Computing Stackelberg Equilibria in Discounted Stochastic Games (Corrected Version)

Yevgeniy Vorobeychik
Sandia National Laboratories
Livermore, CA
yvorobe@sandia.gov

Satinder Singh
Computer Science and Engineering
University of Michigan
Ann Arbor, MI
baveja@umich.edu

Abstract

Stackelberg games increasingly influence security policies deployed in real-world settings. Much of the work to date focuses on devising a fixed randomized strategy for the defender, accounting for an attacker who optimally responds to it. In practice, defense policies are often subject to constraints and vary over time, allowing an attacker to infer characteristics of future policies based on current observations. A defender must therefore account for an attacker's observation capabilities in devising a security policy. We show that this general modeling framework can be captured using stochastic Stackelberg games (SSGs), where a defender commits to a dynamic policy to which the attacker devises an optimal dynamic response. We then offer the following contributions. 1) We show that Markov stationary policies do not suffice in SSGs, except in several very special cases; 2) present a finite-time mixed-integer nonlinear program for computing a Stackelberg equilibrium in SSGs when the leader is restricted to Markov stationary policies, and 3) present a mixed-integer linear program to approximate it. 4) We illustrate our algorithms on a simple SSG representing an adversarial patrolling scenario, where we study the impact of attacker patience and risk aversion on optimal defense policies.

Introduction

Recent work using Stackelberg games to model security problems in which a defender deploys resources to protect targets from an attacker has proven very successful both in yielding algorithmic advances (Conitzer and Sandholm 2006; Paruchuri et al. 2008; Kiekintveld et al. 2009; Jain et al. 2010a) and in field applications (Jain et al. 2010b; An et al. 2011). The solution to these games are Stackelberg Equilibria, or SE, in which the attacker is assumed to know the defender's mixed strategy and plays a best response to it (breaking ties in favor of the defender makes it a Strong SE, or SSE). The defender's task is to pick an optimal (usually mixed) strategy given that the attacker is going to play a best-response to it. This ability of the attacker to know the defender's strategy in SE is motivated in security problems by the fact that the attacker can take advantage of surveillance prior to the actual attack. The simplest Stackelberg games are single-shot zero-sum games. These assumptions keep the computational complexity of finding solutions manageable but limit applicability. In this paper we approach the problem from the other extreme of generality by addressing *SSE computation in general-sum discounted stochastic Stackelberg games (SSGs)*. Our main contributions are: 1) showing that there need not exist SSE in Markov stationary strategies, 2) providing a finite-time general MINLP (mixed-integer nonlinear program) for computing SSE when the leader is restricted to Markov stationary policies, 3) providing an MILP (mixed-integer linear program) for computing approximate SSE in Markov stationary policies with provable approximation bounds, and 4) a demonstration that the generality of SSGs allows us to obtain qualitative insights about security settings for which no alternative techniques exist.

Notation and Preliminaries We consider two-player infinite-horizon discounted stochastic Stackelberg games (SSGs from now on) in which one player is a "leader" and the other a "follower". The leader commits to a policy that becomes known to the follower who plays a best-response policy. These games have a finite state space S, finite action spaces A_L for the leader and A_F for the follower, payoff functions $R_L(s, a_l, a_f)$ and $R_F(s, a_l, a_f)$ for leader and follower respectively, and a transition function $T_{ss'}^{a_l a_f}$, where $s, s' \in S$, $a_l \in A_L$ and $a_f \in A_F$. The discount factors are $\gamma_L, \gamma_F < 1$ for the leader and follower, respectively. Finally, $\beta(s)$ is the probability that the initial state is s.

The history of play at time t is $h(t) = \{s(1)a_l(1)a_f(1) \ldots s(t-1)a_l(t-1)a_f(t-1)s(t)\}$ where the parenthesized indices denote time. Let Π (Φ) be the set of unconstrained, i.e., nonstationary and non-Markov, policies for the leader (follower), i.e., mappings from histories to distributions over actions. Similarly, let Π_{MS} (Φ_{MS}) be the set of Markov stationary policies for the leader (follower); these map the last state $s(t)$ to distributions over actions. Finally, for the follower we will also need the set of deterministic Markov stationary policies, denoted Φ_{dMS}.

Let U_L and U_F denote the utility functions for leader and follower respectively. For arbitrary policies $\pi \in \Pi$ and $\phi \in$

Copyright © 2013, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Φ, $U_L(s, \pi, \phi)$

$$= \mathbb{E}\Big[\sum_{t=1}^{\infty} \gamma_L^{t-1} R_L(s(t), \pi(h(t)), \phi(h(t))) | s(1) = s\Big],$$

where the expectation is over the stochastic evolution of the states, and where (abusing notation) $R_L(s(t), \pi(h(t)), \phi(h(t)))$

$$= \sum_{a_l \in A_L} \sum_{a_f \in A_F} \pi(a_l | h(t)) \phi(a_f | h(t)) R_L(s(t), a_l, a_f),$$

and $\pi(a_l | h(t))$ is the probability of leader-action a_l in history $h(t)$ under policy π, and $\phi(a_f | h(t))$ is the probability of follower-action a_f in history $h(t)$ under policy ϕ. The utility of the follower, $U_F(s, \pi, \phi)$, is defined analogously.

For any leader policy $\pi \in \Pi$, the follower plays the best-response policy defined as follows:

$$\phi_\pi^{BR} \overset{\text{def}}{\in} \arg\max_{\phi \in \Phi} \sum_s \beta(s) U_F(s, \pi, \phi).$$

The leader's optimal policy is then

$$\pi^* \overset{\text{def}}{\in} \arg\max_{\pi \in \Pi} \sum_s \beta(s) U_L(s, \pi, \phi_\pi^{BR})$$

Together $(\pi^*, \phi_{\pi^*}^{BR})$ constitute a Stackelberg equilibrium (SE). If, additionally, the follower breaks ties in the leader's favor, these are a Strong Stackelberg equilibrium (SSE).

A crucial question is: must we consider the complete space of non-stationary non-Markov policies to find a SE? Before presenting an answer, we briefly discuss related work and present an example problem modeled as an SSG.

Related Work and Example SSG While much of the work on SSE in security games focuses on one-shot games, there has been a recent body of work studying patrolling in adversarial settings that is more closely related to ours. In general terms, adversarial patrolling involves a set of targets which a defender protects from an attacker. The defender chooses a randomized patrol schedule which must obey exogenously specified constraints. As an example, consider a problem that could be faced by a defender tasked with using a single boat to patrol the five targets in Newark Bay and New York Harbor shown in Figure 1, where the graph roughly represents geographic constraints of a boat patrol. The attacker observes the defender's current location, and knows the probability distribution of defender's next moves. At any point in time, the attacker can wait, or attack immediately any single target, thereby ending the game. The number near each target represents its value to the defender and attacker. What makes this problem interesting is that two targets have the highest value, but the defender's patrol boat cannot move directly between these.

Some of the earliest work (Agmon, Kraus, and Kaminka 2008; Agmon, Urieli, and Stone 2011) on adversarial patrolling was done in the context of robotic patrols, but involved a highly simplified defense decision space (for example, with a set of robots moving around a perimeter, and a single parameter governing the probability that they move forward or back). Basilico *et al.* (Basilico, Gatti, and Amigoni 2009; Basilico et al. 2010; Basilico, Gatti, and Villa 2011; Basilico and Gatti 2011; Bosansky et al. 2011) studied general-sum patrolling games in which they assumed that the attacker is infinitely patient, and the execution of an attack can take an arbitrary number of time steps. Recent work by Vorobeychik, An, and Tambe (2012) considers only zero-sum stochastic Stackelberg games.

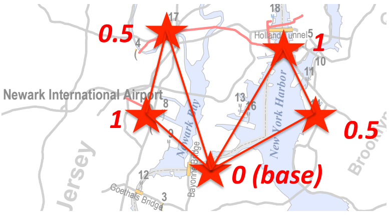

Figure 1: Example of a simple Newark Bay and New York Harbor patrolling scenario.

Considering SSGs in full generality, as we do here, yields the previous settings as special cases (modulo the discount factor). Our results, for example, apply directly to discounted variants of adversarial patrolling settings studied by Basilico et al. Moreover, our use of discount factors makes our setting more plausible: it is unlikely that an attacker is entirely indifferent between now, and an arbitrarily distant future. Finally, Basilico et al. policies are restricted to depend only on previous defender move, even when the attacks take time to unfold; this restriction is approximate, whereas the generality of our formulations allows an exact solution by representing states as finite sequences of defender moves.

Adversarial Patrolling as an SSG We illustrate how to translate our general SSG model to adversarial patrolling on graphs for the example of Figure 1. The state space is the nodes in the graph plus a special "absorbing" state; the game enters this state when the attacker attacks, and remains there for ever. At any point in time, the state is the current location of the defender, the defender's actions A_F are a function of the state and allow the defender to move along any edge in the graph, the attacker's actions A_F are to attack any node in the graph or to wait. Assuming that the target labeled as "base" is the starting point of the defender defines the initial distribution over states. The transition function is a deterministic function of the defender's action (since state is identified with defender's locations) except after an attack, which transitions the game into the absorbing state. The payoff function is as follows: if the attacker waits, both agents get zero payoff; if the attacker attacks node j valued H_j, while the defender chooses action $i \neq j$, the attacker receives H_j, which is lost to the defender. If, on the other hand, defender also chooses j, both receive zero. Thus, as constructed, it is a zero-sum game. We will use the problem of Figure 1 below for our empirical illustrations.

The form of a SSE in Stochastic Games

It is well known that in general-sum stochastic games there always exists a Nash equilibrium (NE) in Markov stationary policies (Filar and Vrieze 1997). The import of this result is that it allows one to focus NE computation on this very restricted space of strategies. In the version of the paper published in AAAI proceedings, we provided a “proof” of the following result:

Theorem 1 (FALSE[1]). *For any general-sum discounted stochastic Stackelberg game, there exist a leader's Markov stationary policy and a follower's deterministic Markov stationary policy that form a strong Stackelberg equilibrium.*

Unfortunately, this result is false at the stated level of generality, as we now proceed to demonstrate (we are grateful to Vincent Conitzer for providing the counterexample we use below).

Before we demonstrate the falsehood of the above theorem, let us state a very basic, and weak, result that does hold in general:

Lemma 1. *For any general-sum discounted stochastic Stackelberg game, if the leader follows a Markov stationary policy, then there exists a deterministic Markov stationary policy that is a best response for the follower.*

This follows from the fact that if the leader plays a Markov stationary policy, the follower faces a finite MDP. A slightly weaker result is, in fact, at the core of proving the existence of Markov stationary NE: it allows one to define a best response correspondence in the space of (stochastic) Markov stationary policies of each player, and an application of Kakutani's fixed point theorem completes the proof. The difficulty that arises in SSGs is that, in general, the leader's policy *need not be a best response to the follower's*.

We now show that Theorem 1 fails to hold even in highly restricted special cases of SSGs.

Example 1. The leader's optimal policy may not be Markov stationary even if transitions are deterministic and independent of player actions. Moreover, the best stationary policy can be arbitrarily suboptimal. *Consider the following counterexample, suggested to us by Vincent Conitzer. Suppose that the SSG has three states, i.e., $S = \{1, 2, 3\}$, and the leader and the follower have two actions each, $A_L = \{U, D\}$ for the leader and $A_F = \{L, R\}$ for the follower. Let initial state be $s = 1$ and suppose that the following transitions happen deterministically and independently of either player's decisions: $T_{12} = 1$, $T_{23} = 1$, $T_{33} = 1$, that is, the process starts at state 1, then moves to state 2, then, finally, to state 3, which is an absorbing state. In state $s = 1$ only the follower's actions have an effect on payoffs, which is as follows: $R_L(1, \cdot, L) = -M$, $R_L(1, \cdot, R) = 0$, $R_F(1, \cdot, L) = \epsilon$, $R_F(1, \cdot, R) = 0$, where M is an arbitrarily large number and $\epsilon << M$. In state $s = 2$, in contrast, only the leader's actions have an effect on payoffs: $R_L(2, U, \cdot) = R_L(2, D, \cdot) = 0$, $R_F(2, U, \cdot) = -M$, $R_F(2, D, \cdot) = 0$. Suppose that the discount factors $\gamma = \delta$ are close to 1. First, note that a Markov stationary policy for the leader would be independent of the follower's action in state 2, and, consequently, the follower's best response is to play L, giving the leader a payoff of $-M$. On the other hand, if the leader plays the following non-Markov policy: play U when the follower plays L and D otherwise, the follower's optimal policy is to play R, and the leader receives a payoff of 0. Since M is arbitrarily large, the difference between an optimal and best stationary policy is arbitrarily large.* □

A natural question is whether there is any setting where a positive result is possible, besides zero-sum games where there is no distinction between Nash equilibria and SSE. Indeed, there is: team games.

Definition 1. *A* team game *is a SSG with $R_L(s, a_l, a_f) = R_F(s, a_l, a_f) = R(s, a_l, a_f)$ and $\gamma_L = \gamma_F$.*

Proposition 1. *For any general-sum discounted team game, there exist a leader's Markov stationary policy and a follower's deterministic Markov stationary policy that form a strong Stackelberg equilibrium. Moreover, these are both deterministic.*

Proof. Construct an MDP with the same state space as the team game, but the actions space $A = A_L \times A_F$ (which is still finite), the reward function is $R(s, a)$ where $a = (a_l, a_f) \in A$, and the transition probabilities are as in the original team game. Let π^*_{MDP} be an optimal deterministic stationary Markov policy of the resulting MDP, which is known to exist. We can decompose this policy into $\pi^*_{MDP} = (\pi^*_L, \phi^*_F)$, where the former simply specifies the leader's and the latter the follower's part in the optimal MDP policy. We now claim that (π^*_L, ϕ^*_F) constitutes a SSE.

First, we show that ϕ^*_F must be the best response to π^*_L. Let $U(\pi, \phi)$ be the expected utility of both leader and follower when following π and ϕ respectively, where expectation is taken also with respect to the initial distribution over states; that these are equal follows by the identity of the payoffs and discount factors in the team game. Note that $U(\pi, \phi) = U(\pi_{MDP} = (\pi, \phi))$, where the latter is the corresponding expected utility of the MDP we constructed above. Now, suppose that there is ϕ' which yields a higher utility to the follower. Then,

$$U(\pi^*, \phi') = U_F(\pi^*, \phi') > U_F(\pi^*, \phi^*) = U(\pi^*, \phi^*),$$

which implies that $U(\pi^*, \phi') > U(\pi^*, \phi^*)$, a contradiction, since (π^*, ϕ^*) are optimal for the MDP.

Second, we show that π^* is leader-optimal. Suppose not. Then there exists (π', ϕ') where ϕ' is a best response to π' and

$$U(\pi', \phi') = U_L(\pi', \phi') > U_L(\pi^*, \phi^*) = U(\pi^*, \phi^*),$$

which implies that $U(\pi', \phi') > U(\pi^*, \phi^*)$, a contradiction, since (π^*, ϕ^*) are optimal for the MDP. □

Computing Markov Stationary SSE Exactly

While in general SSE in Markov stationary strategies do not suffice, we restrict attention to these in the sequel, as general policies need not even be finitely representable. A crucial

[1] Our proof in the proceedings verison of the paper went awry in two ways. First, we assumed that there always exists a leader-optimal policy that is optimal in every state. Second, our approach relied on backwords induction, whereas in SSGs policies have complex inter-temporal dependencies.

consequence of the restriction to Markov stationary strategies is that policies of the players can now be finitely represented. In the sequel, we drop the cumbersome notation and denote leader stochastic policies simply by π and follower's best response by ϕ (with π typically clear from the context). Let $\pi(a_l|s)$ denote the probability that the leader chooses $a_l \in A_L$ when he observes state $s \in S$. Similarly, let $\phi(a_f|s)$ be the probability of choosing $a_f \in A_F$ when state is $s \in S$. Above, we also observed that it suffices to focus on *deterministic* responses for the attacker. Consequently, we assume that $\phi(a_f|s) = 1$ for exactly one follower action a_f, and 0 otherwise, in every state $s \in S$.

At the root of SSE computation are the expected optimal utility functions of the leader and follower starting in state $s \in S$ defined above and denoted by $V_L(s)$ and $V_F(s)$. In the formulations below, we overload this notation to mean the variables which compute V_L and V_F in an optimal solution. Suppose that the current state is s, the leader plays a policy π, and the follower chooses action $a_f \in A_F$. The follower's expected utility is $\tilde{R}_F(s, \pi, a_f)$

$$= \sum_{a_l \in A_L} \pi(a_l|s) \left(R_F(s, a_l, a_f) + \gamma_F \sum_{s' \in S} T_{ss'}^{a_l a_f} V_F(s') \right).$$

The leader's expected utility $\tilde{R}_L(s, \pi, a_f)$ is defined analogously. Let Z be a large constant. We now present a mixed integer non-linear program (MINLP) for computing a SSE:

$$\max_{\pi, \phi, V_L, V_F} \sum_{s \in S} \beta(s) V_L(s) \tag{1a}$$

$$\text{subject to :}$$

$$\pi(a_l|s) \geq 0 \qquad \forall s, a_l \tag{1b}$$

$$\sum_{a_l} \pi(a_l|s) = 1 \qquad \forall s \tag{1c}$$

$$\phi(a_f|s) \in \{0, 1\} \qquad \forall s, a_f \tag{1d}$$

$$\sum_{a_f} \phi(a_f|s) = 1 \qquad \forall s \tag{1e}$$

$$0 \leq V_F(s) - \tilde{R}_F(s, \pi, a_f) \leq (1 - \phi(a_f|s))Z \; \forall s, a_f \tag{1f}$$

$$V_L(s) - \tilde{R}_L(s, \pi, a_f) \leq (1 - \phi(a_f|s))Z \; \forall s, a_f \tag{1g}$$

The objective 1a of the MINLP is to maximize the expected utility of the leader with respect to the distribution of initial states. The constraints 1b and 1c simply express the fact that the leader's stochastic policy must be a valid probability distribution over actions a_l in each state s. Similarly, constraints 1d and 1e ensure that the follower's policy is deterministic, choosing exactly one action in each state s. Constraints 1f are crucial, as they are used to compute the follower best response ϕ to a leader's policy π. These constraints contain two inequalities. The first represents the requirement that the follower value $V_F(s)$ in state s maximizes his expected utility over all possible choices a_f he can make in this state. The second constraint ensures that if an action a_f is chosen by ϕ in state s, $V_F(s)$ exactly equals the follower's expected utility in that state; if $\phi(a_f|s) = 0$, on the other hand, this constraint has no force, since the right-hand-side is just a large constant. Finally, constraints 1g are used to compute the leader's expected utility, given a follower best response. Thus, when the follower chooses a_f, the constraint on the right-hand-side will bind, and the leader's utility must therefore equal the expected utility when follower plays a_f. When $\phi(a_f|s) = 0$, on the other hand, the constraint has no force.

While the MINLP gives us an exact formulation for computing SSE in general SSGs, the fact that constraints 1f and 1g are not convex together with the integrality requirement on ϕ make it relatively impractical, at least given state-of-the-art MINLP solution methods. Below we therefore seek a principled approximation by discretizing the leader's continuous decision space.

Approximating Markov Stationary SSE

MILP Approximation What makes the MINLP formulation above difficult is the combination of integer variables, and the non-convex interaction between continuous variables π and V_F in one case (constraints 1f), and π and V_L in another (constraints 1g). If at least one of these variables is binary, we can linearize these constraints using McCormick inequalities (McCormick 1976). To enable the application of this technique, we discretize the probabilities which the leader's policy can use ((Ganzfried and Sandholm 2010) offer another linearization approach for approximating NE).

Let p_k denote a kth probability value and let $\mathcal{K} = \{1, \ldots, K\}$ be the index set of discrete probability values we use. Define binary variables $d_{s,k}^{a_l}$ which equal 1 if and only if $\pi(a_l|s) = p_k$, and 0 otherwise. We can then write $\pi(a_l|s)$ as $\pi(a_l|s) = \sum_{k \in \mathcal{K}} p_k d_{s,k}^{a_l}$ for all $s \in S$ and $a_l \in A_L$. Next, let $w_{s,k}^{a_l a_f} = d_{s,k}^{a_l} \sum_{s' \in S} T_{ss'}^{a_l a_f} V_L(s')$ for the leader, and let $z_{s,k}^{a_l a_f}$ be defined analogously for the follower. The key is that we can represent these equality constraints by the following equivalent McCormick inequalities, which we require to hold for all $s \in S$, $a_l \in A_L$, $a_f \in A_F$, and $k \in \mathcal{K}$:

$$w_{s,k}^{a_l a_f} \geq \sum_{s' \in S} T_{ss'}^{a_l a_f} V_L(s') - Z(1 - d_{s,k}^{a_l}) \tag{2a}$$

$$w_{s,k}^{a_l a_f} \leq \sum_{s' \in S} T_{ss'}^{a_l a_f} V_L(s') + Z(1 - d_{s,k}^{a_l}) \tag{2b}$$

$$-Z d_{s,k}^{a_l} \leq w_{s,k}^{a_l a_f} \leq Z d_{s,k}^{a_l}, \tag{2c}$$

and analogously for $z_{s,k}^{a_l a_f}$. Redefine follower's expected utility as $\tilde{R}_F(s, d, a_f, k) = \sum_{a_l \in A_L} \sum_{k \in \mathcal{K}} p_k \left(R_F(s, a_l, a_f) d_{s,k}^{a_l} - \gamma_F z_{s,k}^{a_l a_f} \right)$, with leader's expected utility $\tilde{R}_L(s, d, a_f, k)$ redefined similarly.

The full MILP formulation is then

$$\max_{\phi, V_L, V_F, z, w, d} \sum_{s \in S} \beta(s) V_L(s) \tag{3a}$$

$$\text{subject to :}$$

$$d^{a_l}_{s,k} \in \{0,1\} \qquad \forall s, a_l, k \tag{3b}$$

$$\sum_{k \in \mathcal{K}} d^{a_l}_{s,k} = 1 \qquad \forall s, a_l \tag{3c}$$

$$\sum_{a_l \in A_L} \sum_k p_k d^{a_l}_{s,k} = 1 \qquad \forall s \tag{3d}$$

$$0 \leq V_F(s) - \tilde{R}_F(s, d, a_f, k) \leq (1 - \phi(a_f|s))Z \forall s, a_f \tag{3e}$$

$$V_L(s) - \tilde{R}_L(s, d, a_f, k) \leq (1 - \phi(a_f|s))Z \; \forall s, a_f \tag{3f}$$

$$\text{constraints } 1d - 1e, \; 2a - 2c.$$

Constraints 3d, 3e, and 3f are direct analogs of constraints 1c, 1f, and 1g respectively. Constraints 3c ensure that exactly one probability level $k \in \mathcal{K}$ is chosen.

A Bound on the Discretization Error The MILP approximation above implicitly assumes that given a sufficiently fine discretization of the unit interval we can obtain an arbitrarily good approximation of SSE. In this section we obtain this result formally. First, we address why it is not in an obvious way related to the impact of discretization in the context of Nash equilibria. Consider a mixed Nash equilibrium s^* of an arbitrary normal form game with a utility function $u_i(\cdot)$ for each player i (extended to mixed strategies in a standard way), and suppose that we restrict players to choose a strategy that takes discrete probability values. Now, for every player i, let $\hat{s}_i$ be the closest point to s^*_i in the restricted strategy space. Since the utility function is continuous, this implies that each player's possible gain from deviating from $\hat{s}_i$ to s^*_i is small when all others play $\hat{s}_{-i}$, ensuring that finer discretizations lead to better Nash equilibrium approximation. The problem that arises in approximating an SSE is that we do not keep the follower's decision fixed when considering small changes to the leader's strategy; instead, we allow the follower to always optimally respond. In this case, the leader's expected utility can be discontinuous, since small changes in his strategy can lead to jumps in the optimal strategies of the follower if the follower is originally indifferent between multiple actions (a common artifact of SSE solutions). Thus, the proof of the discretization error bound is somewhat subtle.

First, we state the main result, which applies to all finite-action Stackelberg games, and then obtain a corollary which applies this result to our setting of discounted infinite-horizon stochastic games. Suppose that L and F are the finite sets of pure strategies of the leader and follower, respectively. Let $u_L(l, f)$ be the leader's utility function when the leader plays $l \in L$ and the follower plays $f \in F$, and suppose that X is the set of probability distributions over L (leader's mixed strategies), with $x \in X$ a particular mixed strategy with x_f the probability of playing a pure strategy $f \in F$. Let $\mathcal{P} = \{p_1, \ldots, p_K\}$ and let $\epsilon(\mathcal{P}) = \sup_{x \in X} \max_f \min_{k \in \mathcal{K}} |p_k - x_f|$. Suppose that $(x^*, f^{BR}(x^*))$ is a SSE of the Stackelberg game in which the leader can commit to an arbitrary mixed strategy $x \in X$. Let $U(x)$ be the leader's expected utility when he commits to $x \in X$.

Theorem 2. *Let $(x^{\mathcal{P}}, f^{BR}(x^{\mathcal{P}}))$ be an SSE where the leader's strategy x is restricted to $\mathcal{P}$. Then*

$$U(x^{\mathcal{P}}) \geq U(x^*) - \epsilon(\mathcal{P}) \max_{f \in F} \sum_l |u^L(l, f)|.$$

At the core of the proof is the multiple-LP approach for computing SSE (Conitzer and Sandholm 2006). The proof is provided in the Appendix.

The result in Theorem 2 pertains to general *finite-action* Stackelberg games. Here, we are interested in SSGs, where pure strategies of the leader and follower have, in general, arbitrarily infinite sequences of decisions. However, if we restrict attention to Markov stationary policies for the leader, we guarantee that the consideration set of the leader is finite, allowing us to apply Theorem 2.

Corollary 1. *In any SSG in which the leader is restricted to Markov stationary policies, the leader's expected utility in a SSE can be approximated arbitrarily well using discretized policies.*

Comparison Between MINLP and MILP

Above we asserted that the MINLP formulation is likely intractable given state-of-the-art solvers as motivation for introducing a discretized MILP approximation. We now support this assertion experimentally.

For the experimental comparison between the two formulations, we generate random stochastic games as follows. We fix the number of leader and follower actions to 2 per state and the discount factors to $\gamma_L = \gamma_F = 0.95$. We also restricted the payoffs of both players to depend only on state $s \in S$, but otherwise generated them uniformly at random from the unit interval, i.i.d. for each player and state. Moreover, we generated the transition function by first restricting state transitions to be non-zero on a predefined graph between states, and generated an edge from each s to another s' with probability $p = 0.6$. Conditional on there being an edge from s to s', the transition probability for each action tuple (a_l, a_f) was chosen uniformly at random from the unit interval.

	Exp Utility	Running Time (s)
MINLP (5 states)	9.83	375.26
MILP (5 states)	10.16	5.28
MINLP (6 states)	9.64	1963.53
MILP (6 states)	11.26	24.85

Table 1: Comparison between MINLP and MILP ($K = 5$), based on 100 random problem instances.

Table 1 compares the MILP formulation (solved using CPLEX) and MINLP (solved using KNITRO with 10 random restarts). The contrast is quite stark. First, even though MILP offers only an approximate solution, the actual solutions it produces are *better* than those that a state-of-the-art solver gets using MINLP. Moreover, MILP (using CPLEX) is more than 70 times faster when there are 5 states and nearly 80 times faster with 6 states. Finally, while MILP solved every instance generated, MINLP successfully found a feasible solution in only 80% of instances.

Extended Example: Patrolling the Newark Bay and New York Harbor

Consider again the example of patrolling the Newark Bay and New York Harbor under the geographic constraints shown in Figure 1. We now study the structure of defense policies in a variant of this example patrolling problem that is a deviation from the zero-sum assumption. This departure is motivated by the likely possibility that even though the players in security games are adversarial (we assume that the actual values of targets to both players are identical and as shown in the figure), they need not have the same degree of risk aversion. In our specific example, the departure from strict competitiveness comes from allowing the attacker (but not the defender) to be risk averse.

To model risk aversion, we filter the payoffs through the exponential function $f(u) = 1 - e^{-\alpha u}$, where u is the original payoff. This function is well known to uniquely satisfy the property of constant absolute risk aversion (CARA) (Gollier 2004). The lone parameter, α, controls the degree of risk aversion, with higher α implying more risk averse preferences.

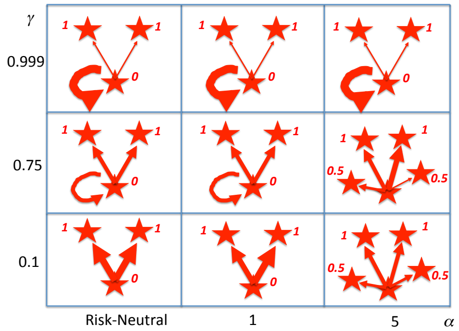

Figure 2: Varying discount factors $\gamma = \gamma_L = \gamma_F$ and the degree of risk aversion α.

In Figure 2 we report the relevant portion of the defense policy in the cross-product space of three discount factor values (0.1, 0.75, and 0.999) and three values of risk aversion (risk neutral, and $\alpha = 1$ and 5). We can make two qualitative observations. First, as the attacker becomes increasingly risk averse, the entropy of the defender's policy increases (i.e., the defender patrols a greater number of targets with positive probability). This observation is quite intuitive: if the attacker is risk averse, the defender can profitably increase the attacker's uncertainty, even beyond what would be optimal with a risk neutral attacker. Second, the impact of risk aversion diminishes as the players become increasingly patient. This is simply because a patient attacker is willing to wait a longer time before an attack, biding his time until the defender commits to one of the two most valued targets; this in turn reduces his exposure to risk, since he will wait to attack only when it is safe.

Conclusion

We defined general-sum discounted stochastic Stackelberg games (SSG). SSGs are of independent interest, but also generalize Stackelberg games which have been important in modeling security problems. We showed that there does not always exist a strong Stackelberg equilibrium in Markov stationary policies. We then provide a MINLP that solves for exact SSE restricted to Markov stationary policies, as well as a more tractable MILP that approximates it, and proved approximation bounds for the MILP. Finally, we illustrated how the generality of our SSGs can be used to address security problems without having to make limiting assumptions such as equal, or lack of, discount factors and identical player risk preferences.

Acknowledgments

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. Satinder Singh was supported by NSF grant IIS 0905146. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the views of the sponsors.

Appendix

Proof of Theorem 2

To prove this theorem, we leverage a particular technique for computing a SSE in finite-action games: one using multiple linear programs, one for each follower strategy $f \in F$ (Conitzer and Sandholm 2006). Each of these linear programs (LP) has the general form

$$\max_x \sum_{l \in L} x_l u^L(l, f)$$
$$s.t.$$
$$x \in \mathcal{D}(f),$$

where $\mathcal{D}(f)$ is the constraint set which includes the restriction $x \in X$ and requires that the follower's choice f is his optimal response to x. To compute the SSE, one then takes the optimal solution with the best value over the LPs for all $f \in F$; the corresponding f is the follower's best response. Salient to us will be a restricted version of these LPs, where we replace $\mathcal{D}(f)$ with $\mathcal{D}^\epsilon(f)$, where the latter requires, in addition, that leader's mixed strategies are restricted to $\mathcal{P}$ (note that $\mathcal{D}^\epsilon(f) \subseteq \mathcal{D}(f)$). Let us use the notation $P(f)$ to refer to the linear program above, and $P^\epsilon(f)$ to refer to the linear program with the restricted constraint set $\mathcal{D}^\epsilon(f)$. We also use P^ϵ to refer to the problem of computing the SSE in the restricted, discrete, setting.

We begin rather abstractly, by considering a pair of mathematical programs, P_1 and P_2, sharing identical linear objective functions $c^T x$. Suppose that X is the set of feasible solutions to P_1, while Y is the feasible set of P_2, and $Y \subseteq X \subseteq \mathbb{R}^m$. Let OPT_1 be the optimal value of P_1.

Lemma 2. *Suppose that $\forall x \in X$ there is $y \in Y$ such that $\|x - y\|_\infty \leq \epsilon$. Let $\hat{x}$ be an optimal solution to P_2. Then $\hat{x}$ is feasible for P_1 and $c^T\hat{x} \geq OPT_1 - \epsilon \sum_i |c_i|$.*

Proof. Feasibility is trivial since $Y \subseteq X$. Consider an arbitrary optimal solution x^* of P_1. Let $\tilde{x} \in Y$ be such that $\|x^* - \tilde{x}\|_\infty \leq \epsilon$; such $\tilde{x}$ must exist by the condition in the statement of the lemma. Then

$$c^T x^* - c^T \tilde{x} = \sum_i c_i(x_i^* - \tilde{x}_i) \leq |\sum_i c_i(x_i^* - \tilde{x}_i)|$$
$$\leq \sum_i |c_i||x_i^* - \tilde{x}_i| \leq \epsilon \sum_i |c_i|,$$

where the last inequality comes from $\|x^* - \tilde{x}\|_\infty \leq \epsilon$. Finally, since $\hat{x}$ is an optimal solution of P_2 and $\tilde{x}$ is P_2 feasible, $c^T\hat{x} \geq c^T\tilde{x} \geq c^T x^* - \epsilon\sum_i |c_i| = OPT_1 - \epsilon \sum_i |c_i|$. □

We can apply this Lemma directly to show that for a given follower action f, solutions to the corresponding linear program with discrete commitment, P_f^ϵ, become arbitrarily close to optimal solutions (in terms of objective value) of the unrestricted program P_f.

Corollary 2. *Let $OPT(f)$ be the optimal value of $P(f)$. Suppose that $x^\epsilon(f)$ is an optimal solution to $P^\epsilon(f)$. Then x^ϵ is feasible in $P(f)$ and*

$$\sum_{l \in L} x_l^\epsilon u^L(l, f) \geq OPT(f) - \epsilon \sum_l |u^L(l, f)|.$$

We now have all the necessary building blocks for the proof.

Proof of Theorem 2. Let $\hat{x}$ be a SSE strategy for the leader in the restricted, discrete, version of the Stackelberg commitment problem, P^ϵ. Let x^* be the leader's SSE strategy in the unrestricted Stackelberg game and let f^* be the corresponding optimal action for the follower (equivalently, the corresponding $P(f)$ which x^* solves). Letting $\hat{x}^{f^*}$ be the optimal solution to the restricted LP $P(f^*)^\epsilon$, we apply Corollary 2 to get

$$\sum_{l \in L} \hat{x}^{f^*} u^L(l, f^*) \geq OPT(f) - \epsilon \sum_l |u^L(l, f^*)|$$
$$= U(x^*) - \epsilon \sum_l |u^L(l, f^*)|,$$

where the last equality is due to the fact that x^* is both an optimal solution to Stackelberg commitment, and an optimal solution to $P(f^*)$.

Since $\hat{x}$ is optimal for the restricted commitment problem, and letting $\hat{f}$ be the corresponding follower strategy,

$$U(\hat{x}) = \sum_{l \in L} \hat{x}_l u^L(l, \hat{f}) \geq \sum_{l \in L} \hat{x}^{f^*} u^L(l, f^*)$$
$$\geq U(x^*) - \epsilon \sum_l |u^L(l, f^*)|$$
$$\geq U(x^*) - \epsilon \max_{f \in F} \sum_l |u^L(l, f)|.$$

□

References

Agmon, N.; Kraus, S.; and Kaminka, G. A. 2008. Multi-robot perimeter patrol in adversarial settings. In *IEEE International Conference on Robotics and Automation*, 2339–2345.

Agmon, N.; Urieli, D.; and Stone, P. 2011. Multiagent patrol generalized to complex environmental conditions. In *Twenty-Fifth National Conference on Artificial Intelligence*.

An, B.; Pita, J.; Shieh, E.; Tambe, M.; Kiekintveld, C.; and Marecki, J. 2011. Guards and protect: Next generation applications of security games. In *SIGECOM*, volume 10, 31–34.

Basilico, N., and Gatti, N. 2011. Automated abstraction for patrolling security games. In *Twenty-Fifth National Conference on Artificial Intelligence*, 1096–1099.

Basilico, N.; Rossignoli, D.; Gatti, N.; and Amigoni, F. 2010. A game-theoretic model applied to an active patrolling camera. In *International Conference on Emerging Security Technologies*, 130–135.

Basilico, N.; Gatti, N.; and Amigoni, F. 2009. Leader-follower strategies for robotic patrolling in environments with arbitrary topologies. In *Eighth International Conference on Autonomous Agents and Multiagent Systems*, 57–64.

Basilico, N.; Gatti, N.; and Villa, F. 2011. Asynchronous multirobot patrolling against intrusion in arbitrary topologies. In *Twenty-Forth National Conference on Artificial Intelligence*.

Bosansky, B.; Lisy, V.; Jakov, M.; and Pechoucek, M. 2011. Computing time-dependent policies for patrolling games with mobile targets. In *Tenth International Conference on Autonomous Agents and Multiagent Systems*, 989–996.

Conitzer, V., and Sandholm, T. 2006. Computing the optimal strategy to commit to. In *Seventh ACM conference on Electronic commerce*, 82–90.

Filar, J., and Vrieze, K. 1997. *Competitive Markov Decision Processes*. Springer-Verlag.

Ganzfried, S., and Sandholm, T. 2010. Computing equilibria by incorporating qualitative models. In *Nineth International Conference on Autonomous Agents and Multiagent Systems*, 183–190.

Gollier, C. 2004. *The Economics of Risk and Time*. The MIT Press.

Jain, M.; Kardes, E.; Kiekintveld, C.; Tambe, M.; and Ordonez, F. 2010a. Security games with arbitrary schedules: A branch and price approach. In *Twenty-Fourth National Conference on Artificial Intelligence*.

Jain, M.; Tsai, J.; Pita, J.; Kiekintveld, C.; Rathi, S.; Tambe, M.; and Ordóñez, F. 2010b. Software assistants for randomized patrol planning for the lax airport police and the federal air marshal service. *Interfaces* 40:267–290.

Kiekintveld, C.; Jain, M.; Tsai, J.; Pita, J.; Ordóñez, F.; and Tambe, M. 2009. Computing optimal randomized resource allocations for massive security games. In *Seventh International Conference on Autonomous Agents and Multiagent Systems*.

McCormick, G. 1976. Computability of global solutions to factorable nonconvex programs: Part I - convex underestimating problems. *Mathematical Programming* 10:147–175.

Paruchuri, P.; Pearce, J. P.; Marecki, J.; Tambe, M.; Ordonez, F.; and Kraus, S. 2008. Playing games with security: An efficient exact algorithm for Bayesian Stackelberg games. In *Proc. of The 7th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 895–902.

Vorobeychik, Y.; An, B.; and Tambe, M. 2012. Adversarial patrolling games. In *AAAI Spring Symposium on Security, Sustainability, and Health*.