MIT 6.S890 — Additional Projects Ideas

Instructor: Gabriele Farina

Fall 2024

▸ CONTENTS

Last update `2024-10-08`

1. Iterate Convergence to Min-Max Equilibrium in Stochastic Games . . . 1
2. Convergence to an epsilon-well supported Correlated Equilibrium . . . 1
3. Karlin's Conjecture . . . 2
4. Optimistic Boosting . . . 2
5. Fast implementation of Counterfactual Regret Minimization . . . 2
6. Average no-regret policies beyond two-player zero-sum games . . . 2
7. Use of optimism in PPO . . . 3
8. Overfitting in PSRO . . . 3
Bibliography . . . 3

This page contains resources to help you formulate a topic for your semester project for the course. Below you can find a list of semi-concrete directions that we find will lead to interesting projects. You can pick a project inspired by these directions and discuss your concrete plan with us before proceeding to writing your proposal. Alternatively, you can select a project around anything that interests you and is related to the topics of this class. For instance, you may pick a project related to your own research interests, which is also related to the course. However, it is not acceptable to submit something that you have already done as part of your research.

1. Iterate Convergence to Min-Max Equilibrium in Stochastic Games

In recent work [1] it was shown that if players of a two-player zero-sum stochastic-game use the online policy gradient method to iteratively update their policies, the resulting dynamics converges to a min-max equilibrium in a best-iterate sense; see Theorem 1 of this paper. Can you identify an "optimistic variant" of the online policy gradient method that exhibits last-iterate convergence, answering Open Problem 1 in that paper? Can you show that optimistic policy gradient converges faster than non-optimistic policy gradient in the best-iterate sense of Theorem 1? For more context on optimistic methods and references see problem above.

2. Convergence to an epsilon-well supported Correlated Equilibrium

In class we showed that no-swap regret dynamics converge to a correlated equilibrium. Moreover, if the regret of the algorithms is $\mathrm{Reg}^{(T)}$, then after T iterations the empirical distribution is an $\mathrm{Reg}^{(T)}/T$-approximate correlated equilibrium. However, the approximation here is of a weak kind: it is not necessarily the case that the conditional expected utility of a player when recommended an action is ε-close to his conditional optimal utility; instead the guarantee is that these quantities multiplied by the probability of being recommended the action are ε-close to each other. If the probability of being recommended the action is tiny, then there may be a huge difference between the conditional utilities. Can learning dynamics get an ε-

correlated equilibrium of the stronger kind with fast rates? This would have many interesting implications. Or is this as hard as designing learning dynamics with fast convergence to ε-approximate Nash equilibrium? These questions are interesting to study even in two-player settings.

3. Karlin's Conjecture

In class we saw that follow-the-leader (FTL) is not a no-regret learning algorithm. Despite this fact, in two-player zero-sum games FTL is known to converge to Nash equilibrium. When used by all players in a game, FTL goes under the name of "Fictitious Play". The proof that it converges to equilibrium in two-player zero-sum games is due to J. Robinson [2]. However, the convergence rate established by Robinson is terrible in the number of actions available to the players; in particular, it is $T^{-O(1/m)}$ when the players have m actions. An open question suggested by S. Karlin is whether Fictitious Play actually converges at a rate of $1/\sqrt{T}$, which is the rate you would get from FTRL or from FTPL algorithm. This was disproven by C. Daskalakis and Q. Pan [3]. Even more recently, I. Panageas, N. Patris, S. Skoulakis, and V. Cevher [4] showed that even the weak version of the hypothesis (for any tie-breaking scheme) is false in potential games.

Can this latter result be shown for zero-sum polymatrix games, another class of games for which Fictitious Play is known to converge [5] (see also result given by C. Daskalakis and C. H. Papadimitriou [6] for a connection with two-player zero-sum games)?

4. Optimistic Boosting

It is known that a variant of the Multiplicative Weights Update algorithm (a.k.a. Hedge) can attain faster convergence by using optimism [7], [8]. Boosting is a very popular technique in machine learning. From an early work of Y. Freund and R. E. Schapire [9], it was pointed out that boosting can be viewed as no-regret learning in a zero-sum game, where one player uses the Hedge algorithm and the other player approximately best responds using a weak classifier. When (i.e. for what learning settings) can replacing Hedge with optimistic Hedge (and perhaps the best-response player with a smooth best response player) lead to an improved boosting algorithm? What theoretical properties can we prove that outperform the Hedge version of boosting (typically referred to as AdaBoost)? Does optimistic boosting lead to stability in the sample weights of the adversary? Can we combine the analysis of optimistic Hedge by V. Syrgkanis, A. Agarwal, H. Luo, and R. E. Schapire [7] or C. Daskalakis, M. Fishelson, and N. Golowich [8] with the analysis of optimistic FTPL from V. Syrgkanis, A. Krishnamurthy, and R. Schapire [10] to get better boosting theorems? (potentially with stronger versions of the "weak classifier condition"). Do they lead to improved practical performance? This would be a combination of theory and simulation.

5. Fast implementation of Counterfactual Regret Minimization

Counterfactual regret minimization (CFR) is a very popular algorithm for solving sequential zero-sum games of incomplete information and variants of it are at the core of the Libratus AI poker system that recently beat top professionals [11]. Can CFR be implemented efficiently on a GPU? What levels of performance can be extracted?

6. Average no-regret policies beyond two-player zero-sum games

In two-player zero-sum games, the average strategy played by no-regret players is guaranteed to converge to the set of Nash equilibria. Beyond that setting, no guarantees are known, but that has not stopped people from using average strategies with success in practice (for example, in six-player poker that led to superhuman performance [12]. What can be said, in theory, or just experimentally, about the average strategies of no-regret players in multiplayer games? Do they empirically perform well across a wide range of domains? For this exploration, you could consider using OpenSpiel [13], a game library from DeepMind that includes several game settings.

7. Use of optimism in PPO

PPO [14] is a popular algorithm used in deep reinforcement learning. Conceptually, it can be thought of as a function approximation version of the multiplicative weights update (MWU) algorithm. Optimistic variants of the MWU algorithm are known to achieve superior performance. Does optimism help PPO as well?

8. Overfitting in PSRO

The PSRO algorithm [15] operates by gradually constructing an approximation of the game where the players are constrained to only use a certain set of strategies. At each iteration, an equilibrium of the restricted game is constructed, and the players then figure out what strategy would best respond (in the unrestricted game) to such an equilibrium of the restricted game. The best responses are then added to the set of strategies that the players can use, and the process continues. Of course, the equilibrium of the restricted game might be very far from being an equilibrium in the original, unrestricted game, as the players might not have discovered strong strategies yet. In that sense, they might be "overfitting" to the restricted game. The problem of overfitting in machine learning is well-studied, and regularization is often a good solution in practice. Is regularization in PSRO beneficial to avoid overfitting, and to achieve better performance in practice? For example, would the players benefit from computing regularized equilibria of the restricted games? Or from computing regularized (smooth) best responses?

Bibliography

[1] C. Daskalakis, D. J. Foster, and N. Golowich, "Independent policy gradient methods for competitive reinforcement learning," *Advances in neural information processing systems*, vol. 33, pp. 5527–5540, 2020.

[2] J. Robinson, "An Iterative Method of Solving a Game," *Annals of Mathematics*, vol. 54, no. 2, pp. 296–301, 1951, Accessed: Oct. 02, 2024. [Online]. Available: http://www.jstor.org/stable/1969530

[3] C. Daskalakis and Q. Pan, "A counter-example to Karlin's strong conjecture for fictitious play," in *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, 2014, pp. 11–20.

[4] I. Panageas, N. Patris, S. Skoulakis, and V. Cevher, "Exponential lower bounds for fictitious play in potential games," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[5] C. Ewerhart and K. Valkanova, "Fictitious play in networks," *Games and Economic Behavior*, vol. 123, pp. 182–206, 2020, doi: https://doi.org/10.1016/j.geb.2020.06.006.

[6] C. Daskalakis and C. H. Papadimitriou, "On a network generalization of the minmax theorem," in *International Colloquium on Automata, Languages, and Programming*, 2009, pp. 423–434.

[7] V. Syrgkanis, A. Agarwal, H. Luo, and R. E. Schapire, "Fast convergence of regularized learning in games," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[8] C. Daskalakis, M. Fishelson, and N. Golowich, "Near-optimal no-regret learning in general games," *Advances in Neural Information Processing Systems*, vol. 34, pp. 27604–27616, 2021.

[9] Y. Freund and R. E. Schapire, "Game theory, on-line prediction and boosting," in *Proceedings of the ninth annual conference on Computational learning theory*, 1996, pp. 325–332.

[10] V. Syrgkanis, A. Krishnamurthy, and R. Schapire, "Efficient algorithms for adversarial contextual learning," in *International Conference on Machine Learning*, 2016, pp. 2159–2168.

[11] N. Brown and T. Sandholm, "Superhuman AI for heads-up no-limit poker: Libratus beats top professionals," *Science*, vol. 359, no. 6374, pp. 418–424, 2018.

[12] N. Brown and T. Sandholm, "Superhuman AI for multiplayer poker," *Science*, vol. 365, no. 6456, pp. 885–890, 2019.

[13] M. Lanctot *et al.*, "OpenSpiel: A Framework for Reinforcement Learning in Games," *CoRR*, 2019, [Online]. Available: http://arxiv.org/abs/1908.09453

[14] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[15] M. Lanctot *et al.*, "A unified game-theoretic approach to multiagent reinforcement learning," *Advances in neural information processing systems*, vol. 30, 2017.

6.S890 Topics in Multiagent Learning

Fall 2024