Robust commitments and partial reputation

Vidya Muthukumar **Anant Sahai**
EECS, UC Berkeley

Abstract

How should agents shape a finite "reputational" history to their advantage knowing that others will be learning from that history? We focus on one leader interacting with multiple followers and show that in most non-zero-sum games, the traditional Stackelberg mixed commitment is very fragile to observational uncertainty. We propose robust commitment rules that anticipate being learned and show that these approximately preserve Stackelberg-level performance. Further, we show that any leader advantage over and above Stackelberg is only realized in a small-sample regime relative to the effective dimension of the game. Thus we can approximate the optimal payoff in the non-asymptotic regime where learnability matters. We corroborate our theory with illustrative examples and extensive simulations on random ensembles of security games.

1 Introduction

In many application domains like security (Paruchuri et al. [2008]), network routing (Roughgarden [2004]) and law enforcement (Muthukumar and Sahai [2017]), the history of play of an incumbent "leader" agent is public. "Follower" agent(s) observe this history and respond as strategic entities. The question then arises of how the leader should shape her history for maximal advantage. Asymptotically, the answer is given by the *Stackelberg commitment*, which provides often significant advantage over the simultaneous, one-shot equilibrium (Von Stengel and Zamir [2010]). However, in finitely-repeated interaction, followers may not be able to observe the exact nature of commitment, or even believe in it. Our main contribution is to understand the extent of reputational advantage when interaction is finite, and prescribe approximately optimal commitment rules for this regime.

During finitely-repeated interaction, the follower needs to *learn and make inferences about the leader's strategic behavior*. The new twist is that the leader (who generates samples of play), and follower (who learns from these samples), possess selfish, possibly non-aligning, not necessarily adversarial incentives. Our broader intellectual motivation is understanding the incentives of a leader whose strategic behavior is being learned – does she want to reveal, or obfuscate, her strategic nature?

Related work Theoretical analysis of *reputation effects* has been limited to the asymptotics and convergence of an appropriate solution concept, usually the Bayes-Nash equilibrium of a repeated game endowed with a prior on leader incentives, to *Stackelberg equilibrium* (Kreps and Wilson [1982], Milgrom and Roberts [1982], Fudenberg and Levine [1989, 1992]). Our analysis is new in that it studies reputation building in a *non-asymptotic manner* – we prescribe how a leader should play to maximally build a partial reputation, and analyze how much advantage she continues to enjoy.

Uncertainty in observation of commitments has been recently studied in Stackelberg security games played between a defender (leader) and attacker (follower). Proposed models include anchoring support theory (Pita et al. [2010]) and a Bayesian prior on the mixed commitment (An et al. [2012], Shieh et al. [2012]). Heuristic, computationally complex algorithms have been developed to optimize leader payoff assuming these models. Analytically, it was shown that the Stackelberg commitment continued to approximate the optimal payoff for zero-sum games (Blum et al. [2014]); but that reasoning does not extend to non-zero-sum games. Qualitatively different sources of observational uncertainty have also been considered, such as the possibility of the mixed commitment being fully

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

observed or not observed at all (Korzhyk et al. [2011]). We provide analytical insights for a generic non-zero-sum two-player game, where incentives will not align but need not be adversarial.

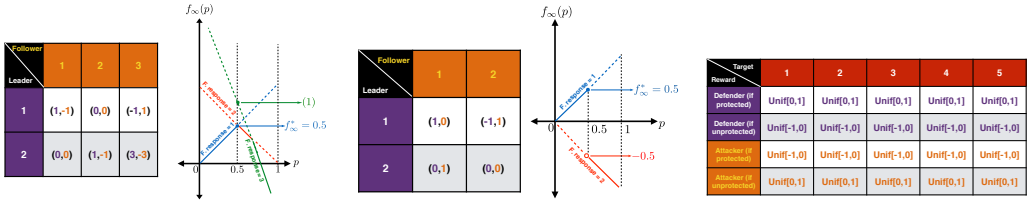
Intellectually related to our contributions are other robust solutions concepts in economics like *Trembling-hand-perfect-equilibria* (Selten [1975]) and *Quantal-response-equilibrium* (McKelvey and Palfrey [1995]). Also related are games with asymmetric private information possessed only by one agent, who can reveal (or obfuscate) this information through repeated play (Aumann et al. [1995]) or explicit signalling (Crawford and Sobel [1982]). Recent work on Bayesian persuasion (Kamenica and Gentzkow [2011]) shows the advantage of manipulating dissemination of private information to elicit a favorable response. Finally, inverse reinforcement learning (Ziebart et al. [2008], Waugh et al. [2013]) is related to the goal of passive learning of strategic behavior. The recent paradigm of *cooperative inverse reinforcement learning* (Hadfield-Menell et al. [2016]) studies a setting where the incentives are not completely aligned, but agents are cooperative. In contrast, we focus on non-cooperative games, but show that learnability is still central.

Problem formulation We represent a two-player game in normal form by the pair of $m \times n$ matrices (A, B), where $A = [\mathbf{a}_1 \quad \mathbf{a}_2 \quad \dots \quad \mathbf{a}_n]$ and $B = [\mathbf{b}_1 \quad \mathbf{b}_2 \quad \dots \quad \mathbf{b}_n]$ denote the leader and follower payoff matrices respectively. We assume that the leader knows about the follower preferences (i.e. B) while the follower does not know about the leader preferences (i.e. A).

Let Δ_m denote the m-dimensional probability simplex, and let a leader *commitment* be denoted by any mixed strategy $\mathbf{x} \in \Delta_m$. We denote the follower's best response to the mixed strategy *that it observes* by $j^*(\mathbf{x}) \in [m]$. Note that $j^*(\mathbf{x})$ is a pure strategy and an important traditional assumption is that in the event that the follower equally prefers more than one pure strategy, the follower breaks ties in favor of the leader (Conitzer and Sandholm [2006]). A leader will expect *ideal payoff* $f_\infty(\mathbf{x}) := \langle \mathbf{x}, \mathbf{a}_{j^*(\mathbf{x})} \rangle$ from mixed commitment $\mathbf{x}$. Therefore, the leader's *ideal Stackelberg payoff* is defined by $f^*_\infty := \max_{\mathbf{x} \in \Delta_m} f_\infty(\mathbf{x})$. The argmax of this optimization problem is denoted as the *Stackelberg commitment* $\mathbf{x}^*_\infty$.

The premise of this paper is that the leader's commitment is partially observed by the follower. Formally, a leader can only reveal her commitment $\mathbf{x}$ through N *pure strategy plays* $I_1, I_2, \dots, I_N$ i.i.d $\sim \mathbf{x}$. The follower assumes that the leader commitment could be any mixture in Δ_m and computes the MLE estimate of leader commitment, $\widehat{\mathbf{X}}_N$. Then, a "rational" follower would play the pure strategy $j^*(\widehat{\mathbf{X}}_N)$ and the leader would expect payoff $f_N(\mathbf{x}) := \mathbb{E}\left[\langle \mathbf{x}, \mathbf{a}_{j^*(\widehat{\mathbf{X}}_N)} \rangle\right]$. We denote $f^*_N := \max_{\mathbf{x} \in \Delta_m} f_N(\mathbf{x})$ and $\mathbf{x}^*_N$ as the argmax.

Under limited observability, the leader thus expects a *mixed response* from the follower. This causes computational difficulties as the optimization becomes non-convex (An et al. [2012]). We take an analytical approach instead to understand how close f^*_N is to f^*_∞, and how close $\mathbf{x}^*_N$ is to $\mathbf{x}^*_\infty$.



(a) 2×3 zero-sum game. (b) 2×2 non-zero-sum game. (c) Security game ensemble.

Figure 1. Illustration of a zero-sum game, a non-zero-sum game, and a random security game ensemble in the form of normal form tables and ideal leader payoff function. p denotes the probability that the leader will play strategy 1, and fully describes leader mixed commitment for the $2 \times n$ games.

2 Main results (shown through examples, simulations and theorems)

The theorems are stated informally for brevity; full formal statements and proofs are contained in the full paper[1]. We choose a 2×3 zero-sum game, a 2×2 non-zero-sum game and a random ensemble

[1] `https://eecs.berkeley.edu/~vidya.muthukumar/robustcommitments.pdf`

of 5×5 security games (An et al. [2012]) to illustrate our results. Figure 1 shows these games in normal form and the *ideal leader payoff functions* illustrating how the follower will respond to leader mixed commitments. Recall that the follower's best-response will be *stochastic*. Our key insight is the potential for leader loss *or* gain compared to the ideal Stackelberg payoff is fundamentally related to the probability that the follower will respond *different than expected*.

(Non)-robustness of Stackelberg commitments As was established in previous work, the Stackelberg commitment $\mathbf{x}^*_\infty$ is robust to observational uncertainty for a zero-sum-game; in fact, there is *benefit* in limited-observation! However, Figure 2 shows the extent of this benefit decays sharply; in fact, exponentially, with N.

The story is different for non-zero-sum games, eg: the game in Figure 1(b). The leader always strictly prefers the follower to respond with his pure strategy 1. Because the Stackelberg commitment ($\mathbf{x}^*_\infty = [1/2 \quad 1/2]$) is on a boundary of the region of commitments for which the follower responds favorably, this response is highly stochastic even for tiny amounts of uncertainty. For this game, the expected payoff is 0 for any value of $N < \infty$, much less than the ideal Stackelberg payoff $f^*_\infty = 0.5$. Figure 3 indicates that this disadvantage also prominently shows up on average for the random security-game ensemble. Clearly, this phenomenon is the norm rather than the exception.

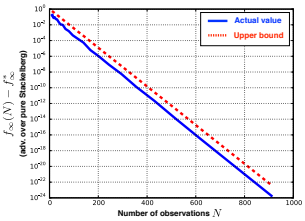


Figure 2. Extent of advantage over Stackelberg in zero-sum game.

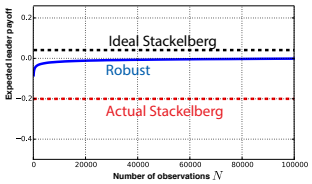


Figure 3. Expected defender payoff using robust commitments and Stackelberg commitment for random security game ensemble.

Simple, robust commitment constructions What made the Stackelberg commitment suboptimal in finite observability was precisely its optimality in infinite observability – the linear program characterization of Stackelberg equilibrium (Conitzer and Sandholm [2006]) pushes the commitment $\mathbf{x}^*_\infty$ all the way to the boundary of a *best-response-region*, which is a convex polytope. If the leader wanted to improve commitment learnability, she could move her commitment to the interior of the polytope to elicit the expected follower response with high probability. She will want to do this while remaining sufficiently close to the ideally optimal Stackelberg commitment $\mathbf{x}^*_\infty$, and a technical tradeoff results. Our robust commitments optimize this tradeoff and ensure that we can approach the ideal Stackelberg payoff at a specific rate with N, provided that N is greater than the effective dimension of the game which we denote by $k(\leq m)$.

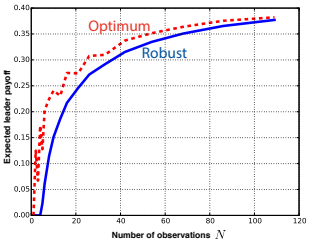


Figure 4. Comparing robust commitment's payoff to optimum f^*_N (brute-forced) for 2×2 non-zero-sum game.

Theorem 1. *For $N > \Omega(k)$, we provide explicit commitments $\{\mathbf{x}_N\}_{N \geq 1}$ such that $f^*_\infty - f_N(\mathbf{x}_N) \leq \widetilde{\mathcal{O}}(\sqrt{\frac{1}{N}})$.*

Further, after we have computed the Stackelberg commitment $\mathbf{x}^*_\infty$ according to the algorithm in (Conitzer and Sandholm [2006]), we can compute these constructions in constant time. Figure 3 displays the performance of the robust commitments as compared to ideal Stackelberg payoff. Their advantage over the non-robust Stackelberg commitment is clearly visible.

Approximation guarantees on maximum obtainable payoff The robust commitment constructions optimized follower learnability. However, the leader could also try and fool the follower into responding favorably, which is what creates the potential for a gain over and above Stackelberg (as in the zero-sum case). To show that our robust commitments are approximately optimal, we show that the potential for such gain decreases with N regardless of the choice of commitments.

Theorem 2. *For any game (A, B), we have $f^*_N \leq f^*_\infty + \mathcal{O}(\frac{1}{\sqrt{N}})$ for $N > \mathcal{O}(k)$.*

Theorems 1 and 2 together imply that our robust commitments $\{f_N(\mathbf{x}_N)\}_{N \geq 1}$ provide an $\mathcal{O}(\frac{1}{\sqrt{N}})$-additive approximation to the optimum f^*_N. Figures 4 compares the performance of the two for the non-zero-sum game example in Figure 1(b), and they are observed to be even closer in practice.

References

B. An, D. Kempe, C. Kiekintveld, E. Shieh, S. Singh, M. Tambe, and Y. Vorobeychik. Security games with limited surveillance. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, pages 1241–1248. AAAI Press, 2012.

R. J. Aumann, M. Maschler, and R. E. Stearns. *Repeated games with incomplete information*. MIT press, 1995.

A. Blum, N. Haghtalab, and A. D. Procaccia. Lazy Defenders Are Almost Optimal against Diligent Attackers. In *AAAI*, pages 573–579, 2014.

V. Conitzer and T. Sandholm. Computing the optimal strategy to commit to. In *Proceedings of the 7th ACM Conference on Electronic Commerce*, pages 82–90. ACM, 2006.

V. P. Crawford and J. Sobel. Strategic information transmission. *Econometrica: Journal of the Econometric Society*, pages 1431–1451, 1982.

D. Fudenberg and D. K. Levine. Reputation and equilibrium selection in games with a patient player. *Econometrica: Journal of the Econometric Society*, pages 759–778, 1989.

D. Fudenberg and D. K. Levine. Maintaining a reputation when strategies are imperfectly observed. *The Review of Economic Studies*, 59(3):561–579, 1992.

D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. Dragan. Cooperative inverse reinforcement learning. In *Advances in neural information processing systems*, pages 3909–3917, 2016.

E. Kamenica and M. Gentzkow. Bayesian persuasion. *The American Economic Review*, 101(6):2590–2615, 2011.

D. Korzhyk, V. Conitzer, and R. Parr. Solving Stackelberg games with uncertain observability. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 3*, pages 1013–1020. International Foundation for Autonomous Agents and Multiagent Systems, 2011.

D. M. Kreps and R. Wilson. Reputation and imperfect information. *Journal of economic theory*, 27(2):253–279, 1982.

R. D. McKelvey and T. R. Palfrey. Quantal response equilibria for normal form games. *Games and economic behavior*, 10(1):6–38, 1995.

P. Milgrom and J. Roberts. Predation, reputation, and entry deterrence. *Journal of economic theory*, 27(2): 280–312, 1982.

V. Muthukumar and A. Sahai. Fundamental limits on ex-post enforcement and implications for spectrum rights. In *Dynamic Spectrum Access Networks (DySPAN), 2017 IEEE International Symposium on*, pages 1–10. IEEE, 2017.

P. Paruchuri, J. P. Pearce, J. Marecki, M. Tambe, F. Ordonez, and S. Kraus. Playing games for security: An efficient exact algorithm for solving Bayesian Stackelberg games. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 2*, pages 895–902. International Foundation for Autonomous Agents and Multiagent Systems, 2008.

J. Pita, M. Jain, M. Tambe, F. Ordóñez, and S. Kraus. Robust solutions to Stackelberg games: Addressing bounded rationality and limited observations in human cognition. *Artificial Intelligence*, 174(15):1142–1171, 2010.

T. Roughgarden. Stackelberg scheduling strategies. *SIAM Journal on Computing*, 33(2):332–350, 2004.

R. Selten. Reexamination of the perfectness concept for equilibrium points in extensive games. *International journal of game theory*, 4(1):25–55, 1975.

E. Shieh, B. An, R. Yang, M. Tambe, C. Baldwin, J. DiRenzo, B. Maule, and G. Meyer. Protect: A deployed game theoretic system to protect the ports of the United States. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 13–20. International Foundation for Autonomous Agents and Multiagent Systems, 2012.

B. Von Stengel and S. Zamir. Leadership games with convex strategy sets. *Games and Economic Behavior*, 69 (2):446–457, 2010.

K. Waugh, B. D. Ziebart, and J. A. Bagnell. Computational rationalization: The inverse equilibrium problem. *arXiv preprint arXiv:1308.3506*, 2013.

B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey. Maximum Entropy Inverse Reinforcement Learning. In *AAAI*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.