Deep Fictitious Play for Games with Continuous Action Spaces

Extended Abstract

Nitin Kamra
University of Southern California
Los Angeles, California
nkamra@usc.edu

Umang Gupta
University of Southern California
Los Angeles, California
umanggup@usc.edu

Kai Wang
University of Southern California
Los Angeles, California
wang319@usc.edu

Fei Fang
Carnegie Mellon University
Pittsburgh, Pennsylvania
feifang@cmu.edu

Yan Liu
University of Southern California
Los Angeles, California
yanliu.cs@usc.edu

Milind Tambe
University of Southern California
Los Angeles, California
milind.tambe11@gmail.com

ABSTRACT

Fictitious play has been a classic algorithm to solve two-player adversarial games with discrete action spaces. In this work we develop an approximate extension of fictitious play to two-player games with high-dimensional continuous action spaces. We use generative neural networks to approximate players' best responses while also learning a differentiable approximate model to the players' rewards given their actions. Both these networks are trained jointly with gradient-based optimization to emulate fictitious play. We explore our approach in zero-sum games, non zero-sum games and security game domains.

KEYWORDS

Fictitious play; Nash equilibrium; Game theory; Multiagent systems; Multiagent learning; Deep Learning

ACM Reference Format:
Nitin Kamra, Umang Gupta, Kai Wang, Fei Fang, Yan Liu, and Milind Tambe. 2019. Deep Fictitious Play for Games with Continuous Action Spaces. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 3 pages.

1 INTRODUCTION

Computing Nash Equilibrium (NE) is an important intermediate step in game theoretic domains and finds major applications in economics, planning, security domains etc. In this work, we consider the problem of finding approximate mixed strategy Nash equilibrium in two-player games with *continuous action spaces* for players.

We are particularly motivated by security domains which involve protecting geographic areas and often lead to continuous action spaces [7, 11, 14, 21]. Though previous approaches focus on discretized action spaces [6, 7, 22], special spatio-temporal structure in games [1, 2, 5, 23] and numerical solutions using approximate differential equations in special cases [11], these do not extend generally to most two-player game settings.

Hence we focus on extending a classic algorithm namely fictitious play (FP) to two-player games with continuous action spaces. Fictitious play involves players repeatedly playing the game and best responding to each other's history of play. FP has been shown to converge to a NE for specific classes of discrete action games with exact best responses [16] and with approximate best responses [3]. We surmise that it can be extended to two-player continuous action space games with approximate best responses. This hypothesis is partly supported by a variant of FP called Stochastic Fictitious Play (SFP) [10] which adds an entropy-maximizing objective to FP and has been shown to converge under more diverse settings: discrete time [10], continuous time [20] and with continuous action sets [19] under reasonable regularity assumptions over underlying domains.

Motivated by these properties, we develop an approximate fictitious play algorithm for two-player games with continuous action spaces. We make the following key contributions: (a) We use novel state-of-the-art generative neural networks to implicitly represent stochastic best responses for players. These networks are very flexible at learning arbitrary distributions with no explicit shape assumptions on players' action spaces, (b) We also learn a game-model neural network which is a differentiable approximation of the players' payoffs given their actions, (c) we train the game-model network and the best response networks end-to-end in a decoupled manner to approximate the Nash equilibrium of games with continuous action spaces.

We also address certain limitations of previous multiagent learning methods. Since *deterministic* player policies work only in collaborative settings [18] but are easily exploited from an adversarial viewpoint, we work in the *stochastic* policy regime. Existing methods which employ stochastic policies do it in domains with discrete action sets since explicit distributions can be maintained over them [4, 8, 9, 17]. However it is challenging to maintain distributions over continuous action spaces and existing approaches often assume explicit distributions for players' strategies which may not span the full space of strategies to which Nash equilibrium distributions belong (e.g. OptGradFP [12, 13] assumes independent multivariate logit-normal distributions for players' strategies). Our use of generative neural networks alleviates this issue and provides stronger modeling capabilities for the best response strategies since they can implicitly approximate arbitrary probability densities. Further, our approach does not require any likelihood estimates and thereby converges stably with minimal or no regularization.

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

2 APPROXIMATING FICTITIOUS PLAY

We consider a two-player game with continuous action sets for players 1 and 2. We will often use the index $p \in \{1, 2\}$ for one of the players and $-p$ for the other player. Letting U_p be the compact, convex action set of player p and $\mathcal{P}(U_p)$ be the set of all probability measures on U_p, the mixed strategy of player p is $\pi_p \in \mathcal{P}(U_p)$ with $\pi_p(B_p)$ denoting the probability of player p selecting an action in the set $B_p \subseteq U_p$. We further denote the probability density function for player p at action $u_p \in U_p$ as $\sigma_p(u_p)$ i.e. $\pi_p(B_p) = \int_{B_p} \sigma_p(u_p) du_p$. An action $u_p \in U_p$ can be sampled from a player p's mixed strategy ($u_p \sim \pi_p$) or from the associated density ($u_p \sim \sigma_p$), and we use these notations interchangeably. We denote joint actions, joint action sets, joint distributions and joint densities without any player subscript i.e. as $u = (u_1, u_2), U = U_1 \times U_2, \pi = (\pi_1, \pi_2)$ and $\sigma = (\sigma_1, \sigma_2)$ respectively.

Each player has a bounded and Lipschitz continuous reward function $r_p : U \to \mathbb{R}$. For zero-sum games, $r_p(u) + r_{-p}(u) = 0 \;\; \forall u \in U$. With players' mixed strategy densities σ_p and σ_{-p}, the expected reward of player p is:

$$\mathbb{E}_{u\sim\sigma}[r_p] = \int_{U_p} \int_{U_{-p}} r_p(u)\sigma_p(u_p)\sigma_{-p}(u_{-p}) du_p du_{-p}$$

The best response of player p against player $-p$'s current strategy σ_{-p} is defined as the set of strategies which maximizes his expected reward:

$$BR_p(\sigma_{-p}) := \arg\max_{\sigma_p} \left\{ \mathbb{E}_{u\sim(\sigma_p, \sigma_{-p})}[r_p] \right\},$$

A pair of strategies $\sigma = (\sigma_1, \sigma_2)$ is said to be a Nash equilibrium if neither player can increase his expected reward by changing his strategy while the other player sticks to his current strategy. In such a case both these strategies belong to the best response sets to each other i.e.

$$\sigma_1 \in BR_1(\sigma_2),$$
$$\sigma_2 \in BR_2(\sigma_1).$$

To compute NE for the game of interest, we introduce an approximate realization of fictitious play in high-dimensional continuous action spaces. Let the empirical distribution of player p's previous actions (a.k.a. belief distribution) be $\bar{\pi}_p$ and the corresponding density function (a.k.a. belief density) be $\bar{\sigma}_p$. Then fictitious play involves player p repeatedly best responding to his opponent's belief density $\bar{\sigma}_{-p}$:

$$BR_p(\bar{\sigma}_{-p}) := \arg\max_{\sigma_p} \left\{ \mathbb{E}_{u\sim(\sigma_p, \bar{\sigma}_{-p})}[r_p] \right\},$$

Repeating this procedure for both players is guaranteed to converge to the Nash equilibrium densities for both players for certain classes of games [16]. This implies that approximating FP in continuous action spaces requires approximations to two essential ingredients:

(1) Belief densities over players' actions, and
(2) Best responses for each player.

In this work we employ two novel ways to approximate both the above mentioned essential ingredients thereby extending fictitious play to games with continuous action spaces.

Maintaining belief densities: Representing belief densities compactly is challenging in continuous action spaces. We propose to maintain belief density $\bar{\sigma}_p$ of each player p via a non-parameterized population based estimate i.e. via memory of all actions played by p so far. Directly sampling u_p from the memory gives an unbiased sample from $\bar{\sigma}_p$.

Approximating best responses: Computing exact best response is intractable for most games. But when the expected reward for a player p is differentiable w.r.t. the player's action u_p and admits continuous and smooth derivatives, best responses can be approximated. We approximate the best response function of player p with deep neural networks represented as $BR_p(\cdot; \theta_p)$ with trainable parameters θ_p and keep them updated with gradient ascent in every iteration of fictitious play. These are essentially implicit density models [15] but are trained differently using another differentiable *game model network* which takes all players' actions i.e. $\{u_p, u_{-p}\}$ as inputs and predicts rewards $\{\hat{r}_p, \hat{r}_{-p}\}$ for each player. The game model can be pre-trained or learnt simultaneously with the best response networks directly from gameplay data.

When the expected reward is not differentiable w.r.t. players' actions or the derivatives are non-smooth or zero in a large part of the action space, one can also employ an approximate best response oracle (BRO_p) for player p. The oracle can be a non-differentiable approximation algorithm employing LP or MIP, since it will never be trained. In many security games, Mixed-integer programming based algorithms are proposed to compute best responses and our algorithm provides a novel way to incorporate them as subroutines in a deep learning framework, as opposed to most existing works which require end-to-end differentiable policy networks and cannot utilize non-differentiable solutions even when available.

3 CONCLUSION

In this work, we focus on an approximate fictitious play algorithm for games with continuous action spaces. Our proposed method implicitly represents players' stochastic best responses via generative neural networks without prior shape assumptions and optimizes them with gradient-based training. It can also utilize approximate best response oracles whenever available, thereby harnessing prowess in approximation algorithms from discrete planning and operations research. Also, our proposed algorithm is off-policy because of the learnt smoothly parameterized game model. It trains significantly faster than on-policy methods like OptGradFP by directly estimating rewards from the game-model network and alleviating the need to replay previously played games thereby significantly speeding up the training and scaling better with growing number of players' resources.

We test our proposed variants of approximate fictitious play in zero-sum, non zero-sum and security game domains with improved results in achieving complex and flexible Nash equilibrium strategies. We further introduce a novel exploitability analysis using a genetic algorithm to evaluate the learnt strategies. Our approach is easily extended to multi-player applications, with each player p best responding to the joint belief density over all other players $\bar{\sigma}_{-p}$ using an oracle or a best response network.

ACKNOWLEDGMENTS

This research was supported in part by NSF Research Grant (IIS-1254206) and Army Research Office (MURI W911NF1810208).

REFERENCES

[1] Soheil Behnezhad, Mahsa Derakhshan, Mohammadtaghi Hajiaghayi, and Saeed Seddighin. 2018. Spatio-Temporal Games Beyond One Dimension. In *Proceedings of the 2018 ACM Conference on Economics and Computation*. 411–428.

[2] Soheil Behnezhad, Mahsa Derakhshan, MohammadTaghi Hajiaghayi, and Aleksandrs Slivkins. 2017. A Polynomial Time Algorithm for Spatio-Temporal Security Games. In *Proceedings of the 2017 ACM Conference on Economics and Computation*. 697–714.

[3] Vincent Conitzer. 2009. Approximation guarantees for fictitious play. In *47th Annual Allerton Conference on Communication, Control, and Computing*. IEEE, 636–643.

[4] Abhishek Das, Théophile Gervet, Joshua Romoff, Dhruv Batra, Devi Parikh, Michael Rabbat, and Joelle Pineau. 2018. TarMAC: Targeted Multi-Agent Communication. *arXiv preprint arXiv:1810.11187* (2018).

[5] Fei Fang, Albert Xin Jiang, and Milind Tambe. 2013. Optimal patrol strategy for protecting moving targets with multiple mobile resources. In *AAMAS*. 957–964.

[6] Jiarui Gan, Bo An, Yevgeniy Vorobeychik, and Brian Gauch. 2017. Security Games on a Plane. In *AAAI*. 530–536.

[7] William Haskell, Debarun Kar, Fei Fang, Milind Tambe, Sam Cheung, and Elizabeth Denicola. 2014. Robust protection of fisheries with COmPASS. In *IAAI*.

[8] Johannes Heinrich, Marc Lanctot, and David Silver. 2015. Fictitious Self-Play in Extensive-Form Games.. In *ICML*. 805–813.

[9] Johannes Heinrich and David Silver. 2016. Deep Reinforcement Learning from Self-Play in Imperfect-Information Games. *CoRR* abs/1603.01121 (2016).

[10] Josef Hofbauer and William H Sandholm. 2002. On the global convergence of stochastic fictitious play. *Econometrica* 70, 6 (2002), 2265–2294.

[11] Matthew P. Johnson, Fei Fang, and Milind Tambe. 2012. Patrol Strategies to Maximize Pristine Forest Area. In *AAAI*.

[12] Nitin Kamra, Fei Fang, Debarun Kar, Yan Liu, and Milind Tambe. 2017. Handling continuous space security games with neural networks. In *IWAISe: First International Workshop on Artificial Intelligence in Security*.

[13] Nitin Kamra, Umang Gupta, Fei Fang, Yan Liu, and Milind Tambe. 2018. Policy Learning for Continuous Space Security Games using Neural Networks. In *AAAI*.

[14] Debarun Kar, Fei Fang, Francesco Delle Fave, Nicole Sintov, and Milind Tambe. 2015. "A Game of Thrones": When Human Behavior Models Compete in Repeated Stackelberg Security Games. In *AAMAS*.

[15] Taesup Kim and Yoshua Bengio. 2016. Deep directed generative models with energy-based probability estimation. *arXiv preprint arXiv:1606.03439* (2016).

[16] Vijay Krishna and Tomas Sjöström. 1998. On the convergence of fictitious play. *Mathematics of Operations Research* 23, 2 (1998), 479–511.

[17] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*. 6379–6390.

[18] Shayegan Omidshafiei, Jason Pazis, Christopher Amato, Jonathan P How, and John Vian. 2017. Deep decentralized multi-task multi-agent reinforcement learning under partial observability. *arXiv preprint arXiv:1703.06182* (2017).

[19] S. Perkins and D.S. Leslie. 2014. Stochastic fictitious play with continuous action sets. *Journal of Economic Theory* 152 (2014), 179 – 213.

[20] Jeff S Shamma and Gürdal Arslan. 2004. Unified convergence proofs of continuous-time fictitious play. *IEEE Trans. Automat. Control* 49, 7 (2004), 1137–1141.

[21] Binru Wang, Yuan Zhang, and Sheng Zhong. 2017. On Repeated Stackelberg Security Game with the Cooperative Human Behavior Model for Wildlife Protection. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS '17)*. 1751–1753.

[22] Rong Yang, Benjamin Ford, Milind Tambe, and Andrew Lemieux. 2014. Adaptive Resource Allocation for Wildlife Protection against Illegal Poachers. In *AAMAS*.

[23] Yue Yin, Bo An, and Manish Jain. 2014. Game-theoretic Resource Allocation for Protecting Large Public Events. In *AAAI*. 826–833.