Thesis proposal: Regret Minimization for Stateful, Cooperative Settings

Chirag Chhablani (chhab2@uic.edu), Ian Kash (iankash@uic.edu)

Abstract

While regret minimization has been extremely effective in two-player, zero-sum stateful settings like poker, results beyond them are much more limited. However, in the stateless case of normal form games, strong results have been proved for cooperative settings (identical interest games and some generalizations). Our position is that the time is ripe to investigate regret minimization in stateful, cooperative settings. As motivation, we discuss our recent preliminary work showing that the updates COMA (a recent algorithm for such settings) uses are quite close to regret minimization. We conclude by discussing three research challenges in this space that seem like natural starting points.

Introduction

Over the past decade there has been tremendous progress in the use of regret minimization in stateful settings. Approaches based on Counterfactual Regret Minimization (CFR) (Zinkevich et al. 2008) have had tremendous success in a variety of poker settings (Moravčík et al. 2017; Brown and Sandholm 2017, 2019). More recently, policy gradient and actor-critic methods have been exploiting connections between advantage functions and regret (Jin, Keutzer, and Levine 2018; Srinivasan et al. 2018; Hennes et al. 2020). Other recent work has connected entropy-regularized reinforcement learning (RL) to mirror descent (Neu, Jonsson, and Gómez 2017), showed how to combine regret minimization with value-iteration-style updates (Kash, Sullins, and Hofmann 2020), studied applications of regret minimization in Monte Carlo Tree Search (Kovařík and Lisỳ 2020), and reduced RL to contextual bandits (Daumé III, Langford, and Sharaf 2018).

The focus of much of this literature has been on competitive settings, often two-player zero-sum ones. In doing so it has built on the rich literature on regret minimization in normal-form games. In particular in two-player, zero-sum games it is well known that algorithms that achieve low regret have an average policy that achieves an equal approximation to Nash equilibrium (Zinkevich et al. 2008; Farina, Kroer, and Sandholm 2017). While this is an appealing result, those for more general normal form games are much more limited. It particular, the most general result is that minimizing regret leads the empirical policy to converge to a coarse correlated equilibrium while algorithms with a stronger guarantee of minimizing internal regret have an empirical policy that coverges to a correlated equilibrium (Foster and Vohra 1997; Gordon, Greenwald, and Marks 2008). These guarantees are often substantially weaker that the guarantee of convergence to Nash from two-player, zero-sum games.

However, there is another smaller literature on regret minimization in congestion games (also known as potential games) and a special case of them known as identical interest games, where all players share the same utility function. Kleinberg, Piliouras, and Tardos (2009) show that players who use Hedge-like updates in congestion games end up playing a pure Nash equilibrium for a fraction of time that is arbitrarily close to 1 with probability also arbitrarily close to 1 after a polynomially small transient stage. Mehta, Panageas, and Piliouras (2015) showed that the multiplicative weights algorithm converges to a pure Nash equilibrium for all but a measure 0 of initial conditions, and hence obtained a stronger guarantee, in identical interest games. Krichene, Drighès, and Bayen (2015) showed that agents converge to Nash equilibrium in all nonatomic potential games if the same algorithm is run with a decreasing step-size. For bandit setting, Coucheney, Gaujal, and Mertikopoulos (2015) showed that a "penalty-regulated" variant of the MW algorithm converges to ϵ-approximate Nash equilibria in congestion games with bandit feedback. Thus, in the stateless, cooperative setting quite strong results like convergence to a pure Nash and convergence of the last iterate rather than the average are possible.

Our position is that the time is ripe to investigate the fundamentals of regret minimization in stateful, cooperative settings. Such settings have seen recent progress in the form of algorithms like COMA (Foerster et al. 2018), QMIX (Rashid et al. 2018), and QDPP (Yang et al. 2020) and Hanabi has recently been proposed as a challenge problem (Bard et al. 2020). To make this case, we first discuss our recent preliminary work showing that the updates COMA uses are quite close to regret minimization (Authors 2020). Furthermore, using a fix to move it closer to regret minimization (Hennes et al. 2020) improves its performance in stateful settings despite a lack of theory. These two facts provide evidence for

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

the promise of regret minimization in stateful, cooperative settings. We conclude by discussing three research directions that seem like natural starting points.

COMA and Regret Minimization

In this section we highlight our recent preliminary work on a connection between a popular cooperative multiagent algorithm, *Counterfactual Multiagent Policy Gradients* (COMA) (Foerster et al. 2018), and regret minimization. As we discuss, this work provides two pieces of evidence for the promise of regret minimization in cooperative, stateful settings.

COMA extends Actor-Critic methods to multiagent settings for learning cooperation among agents when the agents receive a shared reward. The challenge is to create an individual reward signal from the shared reward for personalised training for each agent. Key to COMA's success is efficient multiagent credit assignment through the implementation of *difference rewards* (DR) which were proposed by Wolpert and Tumer (2002) and Tumer and Agogino (2007). For each agent, the difference reward signal represents the benefit of including the agent in the system compared to the counterfactual case when it is excluded from the system. This individual advantage signal is called the *counterfactual advantage baseline* give by:

$$A^a(s, \mathbf{u}) = Q(s, \mathbf{u}) - \sum \pi^a(u'^a|\tau_a) Q(s, (u'^a, \mathbf{u}^{-a})) \quad (1)$$

The first term on right hand side measures expected return $Q(s, \mathbf{u})$ when agent a takes action u^a when others take joint action $\mathbf{u}^{-a}$, while the second term is captures a counterfactual scenario when a is removed from the system by subtracting its expected value based on current policy. This counterfactual baseline is used to update the policy π (parameterized by θ) for each agent a according to the standard policy gradient equation for actor-critic algorithms. For a given state-action pair $(s, \mathbf{u})$) this is given by:

$$\theta_t = \theta_{t-1} + \eta_t \sum \nabla \log \pi^a_\theta(u^a|\tau^a; \theta_{t-1}) A^a(s, \mathbf{u}) \quad (2)$$

Our first observation, as has been observed in more general settings (Jin, Keutzer, and Levine 2018; Srinivasan et al. 2018), is that difference reward implemented in Equation (1) is similar to regret of taking a particular action u^a for an agent a. Given the successes of COMA in stateful, cooperative setttings (in particular cooperative stochastic games, also known as DEC-POMDPs) this connection already provides evidence that regret minimization in such settings is promising and has hope of achieving strong guarantees. For example, COMA only tracks its "last iterate"; it doesn't perform any policy averaging as needed by regret minimization in general settings.

Our second observation, again parallel to previous work in more general settings (Hennes et al. 2020), is that the policy gradient update in Equation (2) is not the exact update that a regret minimizer would use. Instead, the $\log \pi$ term should be replaced with the logits of the neural network which makes the policy update more adaptive to the observation. We call this variant with the fix used by Neural Replicator Dynamics (Hennes et al. 2020) COMA-N. For identical interest games with the usual softmax policy update COMA-N is equivalent to Hedge, which as previously discussed has strong guarantees. Furthermore, experiments show that COMA-N outperforms COMA in stateful settings. This observation, that changing COMA to make it behave more like a regret minimizer improves its performance, provides a second piece of evidence for the promise of regret minimization in cooperative, stateful settings.

We show a subset of our experiments in particular those with four version of the switch game from the ma-gym repository [1]. It is clear from Figures 1-4 that COMA-N leads to faster training and better performance than COMA.

Analyzing Regret Minimization in Stateful, Cooperative Settings

We now turn to the first of our three research directions. As discussed in the introduction, identical interest games are known to be a particularly well-behaved class of games. In particular, they are a special cases of potential games (and even more generally weakly acyclic games) which are relatively forgiving in the sorts of learning dynamics which are guaranteed to converge (Marden, Arslan, and Shamma 2007). We have strong theorems about particular regret minimizers, such as Hedge, in identical interest games. Can we prove similar theorems for stateful identical interest games? For example, suppose we apply Hedge to the set of pure strategies of an extensive form game. What can we guarantee about the result? There is of course the standard reduction of extensive form games to normal form games, but this reduction creates a normal form game that violates some of the assumptions used in prior analyses. For example, prior analysis has required there to be no ties each rows and column of the payoff matrix (Mehta, Panageas, and Piliouras 2015), but this will generally be violated by the reduction. One approach might be to circumvent this by adding a small noise to the probabilities of the players, in the spirit of trembling hand equilibrium, to avoid any ties in payoff matrix. The computation of such solutions, and its relationship to regret minimization approaches, has recently been explored for general games (Farina, Gatti, and Sandholm 2018), but to our knowledge has not been examined in cooperative settings.

A Stronger Analysis of CFR

The standard analysis of CFR shows that when a regret minimization algorithm with an $O(1/\sqrt{T})$ guarantee is used to determine the strategy at each information set then the algorithm achieves a global $O(1/\sqrt{T})$ regret guarantee (Zinkevich et al. 2008). That is, it establishes that CFR is a "generic" regret minimizer in the space of (global) strategies). However, regret minimization alone is not enough to show the desirable properties of algorithms like Hedge in identical interest games (Kleinberg, Piliouras, and Tardos 2009). The reason is that there are trivial algorithms that still satisfy a regret minimization property. For example,

[1] https://github.com/koulanurag/ma-gym

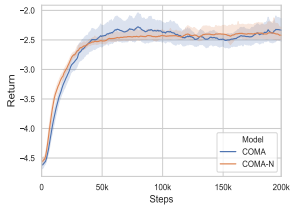

Figure 1: Switch2-v0

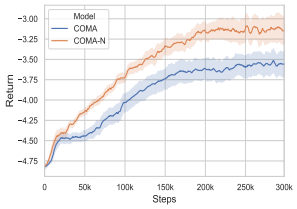

Figure 2: Switch4-v0

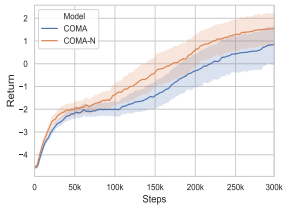

Figure 3: Switch2-v1

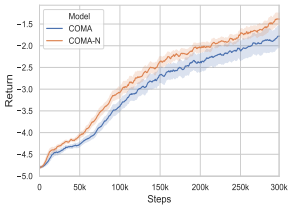

Figure 4: Switch4-v1

a group of agents that collectively coordinate to play any coarse correlated equilibrium will all have zero regret, but this is a much weaker guarantee than the results previously discussed.

A natural question therefore is whether we can give a stronger analysis of CFR, perhaps under the assumption that it uses a particular regret minimizer. For example, if CFR uses Hedge as its choice of regret minimizer, can we show that it is amenable to analysis by the same sorts of continuous time dynamics that have been used to analyze Hedge? This is the clearly related to the previous research direction, but somewhat distinct because of the local regret minimization structure of CFR rather than directly applying global regret minimization in the space of policies.

Competition Between Teams

Another research direction, which straddles the divide between cooperative and competitive settings, is competition between teams. Team competition and its equilibrium guarantees in extensive form and RL settings have received recent attention (Celli and Gatti 2017; Zhang and An 2020; Farina et al. 2018). Recent work has shown that CFR and extensions to it can be useful in Team extensive form like 4-player Kuhn Team poker (2 verses 2), when they are tied through a unified team reward (Hartley, Zheng, and Yue 2017). However the convergence guarantees and properties for such applications is not known. One way to start would be to examine the guarantees that can be provided in zero-sum normal form games with two fully cooperative teams and then extend them to extensive form games with CFR. For larger settings, a natural setting would be to use actor-critic methods like COMA. We have already seen that they are related to regret minimization and recent work has applied them in such settings (Celli et al. 2019).

A related research direction is is exploring the role and efficient use of communicating local state and policies among cooperating agents while using regret minimization. Recent work on cooperative MARL has shown that counterfactual thinking in communication can help in improving coordination amongst agents (Vanneste et al. 2020; Jaques et al. 2019). Hartley, Zheng, and Yue (2017) showed that maintaining a certain degree of communication through an explicit communication channel or through observation from behavior of other agents can be helpful in coordination. A formal analysis of regret minimization with communication is an open research question that we would like explore.

References

Authors, A. 2020. Counterfactual Multiagent Policy Gradients and Regret Minimization in Cooperative Settings. Working Paper.

Bard, N.; Foerster, J. N.; Chandar, S.; Burch, N.; Lanctot, M.; Song, H. F.; Parisotto, E.; Dumoulin, V.; Moitra, S.; Hughes, E.; et al. 2020. The hanabi challenge: A new frontier for ai research. *Artificial Intelligence* 280: 103216.

Brown, N.; and Sandholm, T. 2017. Libratus: the superhuman AI for no-limit poker. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 5226–5228.

Brown, N.; and Sandholm, T. 2019. Superhuman AI for multiplayer poker. *Science* 365(6456): 885–890.

Celli, A.; Ciccone, M.; Bongo, R.; and Gatti, N. 2019. Coordination in Adversarial Sequential Team Games via Multi-Agent Deep Reinforcement Learning. *arXiv preprint arXiv:1912.07712* .

Celli, A.; and Gatti, N. 2017. Computational results for extensive-form adversarial team games. *arXiv preprint arXiv:1711.06930* .

Coucheney, P.; Gaujal, B.; and Mertikopoulos, P. 2015. Penalty-regulated dynamics and robust learning procedures in games. *Mathematics of Operations Research* 40(3): 611–633.

Daumé III, H.; Langford, J.; and Sharaf, A. 2018. Residual loss prediction: Reinforcement learning with no incremental feedback. In *International Conference on Learning Representations*.

Farina, G.; Celli, A.; Gatti, N.; and Sandholm, T. 2018. Ex ante coordination and collusion in zero-sum multi-player extensive-form games. In *Advances in Neural Information Processing Systems*, 9638–9648.

Farina, G.; Gatti, N.; and Sandholm, T. 2018. Practical exact algorithm for trembling-hand equilibrium refinements in games. In *Advances in Neural Information Processing Systems*, 5039–5049.

Farina, G.; Kroer, C.; and Sandholm, T. 2017. Regret minimization in behaviorally-constrained zero-sum games. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1107–1116.

Foerster, J. N.; Farquhar, G.; Afouras, T.; Nardelli, N.; and Whiteson, S. 2018. Counterfactual multi-agent policy gra-

dients. In *Thirty-second AAAI conference on artificial intelligence*.

Foster, D. P.; and Vohra, R. V. 1997. Calibrated learning and correlated equilibrium. *Games and Economic Behavior* 21(1-2): 40.

Gordon, G. J.; Greenwald, A.; and Marks, C. 2008. No-regret learning in convex games. In *Proceedings of the 25th international conference on Machine learning*, 360–367.

Hartley, M.; Zheng, S.; and Yue, Y. 2017. Multi-agent counterfactual regret minimization for partial-information collaborative games. *White paper at CMU* .

Hennes, D.; Morrill, D.; Omidshafiei, S.; Munos, R.; Perolat, J.; Lanctot, M.; Gruslys, A.; Lespiau, J.-B.; Parmas, P.; Duéñez-Guzmán, E.; et al. 2020. Neural Replicator Dynamics: Multiagent Learning via Hedging Policy Gradients. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, 492–501.

Jaques, N.; Lazaridou, A.; Hughes, E.; Gulcehre, C.; Ortega, P.; Strouse, D.; Leibo, J. Z.; and De Freitas, N. 2019. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International Conference on Machine Learning*, 3040–3049. PMLR.

Jin, P.; Keutzer, K.; and Levine, S. 2018. Regret minimization for partially observable deep reinforcement learning. In *International Conference on Machine Learning*, 2342–2351.

Kash, I. A.; Sullins, M.; and Hofmann, K. 2020. Combining No-regret and Q-learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, 593–601.

Kleinberg, R.; Piliouras, G.; and Tardos, E. 2009. Multiplicative updates outperform generic no-regret learning in congestion games. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, 533–542.

Kovařík, V.; and Lisỳ, V. 2020. Analysis of hannan consistent selection for monte carlo tree search in simultaneous move games. *Machine Learning* 109(1): 1–50.

Krichene, W.; Drighès, B.; and Bayen, A. M. 2015. Online learning of nash equilibria in congestion games. *SIAM Journal on Control and Optimization* 53(2): 1056–1081.

Marden, J. R.; Arslan, G.; and Shamma, J. S. 2007. Regret based dynamics: convergence in weakly acyclic games. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, 1–8.

Mehta, R.; Panageas, I.; and Piliouras, G. 2015. Natural selection as an inhibitor of genetic diversity: Multiplicative weights updates algorithm and a conjecture of haploid genetics [working paper abstract]. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, 73–73.

Moravčík, M.; Schmid, M.; Burch, N.; Lisỳ, V.; Morrill, D.; Bard, N.; Davis, T.; Waugh, K.; Johanson, M.; and Bowling, M. 2017. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science* 356(6337): 508–513.

Neu, G.; Jonsson, A.; and Gómez, V. 2017. A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798* .

Rashid, T.; Samvelyan, M.; Schroeder, C.; Farquhar, G.; Foerster, J.; and Whiteson, S. 2018. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In *International Conference on Machine Learning*, 4295–4304.

Srinivasan, S.; Lanctot, M.; Zambaldi, V.; Pérolat, J.; Tuyls, K.; Munos, R.; and Bowling, M. 2018. Actor-critic policy optimization in partially observable multiagent environments. In *Advances in neural information processing systems*, 3422–3435.

Tumer, K.; and Agogino, A. 2007. Distributed agent-based air traffic flow management. In *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, 1–8.

Vanneste, S.; Vanneste, A.; Mercelis, S.; and Hellinckx, P. 2020. Learning to Communicate Using Counterfactual Reasoning. *arXiv preprint arXiv:2006.07200* .

Wolpert, D. H.; and Tumer, K. 2002. Optimal payoff functions for members of collectives. In *Modeling complexity in economic and social systems*, 355–369. World Scientific.

Yang, Y.; Wen, Y.; Chen, L.; Wang, J.; Shao, K.; Mguni, D.; and Zhang, W. 2020. Multi-Agent Determinantal Q-Learning. *arXiv preprint arXiv:2006.01482* .

Zhang, Y.; and An, B. 2020. Computing Ex Ante Coordinated Team-Maxmin Equilibria in Zero-Sum Multiplayer Extensive-Form Games. *arXiv preprint arXiv:2009.12629* .

Zinkevich, M.; Johanson, M.; Bowling, M.; and Piccione, C. 2008. Regret minimization in games with incomplete information. In *Advances in neural information processing systems*, 1729–1736.