Increased Optimism in Multi-Agent Policy Gradients

Michael Sullins
University of Illinois at Chicago
Chicago, IL
sullins2@uic.edu

Ian A. Kash
University of Illinois at Chicago
Chicago, IL
iankash@uic.edu

ABSTRACT

Multi-agent policy gradient algorithms inspired by regret minimization have recently found success in a variety of deep reinforcement learning settings. Stabilizing agents' policies during learning remains challenging and important, but is typically based on the use of entropy regularization. Motivated by optimistic gradient methods with last iterate convergence guarantees, we present our work-in-progress introducing a small modification to existing multi-agent policy gradient algorithms to include (increased) optimism. The particular instantiation we study, an application to Neural Replicator Dynamics, corresponds to Optimistic Hedge in the single-state tabular case and achieves empirical last iterate convergence in the benchmark games of Rock, Paper, Scissors and Kuhn Poker. We also demonstrate that the amount of optimism is robust; increased optimism can increase the rate of convergence.

KEYWORDS

No-regret learning; reinforcement learning; CFR; optimism; multi-agent; policy gradients

1 INTRODUCTION

Designing multi-agent reinforcement learning (MARL) algorithms that converge to equilibria remains a challenging problem. Counterfactual Regret Minimization [24], has empirical successes [3, 15] and strong theoretical convergence guarantees, but only in terms of the time-averaged policy. Recent work has focused on extending CFR to function approximation and policy gradients ([18], [16], [2]). These methods still suffer however from the typical result of no-regret dynamics: it is the time-averaged policy with guarantees of convergence. The current policy lacks any guarantees of converging to anything meaningful, such as a Nash equilibrium. This problem is compounded as the policies are represented by a neural network, where averaging becomes even more problematic. In practice, heuristic approaches such as clipping regrets (CFR+ [21]) or entropy regularization[1] [16, 18] are used to achieve last iterate convergence.

In this paper, representing our work-in-progress, we propose a modification to existing multi-agent policy gradient algorithms that instead achieves last iterate convergence of the policy using optimism. The basic version of optimism involves adding the current gradient twice while subtracting the previous gradient. Of practical importance, our method only requires a small change to store the previous iteration's gradients. Furthermore, our method's tunable hyperparameters allow for increased amounts of optimism that can potentially improve performance in some settings.

More specifically, we demonstrate our approach on the state-of-the-art policy gradient method Neural Replicator Dynamics (NeuRD) [16]. We extend it to allow (increased) optimism and refer to the resulting algorithm as Generalized Optimistic Neural Replicator Dynamics (GO-NeuRD). We show that GO-NeuRD is theoretically grounded as it reduces to Optimistic Hedge using standard optimistic counts of two in the single-state tabular setting, in the same way that NeuRD reduces to Hedge. Since Optimistic Hedge provably achieves last iterate convergence [5] in this setting, it immediately follows this holds for GO-NeuRD as well (at least with exact updates).

On the empirical side, we demonstrate how using increased levels of optimism accelerates last iterate convergence in the benchmark games of Kuhn poker and Rock, Paper, Scissors. To our knowledge, the former represents the first application of increased optimism in settings with imperfect information, although our prior work has demonstrated empirical results using increased optimism in Markov Games [10]. Our results are complementary, suggesting the applicability of (increased) optimism in a broad range of settings.

We conclude this description of our work-in-progress by discussing a number of directions where further work is needed as well as larger open problems.

2 RELATED WORK

When multiple agents interact and learn in the same environment and at the same time, each agent perceives its environment as changing in time. This is the problem of non-stationarity, which makes multi-agent learning considerably more difficult than the single-agent setting. Counterfactual Regret Minimization (CFR) [24] is an algorithm which has had remarkable success in this area, specifically in two-player imperfect information extensive-form games such as Poker [3, 15] and enjoys convergence guarantees in the tabular setting. It has been extended in several ways to include function approximation ([2], [22]). CFR still has some limitations: perfect recall of agents is required, theoretical guarantees don't extend past the original setting, and terminal states are needed for the recursive computation of counterfactuals. However, most CFR-style approaches are not guaranteed to achieve last iterate convergence.

The closest related work to this paper is [6], where zero sum extensive form games were analyzed via the use of dilated distance generating functions. It was shown that the Mirror Descent algorithm, and its optimistic variant, could be decomposed into local regret minimizers in each information set, satisfying the main CFR theorem. They empirically show that optimistic versions of this approach achieve last iterate convergence in settings including

[1] While we were in the final stages of preparing this, a paper was posted on arXiv that gives a principled interpretation of entropy regularization as transforming the game to one with slightly different rewards (and thus equilibria) but better convergence properties [17]. However, the final convergence of the policy to the actual equilibrium still relies on heuristically decaying the weight on the regularization term to effectively solve a sequence of games that are closer and closer approximations of the true game.

Kuhn poker, but do not examine how these results can be applied to policy gradient methods or the importance of increased optimism.

In [10], the Local No-Regret Learning (LONR) algorithm was introduced which uses a copy of a no-regret algorithm in each state to minimize total regret in a value-iteration style. LONR relaxes the assumptions of CFR, primarily the reliance on perfect recall and terminal states and is provably convergent in Markov Decision Processes (and modest extensions). Empirically, it was tested on a class of two-player, general-sum Markov Games specifically designed to be problematic for learning agents. Convergence to equilibrium failed with most regret minimizers, but was achieved in the last iterate through increased optimism using Optimistic Hedge as the underlying no-regret algorithm. (Standard optimism was insufficient.) This paper complements that work by exploring the use of (increased) optimism in policy gradient methods and providing theoretical guarantees in a simple case.

Policy gradient methods based on regret minimization and CFR have attempted to fuse CFR's central idea of local regret minimization into broader settings, specifically model-free online learning. Advantage-based Regret Minimization (ARM) [9] learns a cumulative clipped advantage function that learns well in single-agent partially observable environments. Actor-critic policy gradients (PG) with connections to CFR have been studied in [18], but theoretically require a costly ℓ_2 projection over the simplex. This is problematic enough that softmax policies were used practically instead, which is typical for most policy gradient methods. Neural Replicator Dynamics (NeuRD) addresses this issue by introducing a fix that allows for softmax policies [16]. It has strong connections to PG, decomposes into Hedge in the tabular single-state setting, and has connections to CFR run with Hedge. NeuRD corresponds to PG with a modified update rule that bypasses the gradient through the softmax layer, resulting in its benefits coming at a minimal cost change ('one line fix') to PG. This issue of last iterate convergence is addressed by the authors, who use a non-standard form of entropy regularization to drive the current policy to the equilibrium [17].

Motivated in part by stabilizing training of Generative Adversarial Networks [4, 7] which can exhibit cyclical behavior, modifications to standard gradient descent methods and no-regret algorithms have been studied with the goal of achieving last iterate convergence. In [1], the Multiplicative Weights Update (MWU, referred to from this point as Hedge) algorithm was studied under the KL-divergence between the Nash policy and current policy, where a non-negative lower bound was proven, showing divergence in games with interior equilibria. In [4], Optimistic Mirror Descent (OMD) is proposed to train Wasserstein GANs and OMD is shown to provably converge in the last iterate for a large class of zero sum games. [12] study last iterate convergence to saddle-points in unconstrained convex-concave min-max optimization problems using Optimistic Gradient Descent/Ascent. The constrained setting was studied in [5] with the same last iterate guarantees using Optimistic Multiplicative Weights Update (OMWU, Optimistic Hedge). Optimistic Hedge dynamics are described by the authors as two stages: monotonic improvement of the KL-divergence of the current iterate to the min-max solution, upon which it enters a neighborhood of the solution and becomes a contraction map converging to the exact solution. As our method decomposes to Optimistic Hedge in the single-state, tabular case, examining these dynamics and changes induced by increased optimism in broader settings such as the extensive-form games (EFGs) studied here is a line for future work.

[14] give an analysis of OGDA (and other related algorithms) as an approximation of the proximal point method. Importantly for our work, they also study OGDA with generalized optimism, which provides an approach to extending convergence results to a broader range of parameters.

3 PRELIMINARIES

In this section, we provide the necessary background needed to reach our main results. As we will be building on the results from [16], we closely follow their notation and formulation.

3.1 Normal Form Games

A *normal-form* game (NFG) consists of a finite set of N players each with sets of actions $\{A_1, ..., A_N\}$ and a reward function $\mathbf{u}$: $A_1 \times A_2 \times ... \times A_N \mapsto R^N$ for each joint action $\mathbf{a} \doteq (a_1, a_2, ..., a_N)$ that is a numerical value. The mixed strategy (policy) of player i is $\pi_i \in \Delta(A_i)$, where Δ is the probability simplex over the given set. The strategy profile $\pi = (\pi_1, ..., \pi_N)$, where π_{-i} is the strategy profile of every player except player i. The expected value for player i is $\overline{u}_i(\pi) \doteq \mathbb{E}_\pi[u_i(a)|a \sim \pi]$ Define the best response for player i as $BR_i(\pi_{-i}) \doteq \text{argmax}_{\pi_i}(\overline{u}_i(\pi_i, \pi_{-i}))$ The strategy profile π^* is a Nash equilibrium is every player is best responding $\pi_i^* \in BR_i(\pi_{-i}^*)$ for all $i \in N$. To evaluate policies, the *NashConv* is used to assess their quality. $NashConv(\pi) = \Sigma_{i \in N}\overline{u_i}((BR_i(\pi_{-i}), \pi_{-i})) - \overline{u_i}(\pi)$. *NashConv* is referred to as exploitability in two player games.

3.2 No-regret Learning

One natural goal of an online learning algorithm is to minimize the (external) regret, which compares the actions chosen by an agent with the hindsight optimal action. An algorithm is considered no-regret if its total regret experienced grows at $o(T)$, which implies its average regret goes to zero.

A well-studied class of no-regret algorithms is Follow the Regularized Leader (FTRL) which includes a strongly-convex regularizer term in the update. When the regularization function is chosen to be the negative entropy, this leads to the Hedge algorithm,

$$\pi_T = \prod(\Sigma_{t=1}^{T-1}\eta_t\overline{\mathbf{u}_\mathbf{i}}^t) \tag{1}$$

where $\prod$ is the softmax operator ($\prod(u_{ia}) \propto e^{u_{ia}}$), T is the time steps, and η is the learning rate.

Hedge, along with all FTRL algorithms (including those gradient descent methods falling into this class) are known to only achieve convergence to equilibria in the average sense (i.e., it is the time-averaged policy that approaches a Nash equilibrium) [13]. Optimistic variants of FTRL algorithms have been studied to overcome this limitation and provide convergence guarantees for the *current* policy [4, 5]. The optimistic variant of Hedge accomplishes this by counting the current utility twice:

$$\pi_T = \prod(\Sigma_{t=1}^{T-1}2\eta_t\overline{\mathbf{u}_\mathbf{i}}^t - \eta_{t-1}\overline{\mathbf{u}_\mathbf{i}}^{t-1}) \tag{2}$$

To the best of our knowledge, increased optimistic variants of no-regret algorithms have received little attention. ([6] consider general estimates of the most recent loss/reward in Optimistic FTRL, but do not consider increased amounts. Empirical experiments with increased optimism were performed in [10]). However, this approach has been studied in gradient methods. FTRL, when the regularizer is chosen to be the squared Euclidean norm ($R(w) = \frac{1}{2\eta}||w||_2^2$). corresponds to the Gradient Descent algorithm with a constant learning rate, thus linking gradient methods and no-regret.

Gradient Descent Ascent (GDA) is gradient descent method for problem settings with a minimizer/maximizer in an objective function $f(x, y)$ that captures among other things zero-sum NFGs. Optimistic Gradient Descent Ascent (OGDA) applies optimism, and incorporates what is intuitively a "negative momentum" term which can be seen by writing out the explicit update rules as in [14]

$$x_{t+1} = x_t - \eta\nabla_x f(x_t, y_t) - \eta(\nabla_x f(x_t, y_t) - \nabla_x f(x_{t-1}, y_{t-1}))$$

$$y_{t+1} = y_t + \eta\nabla_y f(x_t, y_t) + \eta(\nabla_y f(x_t, y_t) - \nabla_y f(x_{t-1}, y_{t-1}))$$

[14] then introduce Generalized OGDA, which is parameterized by hyperparemeters α, β which allow for non-standard amounts of optimism:

$$x_{t+1} = x_t - (\alpha+\beta)\nabla_x f(x_t, y_t) + \beta\nabla_x f(x_{t-1}, y_{t-1})$$

$$y_{t+1} = y_t + (\alpha+\beta)\nabla_y f(x_t, y_t) - \beta\nabla_y f(x_{t-1}, y_{t-1})$$

where $\alpha = \beta$ recovers OGDA. In saddle-point problems under certain conditions ([14], Theorem 5), OGDA remains linearly convergent when a factor other than 2 is used.

3.3 Policy Gradients / Replicator Dynamics

Policy Gradient (PG) methods with function approximation are a reinforcement learning technique that updates the policy parameters directly with respect to the gradient of the expected reward [19]. While our approach does not require it, we assume for ease of presentation that the policy is represented as a neural network. More substantively, we assume that the last layer of the network is a softmax layer, so that the policy is determined by the taking the softmax of values in the penultimate layer. We denote these penultimate values $y(a; \theta_t)$.

The goal of reinforcement learning to is maximize the sum of (discounted) expected rewards. A Markov Decision Process is a tuple $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the (finite) action space, $P : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$ is the transition probability kernel, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the (expected) reward function (assumed to be bounded), and $0 < \gamma < 1$ is the discount rate. The expected reward, known as the state value function, is defined as $v^\pi(s) = \mathbb{E}_\pi[\Sigma_{i=t}^\infty \gamma^{i-t} r_t | s_t = s]$ and the action value function is $q^\pi(s, a) = \mathbb{E}_\pi[\Sigma_{i=t}^\infty \gamma^{i-t} r_t | s_t = s, a_t = a]$. In actor-critic methods, the policy π is parameterized by θ and the value function by w. According to the Policy Gradient theorem, updates are made using the gradient $\nabla_\theta \log \pi(a; s, \theta)[q(s, a; w) - b(s; w)]$, where $b(s; w)$ is a variance reducing baseline.

Replicator Dynamics (RD) from Evolutionary Game Theory are a biologically inspired set of operators that describe the evolution of populations. The links between RD and PG are examined in [16] such as FTRL with negative entropy corresponds to RD [16]. As we are now in the RL setting, the previously expected utilities ($\overline{u}$) will be switched to the RL equivalents: namely, to the state-action values q and state values v. The Neural Replicator Dynamics (NeuRD) algorithm updates similarly to PG, with a small modification. First, we restate the parametric update rule of NeuRD to provide some helpful background (Equation 9, [16]):

$$y_{t+1}(a) = y(a; \theta_t) + \eta_t(q^{\pi_t}(a) - v^{\pi_t}) \tag{3}$$

Here, y_{t+1} is the representation of π_{t+1} before the softmax operator is applied. Thus, $y_{t+1}(a)$ can be thought of as a fixed target value that the $y(a; \theta_t)$ are pushed towards. The term in parentheses is referred to in the literature as the *advantange*, which is commonly interpreted as regret. Intuitively, this target value is accumulating regrets, similar to the regret sums tracked in the Regret Matching [8] algorithm. This leads to the NeuRD update rule:

$$\theta_t = \theta_{t-1} + \eta_t\Sigma_a\nabla_\theta y(a; \theta_{t-1})[q^\pi(a) - v^\pi] \tag{4}$$

The NeuRD update rule is thus almost equivalent to the PG update, differing only with respect to where the gradient is taken (i.e., on the y_t rather than the π_t).

4 GENERALIZED OPTIMISTIC NEURAL REPLICATOR DYNAMICS

In this section, we propose a modification to the objective function of the Neural Replicator Dynamics (4) and derive a new update rule from it. Inspired by the results of OGDA presented in Section 3, we propose a composite objective that includes an $(\alpha+\beta)$-weighted Euclidean distance of the current iterate objective together with a β-weighted Euclidean distance of the previous iterate objective as one update:

$$\begin{aligned}\theta_t = \theta_{t-1} &- \Sigma_a\nabla_\theta\frac{\alpha+\beta}{2}||y_t(a) - y(a; \theta_{t-1})||^2 \\ &+ \Sigma_a\nabla_\theta\frac{\beta}{2}||y_{t-1}(a) - y(a; \theta_{t-2})||^2\end{aligned} \tag{5}$$

This composite objective now includes a weighted combination of the regular NeuRD update as well as a portion of the update from the previous iteration. We refer to updates using Equation (5) as Generalized Optimistic Neural Replicator Dynamics (GO-NeuRD). We now state our main result:

Theorem 4.1. *Generalized Optimistic Neural Replicator Dynamics (GO-NeuRD), with $\alpha = \beta = 1$, is equivalent to Optimistic Hedge in the single-state all-actions tabular setting.*

Proof. We closely follow the argument from [16] with minor changes to permit optimism:

$$\begin{aligned}\theta_t &= \theta_{t-1} - \Sigma_a \nabla_\theta \frac{\alpha+\beta}{2} ||y_t(a) - y(a;\theta_{t-1})||^2 \\ &\quad + \Sigma_a \nabla_\theta \frac{\beta}{2} ||y_{t-1}(a) - y(a;\theta_{t-2})||^2 \\ &= \theta_{t-1} + (\alpha+\beta)\Sigma_a (y_t(a) - y(a;\theta_{t-1}))\nabla_\theta y(a;\theta_{t-1}) \\ &\quad - \beta \Sigma_a (y_{t-1}(a) - y(a;\theta_{t-2}))\nabla_\theta y(a;\theta_{t-2}) \\ &= \theta_{t-1} + (\alpha+\beta)\eta_t \Sigma_a \nabla_\theta y(a;\theta_{t-1})(q^{\pi^t} - v^{\pi^t}) \\ &\quad - \beta \eta_{t-1} \Sigma_a \nabla_\theta y(a;\theta_{t-2})(q^{\pi^{t-1}} - v^{\pi^{t-1}})\end{aligned}$$

In the tabular case with only a single state, $\nabla_\theta y(a;\theta_t)$ is the identity matrix, so unrolling the GO-NeuRD update across $T-1$ rounds, we see that the GO-NeuRD policy is

$$\pi_T = \prod(\Sigma_{t=1}^{T-1}(\alpha+\beta)\eta_t(\mathbf{u}^t - \mathbf{u}^t \cdot \pi_t) - \beta\eta_{t-1}(\mathbf{u}^{t-1} - \mathbf{u}^{t-1} \cdot \pi_{t-1}))$$

$$\pi_T = \prod(\Sigma_{t=1}^{T-1}(\alpha+\beta)\eta_t \mathbf{u}^t - \beta\eta_{t-1}\mathbf{u}^{t-1})$$

since $\prod$ is shift invariant. This is equivalent to Optimistic Hedge (2) when $\alpha = \beta = 1$, thus the same policy on every round is used and are therefore equivalent in this setting.

□

Theorem 4.1 links GO-NeuRD with Optimistic Hedge. Since Optimistic Hedge is known to achieve last iterate convergence when used by all players in a NFG [5, 20] it follows that GO-NeuRD does as well when exact updates are used. More broadly, this allows us to tap into the larger literature on last iterate convergence of optimistic methods both for intuition about why adding optimism to policy gradient methods should lead to last iterate convergence more broadly as well as proof techniques that may allow us to make this intuition precise.

The derivations in the proof of Theorem 4.1 also show that when $\eta_t = \eta$ independent of t, it can be absorbed into the choice of α and β. However, we maintain it as a separate parameter for ease of comparison with prior experimental setups.

Algorithm 1 Generalized Optimistic Neural Replicator Dynamics (GO-NeuRD)

```
1:  Initialize policy weights $\theta_0$ and critic weights $w_0$
2:  for $t$ from 0 to $T$ do
3:      $\pi_{t-1}(\theta_{t-1}) \leftarrow \prod(y(\theta_{t-1})$
4:      for $\tau \in$ SampleTrajectories($\pi_{t-1}$) do
5:          for s,a $\in \tau$ do
6:              R $\leftarrow$ Returns($s, \tau, \gamma$)
7:              $w_t \leftarrow$ UpdateCritic($w_{t-1}, s, a, R$)
8:          for $s \in \tau$ do
9:              $v(s; w_t) \leftarrow \Sigma_{a'}\pi(s,a';\theta_{t-1})q_t(s,a';w_t)$
10:             $\theta_t \leftarrow \theta_{t-1}+$
11:             $(\alpha+\beta)\eta_t\Sigma_{a'}\nabla_\theta y(s,a';\theta_{t-1})(q_t(s,a';w_t) - v(s;w_t)$
12:             $-\beta\eta_{t-1}\Sigma_{a'}\nabla_\theta y(s,a';\theta_{t-2})(q_{t-1}(s,a';w_{t-1})$
13:             $-v(s;w_{t-1}))$
```

In our analysis we have described the implementation of GoNeuRD with a single state. In Algorithm 1 we give pseudocode that adapts the NeuRD algorithm. All changes are in the final step (lines 10-13), where the additional gradient terms are included. Following [16], the pseudocode presented is a version that samples trajectories. Our experimental results that follow do a full traversal of all states each iteration (for Kuhn poker). See the discussion and footnote that follows.

Theorem 4.2.

5 EXPERIMENTS

5.1 Experimental Setup

We apply increased optimism to two benchmark settings. The first is Rock, Paper, Scissors, a two-player zero-sum benchmark game. Next we show results for increased optimism in Kuhn poker, a simplified version of Poker. We base our experiments on the implementations of the relevant games and algorithms in OpenSpiel: A Framework for Reinforcement Learning in Games [11].

As our method is an extension of the Neural Replicator Dynamics algorithm,[2] we use of the deep NeuRD feed-forward network provided, as well as the Counterfactual Solver. To simplify the comparison as much as possible, all optional features of the networks were disabled. This includes no hidden features, no skip connections, and no autoencoder. The network for Kuhn Poker consists of two hidden layers of size 128, with ReLU activations. For RPS, we used a smaller network of one hidden layer of size 13. No entropy regularization was added, as our goal is to demonstrate the efficacy of optimism as an alternative. A batch size of 100 was used, which was shuffled and repeated once. A threshold of [-3, 3] was used (we did not thoroughly investigate the impact that thresholding the logit-gap has on the policy.) Training is done in a centralized manner through self-play.

In all experiments, the representational power of the neural network is fixed (i.e. the logit-gap is constant and low) We suspect this limits the precision to which any of the policies can reach. We leave analysis of these factors for future work and instead focus on the effects of optimism.

We leave out a comparison to the entropy regularization used in the original NeuRD experiments as it was not available in time. In the future, we would like to compare with the entropy they used.

When increased optimism is used, this corresponds to setting the parameter $\alpha = 1$ and putting the remainder in β. For example, an optimistic count of 5 would set $\alpha = 1$, $\beta = 4$.

5.2 Rock, Paper, Scissors

Here we present results on Rock, Paper, Scissors (RPS), a simple but widely used benchmark game. Each example uses a learning rate $\eta = 0.1$. RPS is known to cause the policy of FTRL algorithms to diverge and oscillate around its interior Nash equilibrium $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. NeuRD thus exhibits this behavior.

We show three plots with varying ranges of optimism. In Figure 1a, we compare standard optimism of 2 with increased amounts

[2] The OpenSpiel implementation of NeuRD differs from the one presented in [16] in several respects, most notably that it performs updates with all states rather than sampling. As an implementation of the sampling version is not yet available, we leave experiments with it to our future work.

4, 6, 8, and 10. In each case, the level of optimism used converges faster than any of the optimistic counts less than it. An optimistic count of 10 converges very quickly, followed by 8 and 6. Within the total iterations shown, the optimistic count of 2 has not yet converged.

In Figure 1b, we continue to increase the optimism. While each value eventually converges, the pattern of larger counts converging faster does not continue to hold.

In Figure 1c, very high levels of optimism are shown. All counts converge except 80, which converges to 0.0, suggesting that very high levels of optimism are unstable and can diverge.

These experiments demonstrate the benefits of increased optimism that show a clear pattern where increased amounts of optimism cause the policy to converge to the optimal (Nash) policy faster. They also show that very high levels of optimism can be used. The use of optimism is not limitless however; beyond a certain point, convergence begins to take longer. When optimism is too high, the policy fails to converge at all (at least within the window of iterations shown). Identifying how optimistic counts affect the dynamics is an important line for future inquiry, which may be problem-dependent (i.e. may depend on the maximum/minimum reward, the range of reward function, or the condition number of the problem[14].)

5.3 Kuhn Poker

Kuhn Poker is a simplified three card version of poker. The deck consists of a Jack, Queen, and King. Each round, both players begin with 2 chips. The last player remaining, or whoever has the highest rank card at the end, wins. While simple, Kuhn Poker contains all of the interesting properties of full poker such as imperfect information and is modelled as an imperfect-information extensive-form game.

The goal of these experiments is to highlight the benefits of optimism (and increased optimism) in this more complex setting. These are twofold: 1) when optimism is applied, the current policy converges towards a Nash Equilibrium. 2) Increasing optimism can speed up convergence.

In Figure 2a, NeuRD is shown first as a base line. There is a parameterized family of equilibria, and experimentally, we found that the NeuRD policy can vary considerably on different runs. Here we plot a somewhat favorable run that is consistent with other runs. Of particular note is the oscillation of the policy during training. The number of iterations it takes to become relatively stable is high. Thereafter, the policy continues to oscillate (note that the plot line is thicker, and the range in which it oscillates within is larger than the following plots.) In Figure 3a, the exploitability average for 10 runs is shown, and a plateau emerges.

In Figure 2b, GO-NeuRD is plotted with the standard count of optimism of two. The policy stabilizes sooner but continues a smooth drift throughout the rest of the run. Interestingly, while the policy is drifting, the entire strategy profile remains in a (small) $\epsilon-$approximate equilibrium, as can be seen in Figure 3b.

In Figure 2c, GO-NeuRD is plotted with increased optimism of 2.5. The sample policy for this particular run stabilizes marginally faster than with optimism of two, and does not drift as significantly around multiple Nash equilibria. The exploitability in Figure 3c reaches similar values on average to that in Figure 3b. Overall, the benefits of increased optimism in this case are marginal and we found that higher levels of optimism lead to non-convergence.

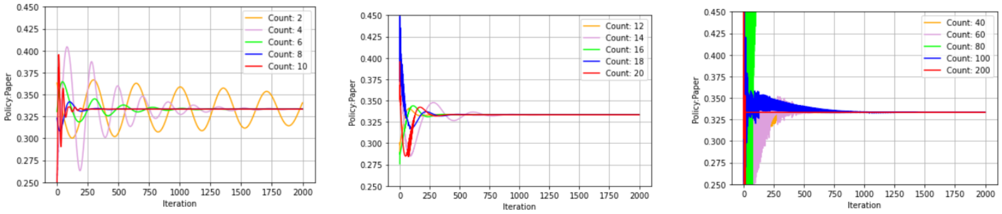
These experiments show that optimism achieves improved convergence. While the optimistic policies are only an order of magnitude more precise, we did not optimize the settings of the network representing the policy or training parameters such as the step size. These may also be a factor in the limited benefits of increased optimism, and we plan to explore this more in the future.

6 CONCLUSION

In this work-in-progress, we have shown how policy gradient approaches to multiagent RL, such as Neural Replicator Dynamics can be extended to use (increased) optimism. We have shown that our extended version of NeuRD corresponds to Optimistic Hedge in the single state case, which provably has last iterate convergence. More broadly, there is a strong theoretical basis for the ability of optimistic methods to achieve last iterate convergence. Additionally, we have emphasized the importance of tuning the degree of optimism to control the rate and quality of convergence.

There are a number of directions for further work. In this work we have experimented with adding optimism in the context of known models. For reinforcement learning, we are interested in the ability of optimism to achieve last iterate convergence with sampling as well, and indeed NeuRD has been shown to work with sampling, providing a natural next set of experiments. Similarly, is adding optimism to actor-critic policy gradient methods (e.g. [18]) which are less precisely tied to regret minimization still effective?

Another important direction is exploring the effects of optimism, both theoretically and empirically, in richer settings. In [10], an increased amount of optimism was applied to local versions of Optimistic Hedge in general sum Markov Games and last iterate convergence was experimentally achieved. This suggests that perhaps policy gradient approaches, such as GO-NeuRD, have potential in imperfect information general sum EFGs / Markov Games. Substantial work is also needed to move beyond these empirical results to provide theoretical guarantees in settings substantially beyond zero-sum normal-form games. Relatedly, can we characterize the optimal amount of optimism to apply in each setting? There has been some recent work along these lines for bilinear games using spectral methods [23]. An alternative to adding optimism is "extra-gradient" methods, which calculate 2 gradients rather than one per iteration (the second one can be thought of as first taking a "half" step). This results in a greater computational cost per iteration, but in centralized training setups where this is feasible the decrease in the number of iterations needed may outweigh this in some settings[14].

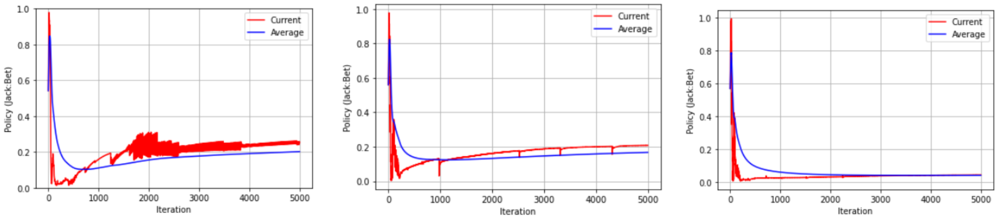


(a) GO-NeuRD optimism [2, 10]

(b) GO-NeuRD optimism [12, 20]

(c) GO-NeuRD optimism [40, 200]

Figure 1: The current policy for the action Scissors in RPS with varying amounts of optimism. (a) Increased optimism leads to faster convergence. (b) Higher levels of optimism still converge (c) Increasing the optimism too high leads to erratic behavior in some counts (optimism of 80 converges to 0.0)

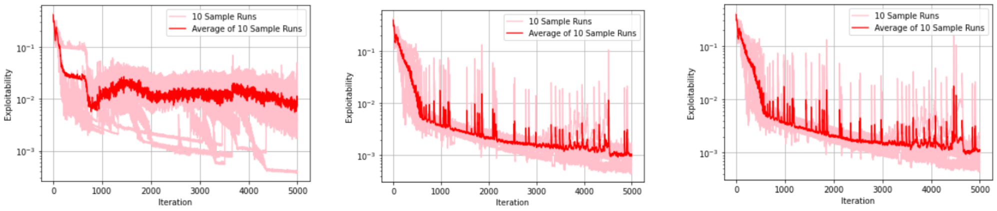


(a) NeuRD (stepsize = 1)

(b) GO-NeuRD (2)

(c) GO-NeuRD (2.5)

Figure 2: The current policy for the action Jack:Bet in Kuhn Poker for a sample run. (a) The NeuRD policy has no convergence guarantees and oscillates. (b) GO-NeuRD with an optimistic count of 2 helps stabilize the current policy. (c) Increasing the optimism to 2.5 stabilizes the current policy sooner in a sample run.



(a) NeuRD (stepsize = 1)

(b) GO-NeuRD (2)

(c) GO-NeuRD (2.5)

Figure 3: The exploitability of the current policy for Kuhn Poker for 10 sample runs (a) NeuRD current policy (b) GO-NeuRD current policy. and (c) increased optimism of 2.5 for GO-NeuRD

REFERENCES

[1] James P Bailey and Georgios Piliouras. 2018. Multiplicative Weights Update in Zero-Sum Games. In *Proceedings of the 2018 ACM Conference on Economics and Computation*. ACM, 321–338.

[2] Noam Brown, Adam Lerer, Sam Gross, and Tuomas Sandholm. 2018. Deep Counterfactual Regret Minimization. *arXiv preprint arXiv:1811.00164* (2018).

[3] Noam Brown and Tuomas Sandholm. 2017. Libratus: the superhuman AI for no-limit poker. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*.

[4] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. 2017. Training GANs with Optimism. *arXiv preprint arXiv:1711.00141* (2017).

[5] Constantinos Daskalakis and Ioannis Panageas. 2018. Last-Iterate Convergence: Zero-Sum Games and Constrained Min-Max Optimization. *arXiv preprint arXiv:1807.04252* (2018).

[6] Gabriele Farina, Christian Kroer, and Tuomas Sandholm. 2019. Optimistic Regret Minimization for Extensive-Form Games via Dilated Distance-Generating Functions. In *Advances in Neural Information Processing Systems*. 5222–5232.

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[8] Sergiu Hart and Andreu Mas-Colell. 2000. A simple adaptive procedure leading to correlated equilibrium. *Econometrica* 68, 5 (2000), 1127–1150.

[9] Peter H Jin, Sergey Levine, and Kurt Keutzer. 2017. Regret Minimization for Partially Observable Deep Reinforcement Learning. *arXiv preprint arXiv:1710.11424* (2017).

[10] Ian A Kash, Michael Sullins, and Katja Hofmann. 2019. Combining no-regret and Q-learning. *arXiv preprint arXiv:1910.03094* (2019).

[11] Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, Daniel Hennes, Dustin Morrill, Paul Muller, Timo Ewalds, Ryan Faulkner, János Kramár, Bart De Vylder, Brennan Saeta, James Bradbury, David Ding, Sebastian Borgeaud, Matthew Lai, Julian Schrittwieser, Thomas Anthony, Edward Hughes, Ivo Danihelka, and Jonah Ryan-Davis. 2019. OpenSpiel: A Framework for Reinforcement Learning in Games. *CoRR* abs/1908.09453 (2019). arXiv:cs.LG/1908.09453 http://arxiv.org/abs/1908.09453

[12] Tengyuan Liang and James Stokes. 2018. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. *arXiv preprint arXiv:1802.06132* (2018).

[13] Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. 2018. Cycles in adversarial regularized learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2703–2717.

[14] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. 2019. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. *arXiv preprint arXiv:1901.08511* (2019).

[15] Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisỳ, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. 2017. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science* 356, 6337 (2017), 508–513.

[16] Shayegan Omidshafiei, Daniel Hennes, Dustin Morrill, Remi Munos, Julien Perolat, Marc Lanctot, Audrunas Gruslys, Jean-Baptiste Lespiau, and Karl Tuyls. 2019. Neural Replicator Dynamics. *arXiv preprint arXiv:1906.00190* (2019).

[17] Julien Perolat, Remi Munos, Jean-Baptiste Lespiau, Shayegan Omidshafiei, Mark Rowland, Pedro Ortega, Neil Burch, Thomas Anthony, David Balduzzi, Bart De Vylder, et al. 2020. From Poincar\'e Recurrence to Convergence in Imperfect Information Games: Finding Equilibrium via Regularization. *arXiv preprint arXiv:2002.08456* (2020).

[18] Sriram Srinivasan, Marc Lanctot, Vinicius Zambaldi, Julien Pérolat, Karl Tuyls, Rémi Munos, and Michael Bowling. 2018. Actor-critic policy optimization in partially observable multiagent environments. In *Advances in Neural Information Processing Systems*. 3422–3435.

[19] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. 2000. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*. 1057–1063.

[20] Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. 2015. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems*. 2989–2997.

[21] Oskari Tammelin, Neil Burch, Michael Johanson, and Michael Bowling. 2015. Solving Heads-Up Limit Texas Hold'em.. In *IJCAI*. 645–652.

[22] Kevin Waugh, Dustin Morrill, James Andrew Bagnell, and Michael Bowling. 2015. Solving Games with Functional Regret Estimation.. In *AAAI*, Vol. 15. 2138–2144.

[23] Guojun Zhang and Yaoliang Yu. 2019. Convergence Behaviour of Some Gradient-Based Methods on Bilinear Games. *arXiv preprint arXiv:1908.05699* (2019).

[24] Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. 2008. Regret minimization in games with incomplete information. In *Advances in neural information processing systems*. 1729–1736.