Policy Improvement for Repeated Zero-Sum Games with Asymmetric Information

Malachi Jones and Jeff S. Shamma

***Abstract*—In a repeated zero-sum game, two players repeatedly play the same zero-sum game over several stages. We assume that while both players can observe the actions of the other, only one player knows the actual game, which was randomly selected from a set of possible games according to a known distribution. The dilemma faced by the informed player is how to trade off the short-term reward versus long-term consequence of exploiting information, since exploitation also risks revelation. Classic work by Aumann and Maschler derives the recursive value equation, which quantifies this tradeoff and derives a formula for optimal policies by the informed player. However, using this model for explicit computations can be computationally prohibitive as the number of game stages increases. In this paper, we derive a suboptimal policy based on the concept of policy improvement. The baseline policy is a non-revealing policy, i.e., one that completely ignores superior information. The improved policy, which is implemented in a receding horizon manner, strategizes for the current stage while assuming a non-revealing policy for future stages. We show that the improved policy can be computed by solving a linear program, and the computational complexity of this linear program is constant with respect to the length of the game. We derive bounds on the guaranteed performance of the improved policy and establish that the bounds are tight.**

I. Introduction

We consider an asymmetric zero-sum game, in which one player is informed about the true state of the world. This state determines the specific game that will be played repeatedly. The other player, the uninformed player, has uncertainty about the true state. Since both players can observe the actions of their opponent, the uninformed player can use his observations of the informed player's actions to estimate the true state of the world. Exploiting/revealing information to achieve a short-term reward risks helping the uninformed player better estimate the true state of the world. A better estimate of the true state can allow the uninformed player to make better decisions that will cost the informed player over the long term. A natural question is how should the informed player exploit his information.

Aumann and Maschler [1] introduced a recursive formulation that characterizes the optimal payoff the informed player can achieve, which is referred to as the value of the game. This formulation evaluates the tradeoff between short-term rewards and long-term costs for all possible decisions of the informed player, and the optimal decision is the decision that provides the best overall game payoff. Determining the optimal decision becomes increasingly difficult as the length of the game grows. Therefore, computing an optimal decision for games of non-trivial lengths can be computational prohibitive. This difficulty extends also to the simplest zero-sum games, which have two states and two possible actions for each player in a given state.

Much of the current work to address the information exploitation issue, which includes the work of Domansky and Kreps [2] and Heur [3], has been limited to finding optimal strategies for special cases of the simplest zero-sum games. Zamir provided a method to generate optimal strategies under certain conditions for 2x2 and 3x3 matrix games [4]. Gilpin and Sandholm [5] propose an algorithm for computing strategies in games by using a non-convex optimization formulation for only the infinitely repeated case. There is work that considers using suboptimal strategies to address the information exploitation issue for all classes of games [6]. In this work, the informed player never uses his information throughout the game, and accordingly is "non-revealing." While the suboptimal strategies are readily computable, only under special circumstances do these strategies offer strong suboptimal payoffs.

In this paper, after introducing basic repeated zero-sum game concepts and definitions, we introduce a suboptimal strategy that we refer to as the one-time policy improvement strategy. We show that the computational complexity of constructing this strategy is constant with respect to the length of game. Next, we provide tight bounds on the guaranteed performance of the constructed suboptimal policy. We then show that the policy improvement strategy can be computed by solving a linear programing problem. Finally, we present an illustrative simulation.

II. Zero-Sum Definitions and Concepts

In this section, we will introduce basic zero-sum repeated game definitions relevant to this paper. The first and most important concept is Aumann and Maschler's dynamic programming equation for evaluating the tradeoff between short-term and long-term payoff. We will then discuss the notion of non-revealing strategies, which will be exploited in our construction of suboptimal policies to reduce computational complexity.

This research was supported by ARO/MURI Project W911NF-09-1-0553.

M. Jones and J.S. Shamma are with the School of Electrical and Computer Engineering, College of Engineering, Georgia Institute of Technology {kye4u, shamma}@gatech.edu

A. Setup

1) Game Play: Two players repeatedly play a zero-sum matrix game over stages $m = 1, 2, ..., N$. The row player is the maximizer, and the column player is the minimizer. The specific game is selected from a finite set of possible games (or states of the world), K. Let $\Delta(L)$ denote the set of probability distributions over some finite set, L. Define S to be the set of pure actions of the row player, and similarly define J to be the set of pure actions of the column player. The game matrix at state $k \in K$ is denoted $M^k \in \mathbf{R}^{|S|\times|J|}$. Before stage $m = 1$, nature selects the specific game according to a probability distribution $p \in \Delta(K)$, which is common knowledge. This selection remains fixed over all stages. The row player is informed of the outcome, whereas the column player is not.

2) Strategies: Mixed strategies are distributions over pure strategies for each player. Since the row player is informed of the state of the world, he is allowed a mixed strategy for each state k. Let $x_m^k \in \Delta(S)$ denote the mixed strategy of the row player in state k at stage m and also denote $x_m^k(s)$ to be the probability that the column player plays pure move s at stage m and state k. In repeated play, this strategy can be a function of the actions of both players during stages $1, ..., m-1$. Likewise, let $y_m \in \Delta(J)$ denote the mixed strategy of the column player at stage m, which again can depend on the players' actions over stages $1, ..., m-1$. Let $x_m = \{x_m^1, ..., x_m^K\}$ denote the collection of the row player's mixed strategies for all states at stage m, and $x = \{x_1, ..., x_m\}$ denote mixed strategies over all states and stages. Likewise, let $y = \{y_1, .., y_m\}$ denote the column player's mixed strategies over all stages.

Define $H_m = [S \times J]^{m-1}$ to be the set of possible histories where an element $h_m \in H_m$ is a sequence $(s_1, j_1; s_2, j_2; ...; s_{m-1}, j_{m-1})$ of the players' moves in the first $m-1$ stages of the game, and let h_m^I denote the history of player 1's moves. Each player can perfectly observe the moves of the other player. Therefore, the histories of each player at stage m are identical. Behavioral strategies are mappings from states and histories to mixed strategies. Let $\sigma_n : k \times h_n \to \Delta(S)$ denote a behavioral strategy of the row player and denote $\hat{\sigma}_m^k[h](s)$ to be the probability that the column player plays pure move s at stage m, history h, and state k. The column player's behavioral strategy can only depend on histories and is denoted by $\tau_n : h_n \to \Delta(J)$. Define $\sigma = \{\sigma_1, ..., \sigma_n\}$ to be the collection of behavioral strategies of the row player over all stages. Likewise, define $\tau = \{\tau_1, ..., \tau_n\}$ to be the collection of behavioral strategies of the column player over all stages. Aumann established that behavioral strategies can be equivalently represented as mixed strategies [7].

3) Beliefs: Since the column player is not informed of the selected state k, he can build beliefs on which state was selected. These beliefs are a function of the initial distribution p and the observed moves of the row player. Therefore, the row player must carefully consider his actions at each stage as they could potentially reveal the true state of the world to the column player. In order to get a worse case estimate of how much information the row player transmits through his moves, he models the column player as a Bayesian player and assumes that the column player has his mixed strategy. The updated belief p^+ is computed as

$$p^+(p, x, s) = \frac{p^k x^k(s)}{\bar{x}(p, x, s)}, \tag{1}$$

where $\bar{x}(p, x, s) := \sum_{k\in K} p^k x^k(s)$ and $x^k(s)$ is the probability of playing pure action s at state k.

4) Payoffs: Let $\gamma_m^p(\sigma, \tau)$=$\mathbf{E}\left[g_m\right]_{p,\sigma^k,\tau}$ denote the expected payoff for the pair of behavioral strategies (σ, τ) at stage m. The payoff for the n-stage game is then defined as

$$\bar{\gamma}_n^p(\sigma, \tau) = \frac{1}{n}\sum_{m=1}^{n} \gamma_m^p(\sigma, \tau). \tag{2}$$

Similarly the payoff for the λ-discounted game is defined as

$$\bar{\gamma}_\lambda^p(\sigma, \tau) = \sum_{m=1}^{\infty} \lambda(1-\lambda)^{m-1}\gamma_m^p(\sigma, \tau). \tag{3}$$

B. Short-term vs long-term tradeoff

The dynamic programming recursive formula

$$v_{n+1}(p) = \frac{1}{n+1}\Bigg[\max_{x_1}\min_{y_1}\sum_{k\in K} p^k x_1^k M^k y_1 + n\sum_{s\in S}\bar{x}_s v_n\Big(p^+(p, x_1, s)\Big)\Bigg] \tag{4}$$

introduced by Aumann and Maschler [1] characterizes the value of the zero-sum repeated game of incomplete information. Note that n is a non-negative integer and for the case where $n = 0$, the problem reduces to

$$v_1(p) = \max_{x_1}\min_{y_1}\sum_{k\in K} p^k x_1^k M^k y_1, \tag{5}$$

which is the value of the 1-shot zero-sum incomplete information game.

A key interpretation of this formulation is that it also serves as a model of the tradeoff between short-term gains and the long-term informational advantage. For each decision x_1 of the informed player, the model evaluates the payoff for the current stage, which is represented by the expression $\sum_{k\in K} p^k x_1^k M^k y_1$, and the long-term cost for decision x_1, which is represented by $n\sum_{s\in S}\bar{x}_s v_n\Big(p^+(p, x_1, s)\Big)$.

It is worth pointing out that the computational complexity for finding the optimal decision x_1^* can be attributed to the cost of calculating the long-term payoff. Since the long-term payoff is a recursive optimization problem that grows with respect to the game length, it can be difficult to find optimal strategies for games of arbitrary length. This difficulty is because the number of decision variables in the recursive optimization problem grows exponentially with respect to the game length. Ponssard and Sorin [8] showed that zero-sum repeated games of incomplete information can be formulated as a linear program (LP) to compute optimal strategies. However, in the LP formulation, it can be immediately seen that the computational complexity is also exponential with respect to the game length. One of the goals of this research is to present a method to compute suboptimal policies with tight lower bounds, and a feature of this method is that its computational complexity remains constant for games of any length.

C. Non-revealing strategies

Revealing information is defined as using a different mixed strategy in each state k at stage m. From (1), it follows that a mixed strategy x_m at stage m does not change the current beliefs of the row player if $x_m^k = x_m^{k'} \ \forall \ k, k'$. As a consequence, not revealing information is equivalent to not changing the column player's beliefs about the true state of the world. In stochastic games, it is possible for the column player's beliefs to change even if the row player uses an identical mixed strategy for each state k.

An optimal non-revealing strategy can be computed by solving

$$u(p) = \max_{x\in \text{NR}} \min_{y} \sum p^k x^k M^k y, \tag{6}$$

where the set of non-revealing strategies is defined as $\text{NR} = \{x_m \mid x_m^k = x_m^{k'} \ \forall k, k' \in K\}$. By playing an optimal non-revealing strategy at each stage of the game, the row player can guarantee a game payoff of $u(p)$. Note that the optimal game payoff for the n-stage game, $v_n(p)$, is equal to $u(p)$ only under special conditions.

III. Policy Improvement strategy

We have previously stated in Section II-B that the difficulty in explicitly computing optimal strategies for an arbitrary game grows exponentially with respect to the number of stages in the game. In this paper, we consider suboptimal strategies whose computational complexity remains constant for games of arbitrary length. One such strategy that we introduce is called one-time policy improvement. In incomplete information games, the row player always has the option of not using his superior information. A "simple" policy for the n-stage game would be for the row player to never use his superior information. As noted in Section II-C, if he uses such a policy, he can only guarantee a payoff of at most $u(p)$. In a one-time policy improvement strategy, the "simple" policy becomes the baseline policy. The key difference is that with the one-time policy improvement strategy, the row player is allowed to deviate from the "simple" policy at the first stage of the game. After the first stage, he is obliged to use the "simple" policy for the next n-1 stages.

The guaranteed payoff for the one-time policy improvement strategy for the n-stage game is denoted by $\hat{v}_n(p)$. Determining $\hat{v}_n(p)$ can be achieved by solving

$$\hat{v}_n(p) = \frac{1}{n}\Bigg[\max_{x_1}\min_{y_1} \sum p^k x_1^k M^k y_1 + (n-1)\sum_{s\in S} \bar{x}_s u\Big(p^+(p, x_1, s)\Big)\Bigg] \tag{7}$$

for $n \geq 1$.

Definition 1: Let cav $u(p)$ denote the point-wise smallest concave function g on $\Delta(K)$ satisfying $g(p) \geq u(p) \ \forall p \in \Delta(K)$.

Definition 2: Denote the one-time policy improvement behavioral strategy by $\hat{\sigma}$, where for each $\hat{\sigma}_{m\geq 2}$, let $\hat{\sigma}_m^k(h)[s] = \hat{\sigma}_m^{k'}(h)[s] \quad \forall k, k' \in K$.

Definition 3: A perpetual policy improvement strategy is a strategy that is implemented in a receding horizon manner and strategizes for the current stage while assuming a non-revealing policy for future stages.

Theorem 1: One-time policy improvement guarantees a payoff of at least cav $u(p)$ for the n-stage zero-sum repeated games of incomplete information on one-side.

Proof: We will devote the remainder of this section to the proof of this theorem. The proof will proceed as follows.

1) We will first prove in Proposition 3.5 that for any initial distribution p there exists a one-time policy improvement behavioral strategy whose lower bound is greater than or equal to $\sum_{l\in L} \alpha_l u(p_l)$, where $|L| < \infty$, $\alpha \in \Delta(L)$, $p, p_l \in \Delta(K)$, and $p = \sum_{l\in L} \alpha_l p_l$.
2) Next we note that cav $u(p) = \sum_{e\in E} \alpha_e u(p_e)$, where $\alpha \in \Delta(E)$, $|E| < \infty$, $p, p_e \in \Delta(K)$, and $p = \sum_{e\in E} \alpha_e p_e$.
3) It then follows that the one-time policy improvement behavioral strategy has a lower bound of cav $u(p)$, which is tight.
4) Recall that behavioral strategies can be equivalently represented as mixed strategies. Therefore a one-time policy improvement behavioral strategy can be equivalently represented as a one-time policy improvement mixed strategy. As a consequence $\hat{v}_n(p) \geq$ cav $u(p)$

■

Corollary 2: Perpetual policy improvement guarantees a payoff of at least cav $u(p)$ for the n-stage zero-sum repeated games of incomplete information on 1-side.

Proof: This can be shown by using standard dynamic programming arguments regarding policy improvement. ■

Lemma 3.1: [9] Let L be a finite set and $p = \sum_{l\in L} \alpha_l p_l$ with $\alpha \in \Delta(L)$, and p,$p_l \in \Delta(K)$ for all l in L Then there exists a transition probability μ from (K, p) to L such that

$$P(l) = \alpha_l \quad \text{and} \quad P(\cdot|l) = p_l,$$

where $P = p \cdot x$ is the probability induced by p and μ on $K \times L : P(k, l) = p^k \mu^k(l)$.

Lemma 3.2: Fix p arbitrarily. Let p be represented as $p = \sum_{l\in L} \alpha_l p_l$, where $|L| < \infty$, $\alpha \in \Delta(L)$, and p,$p_l \in \Delta(K)$

Then their exists a strategy σ, which will be referred to as the splitting strategy, such that

$$\bar{\gamma}^p_n(\sigma,\tau) \geq \sum_{l\in L} \alpha_l u(p_l) \ \ \forall\tau \tag{8}$$

Proof:

1) Introduce strategy σ as follows. Let σ_l be the strategy that guarantees $u(p_l)$ for the n-stage game. Define $\mu^k(l) = \alpha_l \, \frac{p^k_l}{p^k}$. If the state is k, use the lottery μ^k, and if the outcome is l, play σ_l.
2) To get a lower bound on Player 1's payoff, assume even that Player 2 is informed upon l. He is then facing strategy σ_l
3) By Lemma 3.1, this occurs with probability α_l and the conditional probability on K is p_l, hence the game is $\gamma^{p_l}_n$, so that $\bar{\gamma}^p_n(\sigma,\tau) \geq \sum_{l\in L}\alpha_l u(p_l) \ \ \forall\tau$

■

Lemma 3.3: Consider mixed strategy $\tilde{x}$, where $\tilde{x} = \{\tilde{x}^1, \tilde{x}^2, ..., \tilde{x}^{|K|}\}$. There exists a one-time policy improvement behavioral strategy $\hat{\sigma}$ where $\hat{\sigma}^k_m(s) = \tilde{x}^k(s) \ \forall \ k$.

Proof:

1) Consider the following behavioral strategy. At stage $m = 1$, Player 1 uses mixed strategy $\tilde{x}^k$, where k is the true state of the world (i.e. $\hat{\sigma_1} : k \times \{\emptyset\} \mapsto \tilde{x}^k$). Whatever move $s' \in S$ that is realized in stage $m = 1$, Player 1 plays s' at each stage for the remainder of the game (i.e. $\hat{\sigma_n} : h_{m-1} \mapsto h_1 \quad$ where $m \geq 2$ and h_{m-1} is player 1's history at stage $m-1$).
2) Clearly $\hat{\sigma}^k_1(s) = \tilde{x}^k(s)$ by definition.
3) Consider stage 2. Suppose the probability of playing move s' in state k at stage 1 is α. Recall that whatever move that the row player realizes in stage 1 is played for the rest of the game. It follows that if the state of world is state k, the probability of playing move s' in stage 2 is also α.
4) The same argument can be applied to stage m.
5) We have now established that the following equality $\hat{\sigma}^k_m(s) = \tilde{x}^k(s)$ holds for all m, where $\hat{\sigma}$ is the one-time policy improvement behavioral strategy.

■

Proposition 3.4: Fix p arbitrarily. Let p be represented as $p = \sum_{l\in L}\alpha_l p_l$, where $|L| < \infty$, $\alpha \in \Delta(L)$, and p,p_l $\in \Delta(K)$. Suppose we construct a behavioral strategy $\bar{\sigma}$ as follows. Define σ_l to be the optimal behavioral strategy to the non-revealing n-stage game $u(p_l)$. Let $\bar{\sigma}^k = \sum_{l\in L}\mu^k(l)\sigma^k_l$ define the mixed behavioral strategy for the n-stage game, where $\mu^k(l) = \alpha_l \, \frac{p^k_l}{p^k}$. The lower bound for the payoff of behavioral strategy $\bar{\sigma}$ is $\sum_{l\in L}\alpha_l u(p_l)$. Explicitly

$$\bar{\gamma}^p_n(\bar{\sigma},\tau) = \sum_{k\in K} p^k \left[\frac{1}{n}\sum_{m=1}^{n} \mathbf{E}\left[g_m\right]_{k,\bar{\sigma}^k,\tau}\right]$$

$$= \frac{1}{n}\sum_{k\in K} p^k \left[\sum_{m=1}^{n}\sum_{l\in L}\mu^k(l)\mathbf{E}\left[g_m\right]_{k,\sigma^k_l,\tau}\right] \tag{9}$$

$$\geq \sum_{l\in L}\alpha_l u(p_l) \ \ \forall\tau$$

where τ is the behavioral strategy of player 2

Proof: We will first establish that the splitting strategy has the equivalent payoff of mixed behavioral strategy $\bar{\sigma}$ for arbitrary behavioral strategies τ of the column player. We then note that the splitting strategy has a lower bound of $\sum_{l\in L}\alpha_l u(p_l)$. We conclude by making the following observation. Since the splitting strategy has an identical payoff as strategy $\bar{\sigma}$ for each strategy τ, it also has the same lower bound $\sum_{l\in L}\alpha_l u(p_l)$.

1) Recall the splitting strategy as defined in Lemma 3.2.
2) The payoff for this strategy can be expressed as follows:

$$\sum_{k\in K}\sum_{l\in L} p^k \mu^k(l)\left[\frac{1}{n}\sum_{m=1}^{n}\mathbf{E}\left[g_m\right]_{k,\sigma^k_l,\tau}\right]$$

$$= \frac{1}{n}\sum_{k\in K} p^k \sum_{l\in L}\mu^k(l)\left[\sum_{m=1}^{n}\mathbf{E}\left[g_m\right]_{k,\sigma^k_l,\tau}\right]$$

$$= \frac{1}{n}\sum_{k\in K} p^k \sum_{m=1}^{n}\sum_{l\in L}\mu^k(l)\mathbf{E}\left[g_m\right]_{k,\sigma^k_l,\tau} \tag{10}$$

3) Observe that the payoff for strategy $\bar{\sigma}$ in (9) is equivalent to the payoff for the splitting strategy σ in (10) for arbitrary τ. Explicitly $\bar{\gamma}^p_n(\bar{\sigma},\tau) = \bar{\gamma}^p_n(\sigma,\tau) \ \forall\tau$.
4) Recall Lemma 3.2, which states that the splitting strategy has a payoff with lower bound $\sum_{l\in L}\alpha_l u(p_l)$.
5) *Conclusion:* Given a behavioral strategy τ of player 2, we have established that the payoff of strategy $\bar{\sigma}$ is equivalent to that of the splitting strategy. We show that the payoff of splitting strategy is lower bounded by $\sum_{l\in L}\alpha_l u(p_l)$. This implies that strategy $\bar{\sigma}$ also has this lower bound.

■

Proposition 3.5: There exists a one-time policy improvement strategy $\hat{\sigma}$ such that the following inequality holds.

$$\bar{\gamma}^p_n(\hat{\sigma},\tau) \geq \sum_{l\in L}\alpha_l u(p_l)$$

Proof:

1) Recall the mixed behavioral strategy $\bar{\sigma}$ as defined in Proposition 3.4, where $\bar{\sigma}^k = \sum_{l\in L}\mu^k(l)\sigma^k_l$.
2) Note first that since σ_l is an optimal non-revealing strategy, behavioral strategy σ^k_l is the same for every state k (i.e. $\sigma^k_l = \sigma^{k'}_l \ \forall k, k'$). Furthermore, the NR mixed strategy x^*_l is constant for each stage
3) Therefore $\sigma^k_{l_n} = x^*_l \ \ \forall k, \ \forall n$.
4) Define $\alpha^k = \sum_{l\in L}\mu^k(l)x^*_l$.
5) We can then express $\bar{\sigma}_n$ as $\bar{\sigma}_n : K \times H_{n-1} \mapsto \alpha^k \quad \forall n$, which is a stationary strategy.
6) Define $\tilde{x}$ as follows: $\tilde{x} = \{\alpha^1, \alpha^2, ..., \alpha^{|K|}\}$.

7) By Lemma 3.3, there exists a one-time policy improvement strategy $\hat{\sigma}$ such that $\hat{\sigma}^k_m(s) = \tilde{x}^k(s)\ \forall\ s, m$
8) We have now established that $\bar{\gamma}^p_n(\hat{\sigma}, \tau) = \bar{\gamma}^p_n(\bar{\sigma}, \tau)\ \forall\ \tau$
9) Therefore it follows that

$$\bar{\gamma}^p_n(\hat{\sigma}, \tau) = \bar{\gamma}^p_n(\bar{\sigma}, \tau) \geq \sum_{l \in L} \alpha_l u(p_l)\ \ \forall \tau \qquad (11)$$

■

IV. Policy Improvement in Infinite Horizon Games

In the previous section, we have shown that the one-time policy improvement strategy guarantees cav $u(p)$ for the n-stage game. In this section we will show that the guarantee also holds in the infinite horizon games. The guaranteed payoff for the one-time policy improvement strategy in the λ-discounted infinite horizon games can be computed by solving

$$\max_{x_1} \min_{y_1} \Big\{ \lambda \sum p^k x^k_1 M^k y_1 + (1 - \lambda) \sum_{s \in S} \bar{x}_s u\Big(p^+(p, x_1, s)\Big) \Big\} \qquad (12)$$

for $\lambda \in (0, 1)$.

Theorem 3: One-time policy improvement guarantees a payoff of at least cav $u(p)$ for the λ-discounted infinite horizon zero-sum repeated games of incomplete information on one-side.

Proof: We will show the existence of a one-time policiy improvement strategy that guarantees at least cav $u(p)$.

1) Fix $\lambda \in (0, 1)$ arbitrarily.
2) There exists N' s.t. $\lambda > \frac{1}{N'}$. Consider this N'.
3) Let $\tilde{\lambda} = \frac{1}{N'}$, then $\hat{v}_{\tilde{\lambda}}(p) = \hat{v}_{N'}(p)$.
4) Invoking Theorem 1 yields $\hat{v}_{\tilde{\lambda}}(p) = \hat{v}_{N'}(p) \geq$ cav $u(p)$
5) *Claim:* The optimal one-time policy improvement strategy for the discounted game $\hat{v}_{\tilde{\lambda}}(p)$ also guarantees at least cav $u(p)$ for $\hat{v}_{\lambda}(p)$.

   Proof:
   Since $\hat{v}_{\tilde{\lambda}}(p) \geq$ cav $u(p)$ and $\sum_{s \in S} \bar{x}_s u(p^+ \mid x_1, s) \leq$ cav $u(p)$, it follows that the optimal stage 1 strategy x^*_1 for $\hat{v}_{\tilde{\lambda}}(p)$ has the following lower bound: $\sum p^k {x^k_1}^* M^k y_1 \geq cav\, u(p)$. Note that $\lambda > \tilde{\lambda}$. Therefore

$$\lambda \sum p^k {x^k_1}^* M^k y_1 + (1 - \lambda) \sum_{s \in S} \bar{x}_s u\Big(p^+(p, x_1, s)\Big)$$

$$\geq \tilde{\lambda} \sum p^k {x^k_1}^* M^k y_1 + (1 - \tilde{\lambda}) \sum_{s \in S} \bar{x}_s u\Big(p^+(p, x_1, s)\Big)$$

$$\geq \text{cav}\, u(p).$$

   Remark: If $\sum_{s \in S} \bar{x}_s u(p^+(p, x^*_1, s)) =$ cav $u(p)$ and $\sum p^k {x^k_1}^* M^k y_1 =$ cav $u(p)$ then $\hat{v}_{\lambda}(p) = \hat{v}_{\tilde{\lambda}}(p) =$ cav $u(p)$.
6) By using the optimal stage 1 strategy x^*_1 obtained from $\hat{v}_{\tilde{\lambda}}(p)$ and playing an optimal non-revealing strategy thereafter, we have constructed a one-time policy update strategy for $\hat{v}_{\lambda}(p)$ that guarantees cav $u(p)$.

■

Theorem 4: One-time policy improvement is an optimal strategy for infinitely-repeated zero-sum games of incomplete information on one-side.

Proof: Using an argument similar to that of Theorem 3, one can establish that there exists a one-time policy improvement strategy that guarantees a payoff of at least cav $u(p)$. Note that cav $u(p)$ is the optimal payoff for the infinitely-repeated game. ■

V. LP Formulation

A one-time policy-improvement strategy that guarantees cav $u(p)$ can be computed by solving a linear programing problem, and the computational complexity of the linear program is constant with respect to the number of stages of the game.

The following outlines a procedure to construct the appropriate linear program.

1) Let $\tilde{\Sigma}$ denote the set of "pure" one-time policy-improvement behavioral strategies (i.e. $\tilde{\sigma}_1 : k \mapsto s$, $\tilde{\sigma}_m : h^I_1 \mapsto s\ \ \forall m \geq 2$, and $\tilde{\sigma}_m = \tilde{\sigma}_{m'}\ \forall m, m'$)
2) Let $\tilde{T}$ denote the set of "pure" strategies for the uninformed player. (i.e. $\tilde{\tau}_1 : \emptyset \mapsto j$, $\tilde{\sigma}_m : h^I_1 \mapsto j$ $\forall m \geq 2$, and $\tilde{\tau}_m = \tilde{\tau}_{m'}\ \forall m, m'$)
3) Note that the size of the strategy sets $\tilde{\Sigma}$ and $\tilde{T}$ are invariant with respect to the number of stages of the game.
4) Observe that since a one-time policy-improvement strategy is used, the strategy and the corresponding payoff remains constant for $m \geq 2$, so that $\gamma^p_m(\tilde{\sigma}, \tilde{\tau}) = \gamma^p_2(\tilde{\sigma}, \tilde{\tau})\ \forall m \geq 2$.
5) Therefore, the game payoff for the strategy pair is $\tilde{\gamma}^p_{\lambda}(\tilde{\sigma}^i, \tilde{\tau}^l) = \lambda \gamma^p_1(\tilde{\sigma}^i, \tilde{\tau}^l) + (1 - \lambda)\gamma^p_2(\tilde{\sigma}^i, \tilde{\tau}^l)$. *For the n-stage game, set $\lambda = \frac{1}{n}$.*
6) Consider a matrix M, where element (i, l) denotes the game payoff $\tilde{\gamma}^p_{\lambda}(\tilde{\sigma}^i, \tau^l)$ for the strategy pair $(\tilde{\sigma}^i, \tilde{\tau}^l)$.
7) Since this is a zero sum game with finite strategies for each player and a payoff matrix M, a classic zero-sum game result can be used to solve this zero-sum game as a LP.

As a consequence, a perpetual policy-improvement strategy that guarantees cav $u(p)$ can be computed by solving a linear programing problem at each stage of the game, and the computational complexity of the linear program is constant with respect to the number of stages of the game.

The following outlines a procedure for this construction.

1) Note that the stage 1 behavioral strategy σ_1 is the one-time policy-improvement strategy that can be computed by solving a LP.
2) Consider stage 2. Since this is a game of perfect recall, the behavioral strategy σ_1 can be equivalently represented as a mixed strategy x_1.

3) A move of player 1 was realized in stage 1, and since the mixed strategy x_1 is known, the posterior probability p^+ can be computed.
4) Compute the stage 2 strategy by solving the optimization problem $\hat{v}_\lambda(p^+)$. If it is a n-stage game, set $\lambda = \frac{1}{N-1}$.
5) The same techniques that was used to compute the Stage 1 strategy by solving a LP, can be used for Stage 2.
6) By a similar argument, the Stage m strategy can be computed by solving a LP.

VI. Simulation: Cyber Security Example

The network administrator, the row player, manages two web applications $wapp_1$ and $wapp_2$ (i.e. e-mail and remote login). Each web application is run on its own dedicated server (H_1 and H_2). The attacker would like to prevent users from accessing the web applications via a Denial of Service (DOS) attack. In order to help mitigate DOS attacks, a spare server (H_{spare}) can be dynamically configured daily to also run either $wapp_1$ or $wapp_2$. We assume that the attacker can observe which web application the network admin decides to run on the spare server. The attacker has a choice of which web application to execute a Denial of Service.

	$wapp_1$	$wapp_2$
$wapp_1$	0	1
$wapp_2$	0	$\frac{1}{2}$

State α

	$wapp_1$	$wapp_2$
$wapp_1$	$\frac{1}{2}$	0
wap_2	1	0

State β

In state α, all of the legitimate users on the network are using only $wapp_1$. Conversely in state β, they are all only using $wapp_2$. The payoff for each of the states are in terms of of quality of service. Suppose that $p^\alpha = .5$ and $p^\beta = .5$, where p^α denotes the initial probability of being in state α and let $p = (p^\alpha, p^\beta)$. In this example, we will set the number of stages of the game to 2. If the network admin. plays his dominant strategy on day 1, he will have fully revealed the state of the network to the attacker on day 2. *The dominant strategy for the network admin. is to run web application* 1 *on the backup server in state* α *and to run web application* 2 *on the backup server in* β. The network admin. will achieve an expected payoff of 0.5 on day 1 and a payoff of 0 on day 2 for a total payoff of $.25$.

If the network admin uses a 1-time policy improvement strategy, will he do better? Solving numerically we compute a mixed strategy of $x^\alpha = (0.50, .50)$ and $x^\beta = (.50, .50)$, with an expected game payoff of 0.375 over the two day period. Suppose, on day 2 that the network admin. decides to again improve upon the baseline policy of playing non-revealing. By exploiting his information on day 2, he can yield an expected stage payoff of $.50$ and a game payoff of $.4375$.

VII. Conclusion

Computing optimal strategies for zero-sum repeated games with asymmetric information can be computationally prohibitive. Much of the current work to address this issue has been limited to special cases. In this work, we present policy improvement methods to compute suboptimal strategies that have tight lower bounds. We show that the policy improvement strategy can be computed by solving a linear program, and the computational complexity of the linear program remains constant with respect to the number of stages in the game.

In this paper, we focused on computational results for repeated games, where the state of the world remains fixed once it has been selected by nature. Repeated games are a special case of stochastic games. In stochastic games, the state of the world can change at each stage of the game and the state transitions can be a function of the actions of the players. We would like to consider computational results for stochastic games and also extend our current results to stochastic games.

References

[1] R. J. Aumann and M. Maschler, *Repeated Games with Incomplete Information*. MIT Press, 1995.

[2] V. C. Domansky and V. L. Kreps, “Eventually revealing repeated games of incomplete information,” *International Journal of Game Theory*, vol. 23, pp. 89–109, 1994.

[3] M. Heur, “Optimal strategies for the uninformed player,” *International Journal of Game Theory*, vol. 20, pp. 33–51, 1991.

[4] S. Zamir, “On the relation between finitely and infinitely repeated games with incomplete information,” *International Journal of Game Theory*, vol. 23, pp. 179–198, 1971.

[5] A. Gilpin and T. Sandholm, “Solving two-person zero-sum repeated games of incomplete information,” *International Joint Conference on Autonomous Agents and Multiagent Systems*, vol. 2, pp. 903–910, 2008.

[6] S. Zamir, “Repeated games of incomplete information: Zero-sum,” *Handbook of Game Theory*, vol. 1, pp. 109–154, 1999.

[7] R. Aumann, *Mixed and behavior strategies in infinite extensive games*. Princeton University, 1961.

[8] J. Ponssard and S. Sorin, “The l-p formulation of finite zero-sum games with incomplete information,” *International Journal of Game Theory*, vol. 9, pp. 99–105, 1999.

[9] S. Sorin, *A First Course on Zero-Sum Repeated Games*. Springer, 2002.

[10] D. Blackwell, “An analog of the minimax theorem for vector payoffs.” *Pacific Journal of Mathematics*, vol. 1956, no. 1, pp. 1–8, 1956.

[11] Y. Freund and R. E. Schapire, “Game theory, on-line prediction and boosting,” in *Proceedings of the ninth annual conference on Computational learning theory*, ser. COLT ’96. New York, NY, USA: ACM, 1996, pp. 325–332. [Online]. Available: http://doi.acm.org/10.1145/238061.238163

[12] D. Rosenberg, E. Solan, and N. Vieille, “Stochastic games with a single controller and incomplete information,” Northwestern University, Center for Mathematical Studies in Economics and Management Science, Tech. Rep. 1346, May 2002.

[13] J.-F. Mertens and S. Zamir, “The value of two-person zero-sum repeated games with lack of information on both sides,” in *Institute of Mathematics, The Hebrew University of Jerusalem*, 1970, pp. 405–433.

[14] J.-F. Mertens, “The speed of convergence in repeated games with incomplete information on one side,” Universit catholique de Louvain, Center for Operations Research and Econometrics (CORE), Tech. Rep. 1995006, Jan. 1995.