Three Strategies to Success: Learning Adversary Models in Security Games

Nika Haghtalab
Carnegie Mellon U.
nhaghtal@cs.cmu.edu

Fei Fang
U. Southern California
feifang@usc.edu

Thanh H. Nguyen
U. Southern California
thanhhng@usc.edu

Arunesh Sinha
U. Southern California
aruneshs@usc.edu

Ariel D. Procaccia
Carnegie Mellon U.
arielpro@cs.cmu.edu

Milind Tambe
U. Southern California
tambe@usc.edu

Abstract

State-of-the-art applications of Stackelberg security games — including wildlife protection — offer a wealth of data, which can be used to learn the behavior of the adversary. But existing approaches either make strong assumptions about the structure of the data, or gather new data through online algorithms that are likely to play severely suboptimal strategies. We develop a new approach to learning the parameters of the behavioral model of a bounded rational attacker (thereby pinpointing a near optimal strategy), by observing how the attacker responds to only three defender strategies. We also validate our approach using experiments on real and synthetic data.

1 Introduction

The Stackelberg security game (SSG) model [Tambe, 2012], a well-established model for deployment of limited security resources, has recently been applied to assist agencies protecting wildlife, fisheries and forests. This *green security game* (GSG) research [Fang *et al.*, 2015; Nguyen *et al.*, 2015] differs from earlier work in SSGs applied to counter-terrorism [Pita *et al.*, 2009], as GSGs are accompanied with significant historical data (e.g., wildlife crime, arrests of poachers) from repeated defender-adversary interactions. Given this data, machine learning has now begun to play a critical role in improving defender resource allocation in GSGs, taking the place of domain experts as the leading method for estimating adversary utilities and preferences.

Indeed, inspired by GSGs, researchers have focused on learning adversary bounded rationality models and designing optimal defender strategies against such adversaries [Yang *et al.*, 2014; Sinha *et al.*, 2016]. While existing results are encouraging, as we explain in detail below (and as discussed by Sinha et al. [2016]), a shortcoming of the state-of-the-art approach is that its effectiveness implicitly relies on the properties of the underlying distribution from which data is obtained — if near-optimal strategies have not already been played, a near-optimal strategy cannot be learned.

1.1 Our Results

We take the next step in developing the theory and practice of learning adversary behavior. Our first contribution is a theoretical analysis of the learnability of (a generalization of) the most well-studied bounded rationality adversary behavior model in SSGs: Subjective Utility Quantal Response (SUQR) [Nguyen *et al.*, 2013], which is a parametric specification of the adversary's behavior. We find, perhaps surprisingly, that if the data contains a polynomial number of adversary responses to each of only *three* defender strategies that are sufficiently different from each other (precise statement is given in Theorem 3.1), then we can learn the model parameters with high accuracy. Qualitatively, this means that we can expect to learn an excellent strategy from real-world historical data, as typically each defender strategy is played repeatedly over a period of time. It also means that, even if we collect additional data in the future by playing whatever strategy seems good based on existing data, the new data will quickly lead to an *optimal* strategy, under the very mild assumption that the new strategies that are played are somewhat different from the previous ones.

Building on our analysis of the generalized SUQR model, as part of our second contribution, we analyze the learnability of the more general class of adversary behavior models specified by (non-parametric) Lipschitz functions. This is an extremely expressive class, and therefore, naturally, learning an appropriate model requires more data. Even for this class we can learn the adversary response function with high accuracy with a polynomial number of defender strategies — but we make more stringent assumptions regarding how these strategies are selected. Our analysis works by approximating Lipschitz functions using polynomial functions.

Finally, we conduct experiments to validate our approach. We show, via experiments on synthetic data, that a realistic number of samples for each of three strategies suffices to accurately learn the adversary model under generalized SUQR. We also show, using experiments on human subject data, that our main theoretical result provides guidance on selecting strategies to use for learning.

1.2 Related Work

Our work is most closely related to the recent paper of Sinha et al. [2016]. They learn to predict adversary responses in the PAC model of learning. Crucially, following the PAC model, the dataset is assumed to be constructed by drawing defender strategies (and adversary responses) from a fixed but unknown distribution — and the accuracy of the outcome of the learning process is then measured *with respect to that same distribution*. In particular, if the training data is concentrated in suboptimal regions of the defender strategy space, the PAC model approach allows accurate prediction of adversary responses in those regions, but may not help pinpoint a globally optimal strategy (as discussed at length by Sinha et al. [2016]). In contrast, our approach leads to *uniformly* accurate prediction, in the sense that we can accurately predict the adversary responses to *any* defender strategy, even if we only observe suboptimal strategies. As we show, this provides sufficient information to identify a near-optimal strategy.

Other prior work on learning adversary models using SUQR (or variants) using field data in GSGs has yielded practical guidance for generating defender strategies [Haskell *et al.*, 2014; Fang *et al.*, 2015], but has not provided any provable guarantees, nor an analysis of what types of defender strategies could yield improved learning results.

Another line of work [Letchford *et al.*, 2009; Marecki *et al.*, 2012; Blum *et al.*, 2014; Balcan *et al.*, 2015] explores an *online learning* approach, where an optimal strategy is learned by adaptively playing defender strategies and observing the adversary responses. This approach cannot make use of historical data. Moreover, in the most relevant papers [Letchford *et al.*, 2009; Blum *et al.*, 2014], in order to most efficiently pinpoint the optimal strategy, the algorithm may play severely suboptimal strategies, leading to potentially catastrophic losses. In contrast, in the online interpretation of our setting, we can learn an optimal strategy by playing *any* three (sufficiently different) strategies, including ones that seem optimal based on existing evidence. Interestingly, the reason why we are able to do so much better is that we assume a bounded rational adversary that responds probabilistically to defender strategies, while the foregoing papers assume a perfectly rational adversary. It would seem that in our more realistic setting, the learning task should only be harder. But by repeatedly playing a strategy, we gain information about the adversary's utility for all targets, whereas in the perfectly rational case, to gain information about new targets, the learning algorithm has to discover the "best response regions" of those targets.

2 Preliminaries

A Stackelberg security game is a two-player general-sum game between a *defender* and an *attacker*, where the defender commits to a randomized allocation of security resources to defend a set of targets. The attacker, in turn, observes this randomized allocation and responds to it by attacking a target. The defender and the attacker both receive payoffs depending on the target that was attacked and the probability that it was defended.

Formally, we consider a set $T = \{1, \ldots, n\}$ of n *targets*. The defender's action space is defined by a set of vectors $\mathcal{P} \subseteq [0,1]^n$, called the *coverage probability space*, where for each $\mathbf{p} \in \mathcal{P}$, p_t represents the probability with which the defender protects target t. Traditionally, this set is determined by defender's *security resources* and the subsets of targets that each resource can defend simultaneously. A *pure deployment* of the defender is then an assignment of resources to targets, and a *mixed deployment* of the defender is a distribution over pure deployments. In this context, the coverage probability vector induced by a mixed deployment is defined as the probability with which each target is defended under this mixed deployment. As we shall see, the behavior and utility of the attacker only depend on the defender's choice of coverage probabilities, and therefore we choose to represent the defender's action space by the set of coverage probabilities $\mathcal{P}$.

We denote the utility function of the attacker for target $t \in T$ by $u_t : [0,1] \to \mathbb{R}$. Given a coverage probability vector $\mathbf{p} \in \mathcal{P}$, the utility of the attacker under strategy $\mathbf{p}$ is defined as $u_t(p_t)$. Previous work on learning in security games has mainly focused on utilities that are linear functions of the coverage probability [Blum *et al.*, 2014; Marecki *et al.*, 2012], or linear functions with the additional constraint that $u_t(p_t) = wp_t + c_t$ in the generalized SUQR model of Sinha et al. [2016]. Our main result pertains to (unrestricted) linear utility functions, but later we also deal with higher degree polynomials.

Upon observing the defender's strategy $\mathbf{p}$, the attacker computes the utility on each target t, $u_t(p_t)$, and based on these utilities *responds* to the defender's strategy. In this paper, we consider a non-adaptive attacker who attacks target t with probability

$$D^{\mathbf{p}}(t) = \frac{e^{u_t(p_i)}}{\sum_{i \in T} e^{u_i(p_i)}}. \tag{1}$$

This model corresponds to the *Luce* model from quantal choice theory [Luce, 2005; McFadden, 1976], and is a special case of the *logit quantal response* model. Together with our choice of utility functions, our model is a generalization of bounded rationality models considered in previous work, such as SUQR [Nguyen *et al.*, 2013] and generalized SUQR [Sinha *et al.*, 2016].

Suppose the same mixed strategy $\mathbf{p}$ is played for multiple time steps. We denote the empirical distribution of attacks on target t under $\mathbf{p}$ by $\hat{D}^{\mathbf{p}}(\cdot)$. Furthermore, we assume that for the strategies considered in our work, and for all t, $D^{\mathbf{p}}(t) \geq \rho$ for some $\rho = 1/\mathrm{poly}(n)$. This assumption is required to estimate the value of $D^{\mathbf{p}}(t)$ with polynomially many samples.

Our goal is to learn the utility functions, $u_t(\cdot)$ for all $t \in T$, by observing attacker's responses to a choice of coverage probability vectors $\mathbf{p} \in \mathcal{P}$. This allows us to find an approximately optimal defender strategy — the strategy that leads to the best defender utility. We say that $\hat{u}_t : [0,1] \to \mathbb{R}$ *uniformly approximates* or *uniformly learns* $u_t(\cdot)$ within an error of ϵ, if $\forall x \in [0,1],\ |\hat{u}_t(x) - u_t(x)| \leq \epsilon$. Note that the attacker's mixed strategy remains the same when the utility functions corresponding to all targets are increased by the same value. Therefore, we can only hope to learn a "normalized" representation of the utility functions, $\hat{u}_t$, such that for

all t and all x, $|\hat{u}_t(x) + c - u_t(x)| \leq \epsilon$ for some c. Technically, this is exactly what we need to predict the behavior of the attacker. We use this fact in the proof of our main theorem and choose an appropriate normalization that simplifies the presentation of the technical details.

3 Theoretical Results

In this section, we present our theoretical results for learning attacker utility functions. We first state our results in terms of linear utility functions and show that it is possible to uniformly learn the utilities up to error ϵ using only 3 randomized strategies, with $\text{poly}(n, \frac{1}{\epsilon})$ samples for each, under mild conditions. We view these results as practically significant.

In Section 3.2, we extend the results to polynomials of degree d to represent a larger class of utility functions. We show (in Section 3.3) that this allows us to learn the even more expressive class of Lipschitz utility functions. The extension to high-degree polynomials and Lipschitz utilities requires more restrictive conditions, and hence it is of greater theoretical than practical interest.

Finally, in Section 3.4, we show that accurately learning the attacker's utility function allows us to predict the distribution of attacker responses to any defender strategy, and, therefore, to pinpoint a near-optimal strategy.

3.1 Linear Utility Functions

Assume that the utility functions are linear and denoted by $u_t(x) = w_t x + c_t$. As discussed in Section 2, we can normalize the utilities; That is, without loss of generality $c_n = 0$. Our main result is the following theorem.

Theorem 3.1. *Suppose the functions $u_1(\cdot), \ldots, u_n(\cdot)$ are linear. Consider any 3 strategies, $\mathbf{p}, \mathbf{q}, \mathbf{r} \in \mathcal{P}$, such that for any $t < n$, $|(p_t - q_t)(p_n - r_n) - (p_n - q_n)(p_t - r_t)| \geq \lambda$, and for any two different strategies $\mathbf{x}, \mathbf{y} \in \{\mathbf{p}, \mathbf{q}, \mathbf{r}\}$, we have $|x_t - y_t| \geq \nu$. If we have access to $m = \Omega(\frac{1}{\rho}(\frac{1}{\epsilon\nu\lambda})^2 \log(\frac{n}{\delta}))$ samples of each of these strategies, then with probability $1-\delta$, we can uniformly learn each $u_t(\cdot)$ within error ϵ.*

We view the assumptions as being very mild. Indeed, intuitively ν depends on how different the strategies are from each other — a very small value means that they are almost identical on some coordinates. The lower bound of λ is less intuitive, but again, it would not be very small unless there is a very specific relation between the strategies. As a sanity check, if the three strategies were chosen uniformly at random from the simplex, both values would be at least $1/\text{poly}(n)$.

To gain some intuition before proceeding with the proof, note that in the quantal best-response model, for each strategy $\mathbf{p}$, the ratio between the attack probabilities of two targets t and n follows the relation

$$u_t(p_t) = \ln\left(\frac{D^{\mathbf{p}}(t)}{D^{\mathbf{p}}(n)}\right) + u_n(p_n). \tag{2}$$

Therefore, each strategy induces $n-1$ linear equations that can be used to solve for the coefficients of u_t. However, we can only obtain an *estimate* $\hat{D}^{\mathbf{p}}(t)$ of the probability that target t is attacked under a strategy $\mathbf{p}$, based on the given samples. So, the inaccuracy in our estimates of $\ln(\hat{D}^{\mathbf{p}}(t)/\hat{D}^{\mathbf{p}}(n))$ leads to inaccuracy in the estimated polynomial $\hat{u}_t$. For sufficiently accurate estimates $\hat{D}^{\mathbf{p}}(t)$, we show that the value of u_t differs from the true value by at most ϵ.

Let us first analyze the rate of convergence of $\hat{D}^{\mathbf{p}}(t)$ to $D^{\mathbf{p}}(t)$ as the number of observations of strategy $\mathbf{p}$ increases.

Lemma 3.2. *Given $\mathbf{p} \in \mathcal{P}$, let $\hat{D}^{\mathbf{p}}(t)$ be the empirical distribution of attacks based on $m = \Omega(\frac{1}{\rho\epsilon^2}\log(\frac{n}{\delta}))$ samples. With probability $1-\delta$, for all $t \in T$, $\frac{1}{1+\epsilon} \leq \hat{D}^{\mathbf{p}}(t)/D^{\mathbf{p}}(t) \leq 1+\epsilon$.*

Proof. Given $\mathbf{p} \in \mathcal{P}$ and $t \in T$, let $X_1, \ldots, X_m$ be Bernoulli random variables, whose value is 1 if and only if target t is attacked in sample i, under strategy $\mathbf{p}$. These are i.i.d. random variables with expectation $D^{\mathbf{p}}(t)$. Furthermore, $\hat{D}^{\mathbf{p}}(t) = \frac{1}{m}\sum_i X_i$. Therefore, using the Chernoff bound, we have

$$\Pr\left[\frac{1}{1+\epsilon} \leq \frac{\hat{D}^{\mathbf{p}}(t)}{D^{\mathbf{p}}(t)} \leq 1+\epsilon\right] \geq 1 - 2e^{-mD^{\mathbf{p}}(t)\epsilon^2/4}.$$

Since $D^{\mathbf{p}}(t) > \rho$, when $m = \Omega(\frac{1}{\rho\epsilon^2}\log(\frac{n}{\delta}))$, with probability $1 - \frac{\delta}{n}$, $\frac{1}{1+\epsilon} \leq \hat{D}^{\mathbf{p}}(t)/D^{\mathbf{p}}(t) \leq 1+\epsilon$. Taking the union bound over all $t \in T$, with probability $1-\delta$, for all $t \in T$, $\frac{1}{1+\epsilon} \leq \hat{D}^{\mathbf{p}}(t)/D^{\mathbf{p}}(t) \leq 1+\epsilon$. □

Proof of Theorem 3.1. By Equation 2 and using our assumption that $c_n = 0$, for all $t \in T$, $w_t p_t + c_t = \ln\frac{D^{\mathbf{p}}(t)}{D^{\mathbf{p}}(n)} + w_n p_n$. Using the same equation for $\mathbf{q}$ and eliminating c_t, we have

$$w_t(p_t - q_t) = \ln\frac{D^{\mathbf{p}}(t)}{D^{\mathbf{p}}(n)} - \ln\frac{D^{\mathbf{q}}(t)}{D^{\mathbf{q}}(n)} + w_n(p_n - q_n).$$

Repeating the above for $\mathbf{p}$ and $\mathbf{r}$ and solving for w_n, we have

$$w_n = \frac{(p_t - r_t)\ln\frac{D^{\mathbf{p}}(t)D^{\mathbf{q}}(n)}{D^{\mathbf{q}}(t)D^{\mathbf{p}}(n)} - (p_t - q_t)\ln\frac{D^{\mathbf{p}}(t)D^{\mathbf{r}}(n)}{D^{\mathbf{r}}(t)D^{\mathbf{p}}(n)}}{(p_t - q_t)(p_n - r_n) - (p_n - q_n)(p_t - r_t)}. \tag{3}$$

Furthermore, for all $t < n$,

$$w_t = \frac{\ln\frac{D^{\mathbf{p}}(t)}{D^{\mathbf{p}}(n)} - \ln\frac{D^{\mathbf{q}}(t)}{D^{\mathbf{q}}(n)} + w_n(p_n - q_n)}{p_t - q_t} \tag{4}$$

and

$$c_t = \ln\frac{D^{\mathbf{p}}(t)}{D^{\mathbf{p}}(n)} + w_n p_n - w_t p_t \tag{5}$$

Let $\hat{w}_t$ and $\hat{c}_t$ be defined similarly to w_t and c_t but in terms of the estimates $\hat{D}^{\mathbf{p}}(t)$. By Lemma 3.2, for strategy $\mathbf{p}$ (and similarly $\mathbf{q}$ and $\mathbf{r}$) and any t, we have $\frac{1}{1+\epsilon'} \leq \frac{D^{\mathbf{p}}(t)}{\hat{D}^{\mathbf{p}}(t)} \leq 1+\epsilon'$ for $\epsilon' = \epsilon\lambda\nu/128$. Therefore, we have

$$|w_n - \hat{w}_n| = \frac{\left|(p_t - r_t)\ln\frac{D^{\mathbf{p}}(t)D^{\mathbf{q}}(n)\hat{D}^{\mathbf{q}}(t)\hat{D}^{\mathbf{p}}(n)}{\hat{D}^{\mathbf{p}}(t)\hat{D}^{\mathbf{q}}(n)D^{\mathbf{q}}(t)D^{\mathbf{p}}(n)} - (p_t - q_t)\ln\frac{D^{\mathbf{p}}(t)D^{\mathbf{r}}(n)\hat{D}^{\mathbf{r}}(t)\hat{D}^{\mathbf{p}}(n)}{\hat{D}^{\mathbf{p}}(t)\hat{D}^{\mathbf{r}}(n)D^{\mathbf{r}}(t)D^{\mathbf{p}}(n)}\right|}{|(p_t - q_t)(p_n - r_n) - (p_n - q_n)(p_t - r_t)|}$$

$$\leq \frac{|p_t - r_t|\ln(1+\epsilon')^4 + |p_t - q_t|\ln(1+\epsilon')^4}{|(p_t - q_t)(p_n - r_n) - (p_n - q_n)(p_t - r_t)|} \leq 8\frac{\epsilon'}{\lambda} \leq \epsilon/16,$$

where the third transition follows from the well-known fact that $\ln(1+x) \leq x$ for all $x \in \mathbb{R}$. Similarly, for $t < n$, we have

$$|w_t - \hat{w}_t| = \frac{\left|\ln \frac{D^{\mathbf{p}}(t)\hat{D}^{\mathbf{p}}(n)}{\hat{D}^{\mathbf{p}}(t)D^{\mathbf{p}}(n)} - \ln \frac{D^{\mathbf{q}}(t)\hat{D}^{\mathbf{q}}(n)}{\hat{D}^{\mathbf{q}}(t)D^{\mathbf{q}}(n)} + (w_n - \hat{w}_n)(p_n - q_n)\right|}{|p_t - q_t|}$$
$$\leq \frac{1}{\nu}(4\epsilon' + \epsilon/16) \leq \epsilon/8.$$

And,

$$|c_t - c_t| = \left|\ln \frac{D^{\mathbf{p}}(t)\hat{D}^{\mathbf{p}}(n)}{\hat{D}^{\mathbf{p}}(t)D^{\mathbf{p}}(n)} + (w_n - \hat{w}_n)p_n - (w_t - \hat{w}_t)p_t\right|$$
$$\leq 2\epsilon' + \epsilon/4 \leq \epsilon/2.$$

Therefore, for any t and any $x \in [0,1]$, $|u_t(x) - \hat{u}_t(x)| \leq \epsilon$. □

3.2 Polynomial Utility Functions

On the way to learning Lipschitz utilities, we next assume that the utility function is a polynomial of degree at most d (linear functions are the special case of $d = 1$). We show that it is possible to learn these the utility functions using $O(d)$ strategies.

Theorem 3.3. *Suppose the functions $u_1(\cdot), \ldots, u_n(\cdot)$ are polynomials of degree at most d. Consider any $2d+1$ strategies, $\mathbf{q}^{(1)}, \ldots, \mathbf{q}^{(d)}$, $\mathbf{q}^{(d+1)} = \mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(d+1)}$, such that for all $k, k', k \neq k'$, $q_1^{(k)} = q_1^{(k')}$, $p_n^{(k)} = p_n^{(k')}$, $|q_n^{(k)} - q_n^{(k')}| \geq \nu$, and for all $t < n$, $|p_t^{(k)} - p_t^{(k')}| \geq \nu$. If we have access to $m = \Omega(\frac{1}{\rho}(\frac{d}{\epsilon\nu^d})^2 \log(\frac{n}{\delta}))$ samples of each of these strategies, then with probability $1-\delta$, we can uniformly learn each $u_t(\cdot)$ within error ϵ.*

It is important to emphasize that, unlike Theorem 3.1, one would not expect historical data to satisfy the conditions of Theorem 3.3, because it requires different strategies to cover some targets with the exact same probability. It is therefore mostly useful in a setting where we have control over which strategies are played. Strictly speaking, these more stringent conditions are not necessary for learning polynomials, but we enforce them to obtain a solution that is stable against inaccurate observations. Also note that the $d = 1$ case of Theorem 3.3 is weaker and less practicable than Theorem 3.1, because the latter theorem uses tailor-made arguments that explicitly leverage the structure of linear functions.

In a nutshell, the theorem's proof relies on polynomial interpolation. Specifically, consider the relationship between the utility functions of different targets shown in Equation (2). We assume that all the strategies $\mathbf{p}^{(i)}$ have the same coverage probability p_n on target n; since subtracting a fixed constant from all utility functions leaves the distribution of attacks unchanged, we can subtract $u_n(p_n)$ and assume without loss of generality that

$$\forall t < n, \quad u_t(p_t) = \ln\left(\frac{D^{\mathbf{p}}(t)}{D^{\mathbf{p}}(n)}\right). \tag{6}$$

Because u_t is a polynomial of degree d, it can be found by solving for the *unique* degree d polynomial that matches the values of u_t at $d+1$ points. To learn u_n, we can then use the same approach with the exception of using the utility function for targets $1, \ldots, n-1$ in Equation (2) to get the value of $u_n(\cdot)$ on $d+1$ points. As before, we do not have access to the *exact* values of $D^{\mathbf{p}}(t)$, so we use the estimated values $\hat{D}^{\mathbf{p}}(t)$ in these equations.

The next well-known lemma states the necessary and sufficient conditions for existence of a unique degree d polynomial that fits a collection of $d+1$ points [Gautschi, 1962].

Lemma 3.4. *For any values $y_1, \ldots, y_d$ and $x_1, \ldots, x_d$ such that $x_i \neq x_j$ for all $i \neq j$, there is a unique polynomial $f : \mathbb{R} \to \mathbb{R}$ of degree d, such that for all i, $f(x_i) = y_i$. Furthermore, this polynomial can be expressed as*

$$f(x) = \sum_{k=1}^{d+1} y_k \prod_{k':k' \neq k} \frac{x - x_{k'}}{x_k - x_{k'}}. \tag{7}$$

Proof of Theorem 3.3. Let $\hat{y}_t^{(k)} = \ln(\hat{D}^{\mathbf{p}^{(k)}}(t)/\hat{D}^{\mathbf{p}^{(k)}}(n))$ for all $t < n$. We have assumed that $p_t^{(k)} \neq p_t^{(k')}$ for any $k \neq k'$, so the conditions of Lemma 3.4 hold with respect to the pairs $\left(p_t^{(k)}, \hat{y}_t^{(k)}\right)$. Let $\hat{u}_t$ be the unique polynomial described by Equation (7), i.e.,

$$\hat{u}_t(x) = \sum_{k=1}^{d+1} \hat{y}_t^{(k)} \prod_{k':k' \neq k} \frac{x - p_t^{(k')}}{p_t^{(k)} - p_t^{(k')}}.$$

Similarly, for the values $y_t^{(k)} = \ln(D^{\mathbf{p}^{(k)}}(t)/D^{\mathbf{p}^{(k)}}(n))$, by Lemma 3.4 and Equation (6), $u_t(x)$ can be expressed by

$$u_t(x) = \sum_{k=1}^{d+1} y_t^{(k)} \prod_{k':k' \neq k} \frac{x - p_t^{(k')}}{p_t^{(k)} - p_t^{(k')}}.$$

Let ϵ' be such that $\epsilon = 4\epsilon'(d+1)/\nu^d$. By Lemma 3.2 for strategy $\mathbf{p}^{(k)}$ and any t, we have $\frac{1}{1+\epsilon'} \leq \frac{D^{\mathbf{p}^{(k)}}(t)}{\hat{D}^{\mathbf{p}^{(k)}}(t)} \leq 1 + \epsilon'$. Using the fact that $\ln(1+x) \leq x$ for all $x \in \mathbb{R}$, with probability $1-\delta$,

$$|\hat{y}_t^{(k)} - y_t^{(k)}| = \left|\ln \frac{\hat{D}^{\mathbf{p}^{(k)}}(t)}{D^{\mathbf{p}^{(k)}}(t)} - \ln \frac{\hat{D}^{\mathbf{p}^{(1)}(n)}}{D^{\mathbf{p}^{(1)}(n)}}\right| \leq 2\epsilon'.$$

Therefore, for all x and all $t < n$,

$$|\hat{u}_t(x) - u_t(x)| = \left|\sum_{k=1}^{d+1} (\hat{y}_t^{(k)} - y_t^{(k)}) \prod_{k' \neq k} \frac{x - p_t^{(k')}}{p_t^{(k)} - p_t^{(k')}}\right|$$
$$\leq 2\epsilon' \frac{d+1}{\nu^d} \leq \epsilon/2.$$

Similarly, by Equation (2) for target n and $\mathbf{q}^{(k)}$, we have,

$$u_n(q_n^{(k)}) = \ln\left(\frac{D^{\mathbf{q}^{(k)}}(n)}{D^{\mathbf{q}^{(k)}}(1)}\right) + u_1(q_1^{(k)})$$

Since for all k, $q_1^{(k)} = q_1$, using Lemma 3.4, u_n can be described by the unique polynomial passing through points $\left(q_n^{(k)}, \ln \frac{D^{\mathbf{q}^{(k)}}(n)}{D^{\mathbf{q}^{(k)}}(1)}\right)$ translated by the value $u^{(1)}(q_1)$. Similarly, let $\hat{u}^{(1)}$ be defined by the unique polynomial passing

through points $\left(q_n^{(k)}, \ln \frac{\hat{D}^{\mathbf{q}^{(k)}}(n)}{\hat{D}^{\mathbf{q}^{(k)}}(1)}\right)$ translated by the value $\hat{u}_1(q_1)$, then

$$|\hat{u}_n(x) - u_n(x)| \leq |\hat{u}_1(q_1) - u_1(q_1)| + \left|\sum_{k=1}^{d+1}\left(\ln \frac{\hat{D}^{\mathbf{q}^{(k)}}(n)}{\hat{D}^{\mathbf{q}^{(k)}}(1)} - \ln \frac{D^{\mathbf{q}^{(k)}}(n)}{D^{\mathbf{q}^{(k)}}(1)}\right) \prod_{k':k'\neq k} \frac{x - q_n^{(k')}}{q_n^{(k)} - q_n^{(k')}}\right| \leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

This completes our proof. □

3.3 Lipschitz Utilities

We now leverage the results of Section 3.2 to learn any utility function that is continuous and L-Lipschitz, i.e., for all t and values x and y, $|u_t(x) - u_t(y)| \leq L|x - y|$. We argue that such utility functions can be uniformly learned up to error ϵ, using $O(\frac{L}{\epsilon})$ strategies.

To see this, we first state a result that shows that all L-Lipschitz functions can be uniformly approximated within error ϵ using polynomials of degree $O(\frac{L}{\epsilon})$ [French *et al.*, 2003].

Lemma 3.5. *Let $\mathcal{F}_m$ be a family of degree m polynomials defined over $[-1, 1]$, and let $\mathcal{F}$ be the set of all L-Lipschitz continuous functions over $[-1, 1]$. Then, for all $f \in \mathcal{F}$,*

$$\inf_{g\in\mathcal{F}_m} \sup_x |f(x) - g(x)| \leq \frac{6L}{m}.$$

Therefore, for any L-Lipschitz function $u_t(x)$, there is a polynomial of degree $m = 12L/\epsilon$ that uniformly approximates $u_t(x)$ within error of $\epsilon/2$. By applying Theorem 3.3 to learn polynomials of degree $12L/\epsilon$, we can learn all the utility functions using $O(L/\epsilon)$ strategies.

Corollary 3.6. *Suppose the functions $u_1(\cdot), \ldots, u_n(\cdot)$ are L-Lipschitz. For $d = 12L/\epsilon$, consider any $2d + 1$ strategies, $\mathbf{q}^{(1)}, \ldots, \mathbf{q}^{(d)}$, $\mathbf{q}^{(d+1)} = \mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(d+1)}$, such that for all $k, k', k \neq k'$, $q_1^{(k)} = q_1^{(k')}$, $p_n^{(k)} = p_n^{(k')}$, $|q_n^{(k)} - q_n^{(k')}| \geq \nu$, and for all $t < n$, $|p_t^{(k)} - p_t^{(k')}| \geq \nu$. If we have access to $m = \Omega(\frac{L^2}{\rho\epsilon^4\nu^{24L/\epsilon}} \log(\frac{n}{\delta}))$ samples of each of these strategies, then with probability $1-\delta$, we can uniformly learn each $u_t(\cdot)$ within error ϵ.*

3.4 Learning the Optimal Strategy

So far, we have focused on the problem of uniformly learning the utility function of the attacker. We now show that an accurate estimate of this utility function allows us to pinpoint an almost optimal strategy for the defender.

Let the utility function of the defender on target $t \in T$ be denoted by $v_t : [0, 1] \to [-1, 1]$. Given a coverage probability vector $\mathbf{p} \in \mathcal{P}$, the utility the defender receives when target t is attacked is $v_t(p_t)$. The overall expected utility of the defender is

$$\mathcal{V}(\mathbf{p}) = \sum_{t\in T} D^{\mathbf{p}}(t) v_t(p_t).$$

Let $\hat{u}_t$ be the learned attacker utility functions, and $\bar{D}^{\mathbf{p}}(t)$ be the predicted attack probability on target t under strategy $\mathbf{p}$, according to the utilities $\hat{u}_t$, i.e.,

$$\bar{D}^{\mathbf{p}}(t) = \frac{e^{\hat{u}_t(p_i)}}{\sum_{i\in T} e^{\hat{u}_i(p_i)}}.$$

Let $\bar{\mathcal{V}}(\mathbf{p})$ be the predicted expected utility of the defender based on the learned attacker utilities $\bar{D}^{\mathbf{p}}(t)$, that is, $\bar{\mathcal{V}}(\mathbf{p}) = \sum_{t\in T} \bar{D}^{\mathbf{p}}(t) v_t(p_t)$. We claim that when the attacker utilities are uniformly learned within error ϵ, then $\bar{\mathcal{V}}$ estimates $\mathcal{V}$ with error at most 8ϵ. At a high level, this is established by showing that one can predict the attack distribution using the learned attacker utilities. Furthermore, optimizing the defender's strategy against the approximate attack distributions leads to an approximately optimal strategy for the defender.[1]

Theorem 3.7. *Assume for all $\mathbf{p}$ and any $t \in T$, $|\hat{u}_t(p_t) - u_t(p_t)| \leq \epsilon \leq 1/4$. Then, for all $\mathbf{p}$, $|\bar{\mathcal{V}}(\mathbf{p}) - \mathcal{V}(\mathbf{p})| \leq 4\epsilon$. Furthermore, let $\mathbf{p}' = \arg\max_{\mathbf{p}} \bar{\mathcal{V}}(\mathbf{p})$ be the predicted optimal strategy, then $\max_{\mathbf{p}} \mathcal{V}(\mathbf{p}) - \mathcal{V}(\mathbf{p}') \leq 8\epsilon$.*

4 Experimental Results

We conducted extensive experiments based on both synthetic data and real data collected through human subject experiments. The results not only support our theoretical analysis, but also verify the practical implications of our theorems for guiding the learning process. In the experiments, we focus on linear utility functions of the form $u_t(x) = w_t x + c_t$, and learn $u_t(\cdot)$ using closed-form equations (3)–(5) as stated in the proof of Theorem 3.1. We refer to this learning approach as Closed-Form Estimation (CFE).

Theorem 3.1 asserts that one can learn $u_t(\cdot)$ using three strategies with a polynomial number of attack samples. In the first set of experiments, we aim to verify empirically how many attack samples are needed to learn $u_t(\cdot)$ and predict the attack distribution with high accuracy using CFE. To test the performance with a wide range of sample sizes, we use synthetic data. More specifically, we randomly generated 50 different sets of true values of the utility function's parameters (w_t, c_t) from a uniform distribution on $[-3, -1]$ and $[1, 3]$, respectively. For each set of true parameter values, we generated 20 different sets of three defender strategies, each with a total coverage probability of 3 (i.e., there are three defender resources). The three defender strategies in each set are chosen to be sufficiently different such that the minimum difference in coverage probability between the strategies is at least 0.12. The attack samples are drawn from corresponding attack probability distributions with the number of samples ranging from 500 to 10000. For each set of defender strategies and corresponding attack samples, we learned the functions $\hat{u}_t(\cdot)$ using CFE. We generated a test set that consists of 1000 uniformly-distributed random defender strategies, and computed the error ϵ as the maximum difference between $\hat{u}_t(\cdot)$ and $u_t(\cdot)$ on the test set. We also report the error in predicting the attack distribution, measured by the ∞-norm distance between the predicted and the true attack distribution. All of our results are statistically significant (bootstrap-t method with $p < 0.05$) unless otherwise specified.

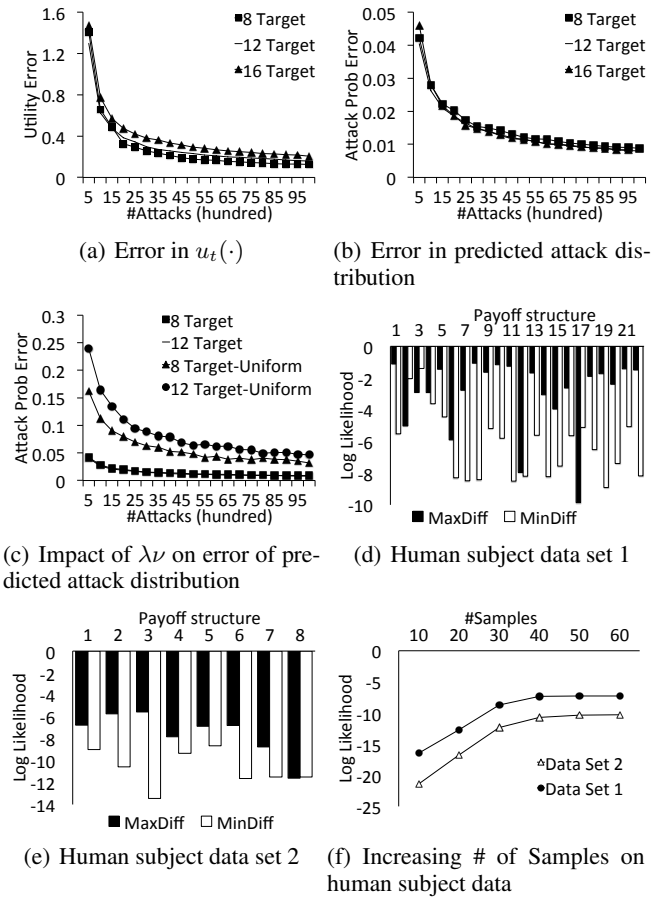
[1]The proof of this theorem appears in the full version of the paper, available at http://www.cs.cmu.edu/~nhaghtal/pubs/3strategies.pdf.

Figure 1(a) and 1(b) show the experimental results with different number of targets: $n = 8, 12, 16$. The x-axis indicates the number of attack samples, and the y-axis is the average error in predicting $u_t(\cdot)$ and the attack distribution. The error significantly decreases when the number of attack samples increases. Moreover, when more than 2000 samples are provided for each of the three defender strategies, the error in predicting the attack distribution is low ($|| \cdot ||_\infty \leq 0.02$). In the wildlife protection domain, more than one thousand snares can be found in a year in some areas [Hashimoto *et al.*, 2007], where each snare is an attack sample. Therefore, these results empirically show that it is possible to predict the attack distribution with high accuracy, based on historical data in domains such as wildlife protection, as long as the deployed defender strategies are sufficiently different.

To understand how the difference between strategies can impact the prediction error, we compare the error of learning from strategies that are sufficiently different to that of learning from uniformly-distributed randomly chosen strategies. The prediction accuracy is shown in Figure 1(c), which is averaged over 50 different samples of the three strategies. The error decreases at a much slower rate when randomly chosen strategies are used. This conclusion is consistent with Theorem 3.1, which states that the number of samples needed depends on the value of $\lambda\nu$, where $\lambda\nu$ can be seen as an indicator of how different the strategies are. In this set of experiments, by using randomly chosen strategies instead of strategies that are sufficiently different, the average value of $\lambda\nu$ decreases from 0.498 to 0.045 for 8 targets and from 0.147 to 0.017 for 16 targets. These results also imply that if we carefully choose the defender strategies to learn from, far fewer samples would be needed to achieve high prediction accuracy.

In addition to testing on synthetic data, we tested on human subject data collected by Nguyen et al. [2013] (data set 1) and Kar et al. [2015] (data set 2). We did not run statistical tests in this experiment since the data is limited. We aim to examine the impact of $\lambda\nu$ on the prediction accuracy with human subject data. We tested on 22 different payoff structures in data set 1 and 8 different payoff structures in data set 2. There are five different strategies of the defender associated with each payoff structure. In both data sets, the number of attack samples is small (30-50 samples for each game) relative to the number of targets (8 targets for data set 1 and around 25 targets for data set 2). For each payoff structure, we selected two sets of three strategies such that $\lambda\nu$ is maximized and minimized, respectively. We consider these two sets as the two different training sets for learning the utility parameters, w_t and c_t. Since the true values of the utility parameters are unknown and the true attack distribution is also unknown due to the lack of attack samples, we report the log likelihood of the observed attack samples averaged over all five strategies.

The results are plotted in Figure 1(d) and 1(e), in which the x-axis indicates different payoff structures, and the y-axis is the average log likelihood value. In data set 1, the set of strategies with maximum $\lambda\nu$ difference obtains higher average log-likelihood values in 19 out of 22 payoff structures. In data set 2, the set of strategies with maximum $\lambda\nu$ difference obtains higher average log likelihood values in 7 out of 8 payoff structures. These results indicate that a larger difference



(a) Error in $u_t(\cdot)$

(b) Error in predicted attack distribution

(c) Impact of $\lambda\nu$ on error of predicted attack distribution

(d) Human subject data set 1

(e) Human subject data set 2

(f) Increasing # of Samples on human subject data

Figure 1: Experimental results.

between the three chosen strategies typically leads to a higher prediction accuracy. As an implication of Theorem 3.1 that is verified by experimental results, this conclusion again provides guidance on selecting the defender's strategies during the learning phase.

With the set of strategies that maximize $\lambda\nu$, we tested how the number of samples affects learning. Figure 1(f) reports the log likelihood value averaged over 1000 sets of randomly selected samples. Not surprisingly, the log likelihood value increases as the sample size grows and the increase from 10 to 30 samples is significant. Note that in some games we have less than 40 samples in total, so the increase from 40 to 60 samples is not significant.

5 Discussion

The human subject data we used in our experiments is convenient, because we know which randomized defender strategy the subjects are responding to. In contrast, in the wildlife protection domain, the relevant historical data is not as neat: we can only observe the rangers' actual patrols (pure deployments), and the corresponding attacks. Therefore, in order to apply SSG-based learning techniques, the patrols must be artificially associated with different randomized strategies. For example, patrols in a single winter can be seen as instantia-

tions of a single randomized strategy.

Despite this difficulty, our results can give guidance for learning. They suggest that partitioning the historical data into only a few strategies, with many samples for each, might be best. And in the future, when new data is collected from deploying randomized strategies, our results will be even more useful, as they impose extremely mild conditions on deployed strategies, which allow learning optimal strategies.

Acknowledgments

This work was supported in part by MURI grant W911NF-11-1-0332; NSF grants CCF-1215883, CCF-1525932, and IIS-1350598; a Sloan Research Fellowship; and an IBM Ph.D. Fellowship.

References

[Balcan *et al.*, 2015] M.-F. Balcan, A. Blum, N. Haghtalab, and A. D. Procaccia. Commitment without regrets: Online learning in Stackelberg security games. In *Proceedings of the 16th ACM Conference on Economics and Computation (EC)*, pages 61–78, 2015.

[Blum *et al.*, 2014] A. Blum, N. Haghtalab, and A. D. Procaccia. Learning optimal commitment to overcome insecurity. In *Proceedings of the 28th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 1826–1834, 2014.

[Fang *et al.*, 2015] F. Fang, P. Stone, and M. Tambe. When security games go green: Designing defender strategies to prevent poaching and illegal fishing. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2589–2595, 2015.

[French *et al.*, 2003] M. French, C. Szepesvári, and E. Rogers. *Performance of nonlinear approximate adaptive controllers*. John Wiley & Sons, 2003.

[Gautschi, 1962] W. Gautschi. On inverses of vandermonde and confluent vandermonde matrices. *Numerische Mathematik*, 4(1):117–123, 1962.

[Hashimoto *et al.*, 2007] C. Hashimoto, D. Cox, and T. Furuichi. Snare removal for conservation of chimpanzees in the Kalinzu forest reserve, Uganda. *Pan Africa News*, 14(1):8–11, 2007.

[Haskell *et al.*, 2014] W. Haskell, D. Kar, F. Fang, M. Tambe, S. Cheung, and E. Denicola. Robust protection of fisheries with COmPASS. In *Proceedings of the 26th Annual Conference on Innovative Applications of Artificial Intelligence (IAAI)*, pages 2978–2983, 2014.

[Kar *et al.*, 2015] D. Kar, F. Fang, F. Delle Fave, N. Sintov, and M. Tambe. "A game of thrones": When human behavior models compete in repeated Stackelberg security games. In *Proceedings of the 14th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 1381–1390, 2015.

[Letchford *et al.*, 2009] J. Letchford, V. Conitzer, and K. Munagala. Learning and approximating the optimal strategy to commit to. In *Proceedings of the 2nd International Symposium on Algorithmic Game Theory (SAGT)*, pages 250–262, 2009.

[Luce, 2005] R. D. Luce. *Individual choice behavior: A theoretical analysis*. Courier Corporation, 2005.

[Marecki *et al.*, 2012] J. Marecki, G. Tesauro, and R. Segal. Playing repeated Stackelberg games with unknown opponents. In *Proceedings of the 11th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 821–828, 2012.

[McFadden, 1976] D. L. McFadden. Quantal choice analaysis: A survey. *Annals of Economic and Social Measurement*, 5(4):363–390, 1976.

[Nguyen *et al.*, 2013] T. H. Nguyen, R. Yang, A. Azaria, S. Kraus, and M. Tambe. Analyzing the effectiveness of adversary modeling in security games. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI)*, pages 718–724, 2013.

[Nguyen *et al.*, 2015] T. H. Nguyen, F. M. Delle Fave, D. Kar, A. S. Lakshminarayanan, A. Yadav, M. Tambe, N. Agmon, A. J. Plumptre, M. Driciru, F. Wanyama, and A. Rwetsiba. Making the most of our regrets: Regret-based solutions to handle payoff uncertainty and elicitation in green security games. In *Proceedings of the 6th Conference on Decision and Game Theory for Security (GameSec)*, pages 170–191, 2015.

[Pita *et al.*, 2009] J. Pita, M. Jain, F. Ordóñez, C. Portway, M. Tambe, C. Western, P. Paruchuri, and S. Kraus. Using game theory for Los Angeles airport security. *AI Magazine*, 3(1):43–57, 2009.

[Sinha *et al.*, 2016] A. Sinha, D. Kar, and M. Tambe. Learning adversary behavior in security games: A PAC model perspective. In *Proceedings of the 15th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2016. Forthcoming.

[Tambe, 2012] M. Tambe. *Security and Game Theory: Algorithms, Deployed Systems, Lessons Learned*. Cambridge University Press, 2012.

[Yang *et al.*, 2014] R. Yang, B. J. Ford, M. Tambe, and A. Lemieux. Adaptive resource allocation for wildlife protection against illegal poachers. In *Proceedings of the 13th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 453–460, 2014.