less than $\frac{2}{2+\epsilon}$ of the time, player 2 will respond with R_C which leads to a utility of 0 for player 1. Furthermore, if player 1 commits to U_C at at least $\frac{2}{2+\epsilon}$ of the time, player 2 will always prefer L_S which again gives a utility of 0 to player 1.

However, being able to signal about one's action at a state

only when reaching that state may not be enough in every game. Consider the game pictured in Figure 3. To achieve a positive utility in this game, player 1 needs to signal what she will be playing in state C before player 2 chooses his action in state F but after he acts in state S. Consider what happens if player 1 commits to $(.5+\epsilon)U_C$ + $(.5-\epsilon)D_C$. If she commits to sending the following signal to player 2 in state F: R_F when she will be playing U_C and L_F when she will be playing D_C, it is possible for the two players to achieve a correlated equilibrium of $(.5+\epsilon)(R_F, U_C)$ + $(.5-\epsilon)L_F$. Given this, player 2 would prefer R_S to L_S as he would achieve an expected utility of $2+2\epsilon$ for R_S. In contrast, if player 1 only sends the signal after the transition to state C, then player 2 will prefer to play R_F(and L_S). On the other hand, the information signaled informs player 2 of what action player 1 will play in state C and if this information is signaled too early, namely before player 2 makes a choice in S, then he would prefer to choose L_S when the signal is to play L_F. Thus, without the ability to signal this information at this specific time, this correlation would not be possible and player 1 would achieve a utility of 0.

4 Value of correlation and commitment

The game in Figure 3 illustrates a situation where player 1 would be unable to achieve a positive utility without both the ability to commit to a mixed strategy and the ability to correlate her actions with those of player 2. In this game commitment and correlation work synergistically. In this section we show it is not always so: in some games commitment by itself obtains all the value whereas correlation by itself obtains none, and in other games the situation is reversed.

Towards this end, we define the following three values. First, we define OptCom as player 1's utility in the optimal commitment strategy[1] that does not use correlation. Second, we define OptCor as player 1's utility in the correlated equilibrium that maximizes payoff for player 1. Finally, we define Opt as player 1's utility in the optimal commitment strategy that uses correlation.

S

	L	R
U	(ϵ,1);E	(1,0);E
D	(0,0);E	(1-ϵ,1);E

E

	N
N	(0,0);E

Figure 4: Game where commitment offers an unbounded advantage over correlation.

Theorem 1. *There exists a stochastic game where* $\frac{OptCom}{OptCor} = \infty$ *and* $OptCom = Opt$ *for any discount factor.*

Proof. Consider the game in Figure 4.[2] This is effectively a one-shot game because of the immediate transition to an absorbing state. To calculate OptCom and OptCor for this game, we can reason as follows. In the normal-form game associated with S, U_S dominates D_S, which allows us to solve this game by iterated strict dominance. Thus the only correlated equilibrium is (U_S,L_S) which gives a utility of ϵ to player 1. However, if player 1 is able to commit to playing D_S with probability 1, then player 2's best response to this is to play R_S (both with and without correlation). This causes player 1 to receive a utility of $1-\epsilon$. Thus, $\frac{OptCom}{OptCor} = \frac{1-\epsilon}{\epsilon}$ which tends to ∞ as epsilon approaches 0 and $OptCom = Opt$.

□

[1] We use the standard definition of optimal here, where a commitment strategy is considered to be optimal if it maximizes the utility for player 1.

[2] We have used the stage game for state S elsewhere to point out that the value of commitment in normal-form games is ∞ (Letchford, Korzhyk, and Conitzer 2012).

S

	L	R
U	(0,0);C	(0,.5);E
D	(2ϵ,0);E	(ϵ,ϵ);E

E

	N
N	(0,0);E

C

	L	R
U	(0,0);E	($\frac{1}{\gamma}$,$\frac{\epsilon}{\gamma}$);E
D	($\frac{\epsilon}{\gamma}$,$\frac{1}{\gamma}$);E	(0,0);E

Figure 5: Game where correlation offers an unbounded advantage over commitment.

Theorem 2. *For any* $\gamma > 0$ *there exists a stochastic game where* $\frac{OptCor}{OptCom} = \infty$ *and* $OptCor = Opt$.

Proof. Consider the game in Figure 5. To calculate OptCor and OptCom for this game, let us start by finding the optimal Stackelberg strategy without correlation. Let us first consider what is necessary to convince player 2 to play L_S. Even if player 1 commits to U_S, player 2 will best respond by playing R_S (which gives player 1 a utility of 0) unless player 1 commits to D_C with probability at least .5. If player 1 does commit to D_C with probability at least .5, then player 2's best response for C will be L_C and at best player 1 can expect at ϵ utility from C. Thus, there is no way for player 1 to achieve more than 2ϵ by convincing player 2 to best respond with L_S. As both outcomes where player 2 plays R_S transition to the absorbing state E and give player 1 at most ϵ utility, we can conclude that player 1 will achieve at most 2ϵ via commitment without correlation.

Next, consider the following correlated equilibrium, which involves: (U_S,L_S), .5(U_C,R_C) + .5(D_C,L_C). This gives a discounted expected value of $.5+.5\epsilon$ for both players for state C, causing player 1 to prefer U_S to D_S and player 2 to prefer L_S to R_S. Player 1's utility under this equilibrium is $.5+.5\epsilon$. Thus, $\frac{OptCor}{OptCom} \geq \frac{.5+.5\epsilon}{2\epsilon}$ which tends to ∞ as ϵ goes to 0. □

5 Hardness results

In this section we consider the difficulty of solving for an optimal Stackelberg strategy. We consider two main dimensions of the problem, the amount of memory about past

	$h = 0$	$0 < h < \infty$	$h = \infty$
Corr.	NP-hard (Th 3)	NP-hard (Th 4)	?[3]
No Corr.	NP-hard (Th 3)	NP-hard (Th 4)	NP-hard (Th 5)

Figure 6: Overview of hardness results. h represents the amount of history that player 1 can remember.

states and actions player 1 can base her actions upon and the ability of player 1 to signal to the follower to enable correlation. For the memory dimension we consider three main cases. In the first case player 1 is constrained to commit to a stationary strategy. In the second case player 1 has some finite memory and can commit to act based upon the states and actions taken in these past timesteps. Finally, in the third case player 1 has an infinite memory, and can commit based on all actions and states that have occurred since the start of the game. For the signaling dimension we consider two cases, both with and without the ability to signal. An overview of our results appears in Figure 6.

Theorem 3. *It is NP-hard to solve for the optimal commitment to a stationary strategy in a stochastic game with or without correlation, for any discount factor $\gamma > 0$.*

Proof. We reduce an arbitrary instance of 3SAT to a stochastic game such that player 1 can obtain utility 1 if and only if the 3SAT instance is satisfiable. The 3SAT instance consists of N variables $x_i \dots x_n$ and M clauses $C_1 \dots C_m$. The construction is pictured in Figure 7 and the details are as follows. We start with an initial state S, one state C_i for each clause, one state x_i for each variable and one final absorbing state E. Our initial state S has no payoff for either player, but transitions uniformly at random to a clause state C_i. Additionally, E, the absorbing state has a single possible outcome, namely (0,0);E.

Clause states: Each clause state is constructed as follows. For each literal we have one row (C_i^x) and two columns (C_i^{+x} and C_i^{-x}). If the literal is positive we have two entries in the game, (C_i^x,C_i^{+x}) is assigned an outcome of ($\frac{1}{\gamma}$,0);x and (C_i^x,C_i^{-x}) is assigned an outcome of (0,1);E. If the literal is instead negative, we include the following two entries, (C_i^x,C_i^{+x}) is assigned an outcome of ($\frac{1}{\gamma}$,0);E and (C_i^x,C_i^{-x}) is assigned an outcome of (0,0);x. The other 12 outcomes of the game are (0,0);E. In Figure 7 we show an example for a clause with the literals $(x_1 \vee \neg x_2 \vee x_3)$.

Variable states: Each variable state has 1 column and two rows $+x_i$ and $-x_i$. The outcomes are as follows, for row $+x_i$ it is (0,$\frac{1}{\gamma}$);E and for row $-x_i$ it is (0,0);E.

Proof of equivalence to 3SAT instance: We now show that player 1 can obtain a utility of 1 from this game if and only if there exists a satisfying assignment to the underlying 3SAT instance. Let us start by considering when player 1 can obtain a utility of 1 from a given clause state. We first consider a row that corresponds to a positive literal. In this case, if player 1 has also committed to $+x$ in the corresponding variable state, then the two non-zero outcomes in this row give payoffs of ($\frac{1}{\gamma}$,1) and (0,1). Since, by assumption, the follower breaks ties in player 1's favor, a signal to play C^{+x} (this could be either committing to play C^x or explicitly signaling this to player 2) will lead to a utility of 1 for player 1. If player 1 instead commits to any other strategy in x then a signal to play C^{+x} will instead lead to player 2 deviating to C^{-x} causing player 1 to receive 0 utility. Next, consider a row corresponding to a negative literal. In this case, if player 1 has committed to $-x$ in the corresponding variable state, then the two potentially non-zero outcomes in this row gives payoffs of (0,0) and ($\frac{1}{\gamma}$,0). With similar logic as before, a signal to play C^{-x} will lead to a utility of 0 for player 1 unless she has committed to $-x$. As there are three literals in each clause, this gives player 1 three potential ways to incentivize the follower to play in a way that is beneficial to player 1. However, later commitment (in the variable states) can remove this potential (namely commitment in such a way to preserve the potential for the opposing literal). If all three of these signals lose their potential, player 1 is left with no way to incentivize the follower to play in a way that gives her utility if play reaches this clause. Since the initial state forces a uniform random choice between these clauses subgames, this game will have expected value for player 1 of 1 if and only if all of the clause states have a utility of 1. □

Theorem 4. *It is NP-hard to solve for the optimal commitment to a strategy that uses a constant h steps of history with or without correlation and any discount factor $\gamma > 0$.*

Proof. Consider the following modification to the reduction used in the proof of Theorem 3. For each variable state, we isnert h buffer states that give no payoffs before that state. Thus, by the time player 1 reaches the variable state, they will have forgotten which clause they originated from and the above reduction will again hold. □

For the case of infinite history, it is impossible to extend the above 3SAT reduction. Consider the construction from Theorem 3 with h buffer states inserted for each variable. If player 1 has a memory of $h + 1$, when choosing a literal to commit to, player 1 can condition this upon the clause the players transitioned from. In this way, player 1 will be able to "satisfy" both the positive and negative values of each literal.

However, we note that stochastic games can model extensive-form games with chance nodes, which allows us to adapt the KNAPSACK reduction from Theorem 5 in (Letchford and Conitzer 2010) with minor changes.

Theorem 5. *It is NP-hard to solve for the optimal commitment in a stochastic game even when the strategy is allowed to use infinite history without correlation and any discount factor $\gamma > 0$.*

Proof. In the KNAPSACK problem, we are given a set of N items, and for each of them, a value p_i and a weight w_i; additionally, we are given a weight limit W. We are asked to find a subset of the items with total weight at most W that maximizes the sum of the p_i in the subset. We reduce an arbitrary KNAPSACK instance to a stochastic game, in

[3] We show how to solve this approximately by a modification of the QPACE algorithm in Section 6.

S

	N
N	(0,0);$\frac{1}{M}C_1 + ... + \frac{1}{M}C_m$

x_i

	N
$+x_i$	(0,$\frac{1}{\gamma}$);E
$-x_i$	(0,0);E

E

	N
N	(0,0);E

C_i $(x_1 \vee \neg x_2 \vee x_3)$

	$C_i^{+x_1}$	$C_i^{-x_1}$	$C_i^{+x_2}$	$C_i^{-x_2}$	$C_i^{+x_3}$	$C_i^{-x_3}$
$C_i^{x_1}$	($\frac{1}{\gamma}$,0);x_1	(0,1);E	(0,0);E	(0,0);E	(0,0);E	(0,0);E
$C_i^{x_2}$	(0,0);E	(0,0);E	($\frac{1}{\gamma}$,0);E	(0,0);x_2	(0,0);E	(0,0);E
$C_i^{x_3}$	(0,0);E	(0,0);E	(0,0);E	(0,0);E	($\frac{1}{\gamma}$,0);x_3	(0,1);E

Figure 7: Stochastic game used in the hardness reduction of Theorem 3.

such a way that the maximal utility obtainable by player 1 with commitment (whether pure or mixed) is equal to the optimal solution value in the KNAPSACK instance. This game is illustrated in Figure 8, and defined formally below.

Initial state S: The first state contains two possible choices by player 2, who chooses between an outcome of $(0, -W); E$ and a outcome which randomizes uniformly over the item states, defined next.

Item states: Each item I_i has two states. At the top level I_i^1 (which can be reached directly from S), there is a state where player 2 acts. It has two potential outcomes: one is an outcome of $(\frac{Np_i}{\gamma}, \frac{-Nw_i}{\gamma}); E$, the other is a transition to a state I_i^2 where only player 1 has a choice. The latter node also has two outcomes: $(0, \frac{-Nw_i}{\gamma^2}); E$ and $(0,0); E$.

Proof of equivalence to KNAPSACK instance: If for an item i, player 1 commits to playing 100% $\text{In}_{I_i^2}$, then player 2, breaking ties in 1's favor, will move $\text{In}_{I_i^1}$, resulting in discounted payoffs of $(Np_i, -Nw_i)$ if I_i^1 is reached. Otherwise, player 2 will move $\text{Out}_{I_i^1}$, and player 1 will get 0 (and player 2 at most 0). Because player 1 wants player 2 to choose K_S, there is no benefit to player 1 in moving $\text{In}_{I_i^2}$ with probability strictly between 0 and 100%, since this will only make K_S less desirable to player 2 without benefiting player 1. Thus, we can assume without loss of optimality that player 1 commits to a pure strategy.

Let X be the set of indices of states where player 1 commits to playing In. Then, player 2's expected utility for choosing K_S is $(1/N)\sum_{i \in X} -Nw_i = -\sum_{i \in X} w_i$. Player 2 will choose K_S if and only if $\sum_{i \in X} w_i \leq W$. Given this, player 1's expected utility is $(1/N)\sum_{i \in X} Np_i = \sum_{i \in X} p_i$. Hence, finding player 1's optimal strategy to commit to is equivalent to solving the KNAPSACK instance. □

6 Empirical Results

While we have shown that the ability to commit can, in the extreme, provide an unbounded increase in utility, these results say little about the value of committing in general. We present the first algorithm which can compute all correlated commitment equilibria of a stochastic game. We use this

S

	K	A
N	$\frac{1}{N}I_1^1 + ... + \frac{1}{N}I_n^1$	(0,$-W$);E

I_i^2

	N
In	(0,$\frac{-Nw_i}{\gamma^2}$);E
Out	(0,0);E

I_i^1

	In	Out
N	($\frac{Np_i}{\gamma}$,$\frac{-Nw_i}{\gamma}$);E	(0,0);I_i^2

E

	N
N	(0,0);E

Figure 8: Stochastic game used in the hardness reduction of Theorem 5.

algorithm to compare a leader's value using correlated commitment to her value using only correlation, and make conclusions about the conditions under which commitment is most important.

6.1 Computing Commitment Equilibria

The QPACE algorithm (MacDermed et al. 2011) efficiently approximates the set of correlated equilibria in stochastic games by iteratively contracting a state's achievable set, by removing policies that violate a player's rationality constraints. Conitzer and Korzhyk (2011) showed that the set of commitment equilibria is equivalent to the set of correlated equilibria without the leader's rationality constraints. Therefore QPACE can be easily modified to approximately compute commitment equilibria by removing the leader rationality constraints in the achievable set contraction step.

Every iteration of QPACE performs a backup similar to single agent value iteration which improves the current estimation of each state's achievable set $V(s)$. Achievable sets are represented as polytopes with halfspace normals H_j and offsets $V(s)_j$. The normals H are fixed at initialization and are the same for all achievable sets. A state's backup is broken down into three steps: (1) Calculate the action achievable sets $Q(s, \vec{a})$, giving us the set of possible continuation utilities. (2) Construct a set of inequalities that defines the set of equilibria. (3) Approximately project this feasible set into value-vector space by solving a linear program for each hyperplane $V(s)_j$. Step two is the only step that needs to be changed to compute feasible commitment policies instead of correlated equilibria. We modify equation 6 in MacDermed

et al. (2011) to not include rationality constraints for the leader. The resulting set of inequalities defining the set of commitment equilibria is shown in equation 1. This gives us our polytope in $\mathbb{R}^{(n+1)|A|^n}$ over variables $\overrightarrow{cu_{\vec{a}i}}$ and $x_{\vec{a}}$ of feasible correlated equilibria:

$$\begin{gathered}\text{For each player } i \text{ who cannot commit,} \\ \text{for each pair of distinct actions } \alpha, \beta \in A_i \\ \sum_{\vec{a} \in A^n} \overrightarrow{cu_{\vec{a}^{(\alpha)}i}} \geq \sum_{\vec{a} \in A^n} x_{\vec{a}^{(\alpha)}} [\overrightarrow{gt_{\vec{a}^{(\beta)}i}} + R(s, \vec{a}^{(\beta)})_i] \\ \sum\nolimits_{\vec{a} \in A^n} x_{\vec{a}} = 1 \text{ and } \forall \vec{a} \in A^n,\ x_{\vec{a}} \geq 0 \\ \text{For each joint-action } \vec{a} \in A^n \text{ and halfspace } j \\ H_j\ \overrightarrow{cu_{\vec{a}}} \leq x_{\vec{a}}\ Q(s, \vec{a})_j \end{gathered} \tag{1}$$

The rest of QPACE remains unchanged. Our modification to QPACE is minor and leaves the strong theoretical properties of the original algorithm intact. Most importantly, the algorithm converges to within ϵ in polynomial time and returns a set of commitment equilibria which is guaranteed to include all exact equilibria with additional solutions being no worse than ϵ-equilibria, where ϵ is the approximation parameter.

6.2 Experiments on random games

We ran suites of experiments over sets of random games. These random games varied over the number of states, actions, stochasticity (the number of successor states with non-zero transition probability), and discount (γ). Unless otherwise noted, games were run with four joint-actions, five states, two successor states, a γ of 0.9 and an ϵ approximation error of 0.01. Results are averaged over 1000 games, which allows our utility results to be accurate within 0.02 with 99% confidence. A random game is generated using the following procedure: for each state joint-action pair k successor states are chosen at random. The simplex over these k states represents all possible probability distributions over these states. A transition probability distribution is chosen uniformly at random from this simplex. Each state joint-action pair is also assigned a reward for each player drawn uniformally at random. Finally, these rewards are normalized such that each player's rewards range between 0 and 1.

Our first set of experiments examines the scalability of the algorithm. Despite our algorithm having fewer constraints for each linear program than the original QPACE algorithm, we found our algorithm to have running time nearly identical to the original. Because QPACE starts each linear program at the solution of the previous iteration's linear program, the total number of basis changes over the course of the entire algorithm is relatively small. Thus, fewer constraints reduces the overall running time by an insignificant amount. Our algorithm appears to scale linearly in the number of states, joint-actions, and $1/\epsilon$ (the first two of these are shown in Figure 9).

Our second set of experiments focuses on determining the importance of commitment vs. equilibrium selection. One

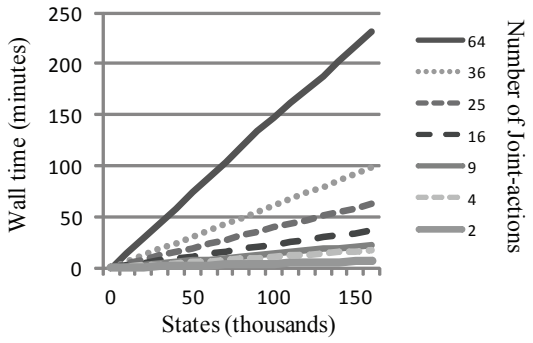

Figure 9: The running time of our algorithm is linear in the number of states and joint-actions.

of the more powerful aspects of committing is being able to dictate which particular equilibrium of the many possible will be chosen. Without commitment, players are faced with a bargaining problem to determine the which equilibrium will be chosen. This may result in significantly less utility for a potential leader. It is important to differentiate between the benefit of being able to commit to sub-rational policies and the benefit of equilibrium selection. Towards this end we compute both the Kalai-Smorodinsky bargaining solution (Kalai and Smorodinsky 1975), which favors equal gains to all parties, and the optimal selfish equilibrium for the potential leader with and without commitment (Figure 10).

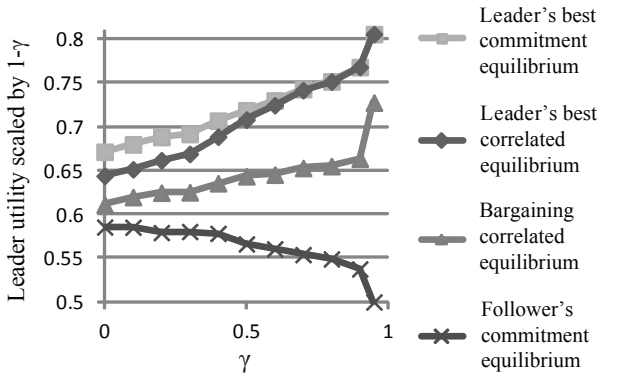

Figure 10: The average utility of being able to select a correlated or commitment equilibrium selfishly as opposed to the leader's Kalai-Smorodinsky bargaining solution or the follower's utility when the leader selected a commitment equilibrium.

The results show that as γ increases, the importance of committing, over and above just being able to select the equilibrium, decreases. This relationship is caused by threats becoming more powerful as the horizon increases. Strong threats act as a binding mechanism, permitting a wider array of possible equilibria by allowing players to punish each other for deviations without the need of someone violating her rationality constraint (as per folk-theorems). On the other hand, the benefit of equilibrium selection remains important regardless of the discount factor.

The effect of equilibrium selection on the follower is even more dramatic than it is for the leader (Figure 10). With a small discount factor, the set of possible equilibria is small and thus likely to provide both players with similar utilities, even when the leader selects selfishly. As γ increases, a leader has more options for forcing the follower down a path more preferable to the leader at the expense of the follower.

Our third set of experiments examines how the number of actions affects the value of committing. We tested random games with a γ of both 0.0 and 0.4 across varying numbers of actions per player (Figure 11). We observe that as the number of actions increases, the relative commitment gain decreases slightly because additional actions increase the probability that a correlated equilibrium without commitment will approach the unrestricted optimum, leaving less room for improvement by committing. While more actions decreases the importance of committing, the importance of being able to select the equilibrium (as opposed to having to bargain) remains high. This effect is stronger for larger values of γ. For games with very high discount factors, increasing the number of actions tends to increase the total value of the game, but not the relative importance of committing.

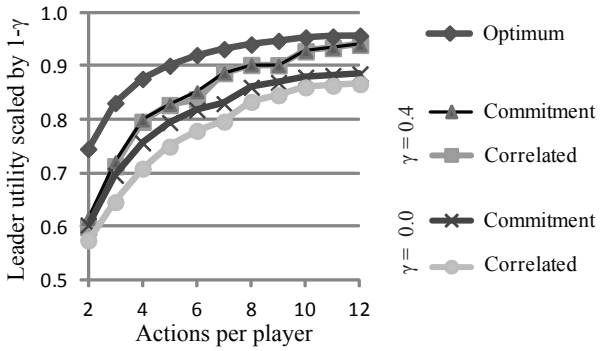

Figure 11: The best utility achievable by the leader using either correlated or commitment equilibria compared to the unrestricted optimum. Result are shown for $\gamma = 0.0$ and 0.4.

7 Conclusion

In this paper we showed how to adapt the QPACE algorithm to approximately solve for the optimal commitment solution in stochastic games. Additionally, we showed that removing the ability to correlate or the ability to use the full history of the game causes solving for the optimal commitment strategy to become NP-hard. Finally, we studied the value that both commitment and correlation provide to the leader, both theoretically and experimentally.

8 Acknowledgements

The authors would like to thank Dmytro Korzhyk for helpful discussions. Letchford and Conitzer gratefully acknowledge NSF Awards IIS-0812113, IIS-0953756, and CCF-1101659, as well as ARO MURI Grant W911NF-11-1-0332 and an Alfred P. Sloan fellowship, for support. MacDermed and Isbell gratefully acknowledge NSF Grant IIS-0644206 for support.

References

Chen, X., and Deng, X. 2006. Settling the complexity of two-player Nash equilibrium. In *FOCS*, 261–272.

Conitzer, V., and Korzhyk, D. 2011. Commitment to correlated strategies. In *AAAI*, 632–637.

Conitzer, V., and Sandholm, T. 2006. Computing the optimal strategy to commit to. In *EC*, 82–90.

Conitzer, V., and Sandholm, T. 2008. New complexity results about Nash equilibria. *Games and Economic Behavior* 63(2):621–641.

Daskalakis, C.; Goldberg, P.; and Papadimitriou, C. H. 2006. The complexity of computing a Nash equilibrium. In *STOC*, 71–78.

Gilboa, I., and Zemel, E. 1989. Nash and correlated equilibria: Some complexity considerations. *Games and Economic Behavior* 1:80–93.

Jain, M.; Pita, J.; Tambe, M.; Ordóñez, F.; Paruchuri, P.; and Kraus, S. 2008. Bayesian Stackelberg games and their application for security at Los Angeles International Airport. *SIGecom Exch.* 7(2):1–3.

Kalai, E., and Smorodinsky, M. 1975. Other solutions to nashs bargaining problem. *Econometrica* 43(3):513–518.

Letchford, J., and Conitzer, V. 2010. Computing optimal strategies to commit to in extensive-form games. In *EC*, 83–92.

Letchford, J.; Korzhyk, D.; and Conitzer, V. 2012. On the value of commitment. *Working Paper*.

MacDermed, L.; Narayan, K. S.; Isbell, C. L.; and Weiss, L. 2011. Quick polytope approximation of all correlated equilibria in stochastic games. In *AAAI*, 707-712.

Murray, C., and Gordon, G. 2007. Finding correlated equilibria in general sum stochastic games. Technical Report CMU-ML-07-113, Carnegie Mellon University.

Paruchuri, P.; Pearce, J. P.; Marecki, J.; Tambe, M.; Ordóñez, F.; and Kraus, S. 2008. Playing games for security: An efficient exact algorithm for solving Bayesian Stackelberg games. In *AAMAS*, 895–902.

Pita, J.; Jain, M.; Ordóñez, F.; Portway, C.; Tambe, M.; and Western, C. 2009. Using game theory for Los Angeles airport security. *AI Magazine* 30(1):43–57.

Shieh, E.; An, B.; Yang, R.; Tambe, M.; Baldwin, C.; DiRenzo, J.; Maule, B.; and Meyer, G. 2012. PROTECT: A deployed game theoretic system to protect the ports of the united states. In *AAMAS*.

Tsai, J.; Rathi, S.; Kiekintveld, C.; Ordonez, F.; and Tambe, M. 2009. IRIS - a tool for strategic security allocation in transportation networks. In *AAMAS - Industry Track*, 37–44.

von Stengel, B., and Zamir, S. 2010. Leadership games with convex strategy sets. *Games and Economic Behavior* 69:446–457.