REGULARIZATION FOR STRATEGY EXPLORATION IN EMPIRICAL GAME-THEORETIC ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

In iterative approaches to empirical game-theoretic analysis (EGTA), the strategy space is expanded incrementally based on analysis of intermediate game models. A common approach to *strategy exploration*, represented by the double oracle algorithm, is to add strategies that best-respond to a current equilibrium. This approach may suffer from overfitting and other limitations, leading the developers of the policy-space response oracle (PSRO) framework for iterative EGTA to generalize the target of best response, employing what they term *meta-strategy solvers* (MSSs). Noting that many MSSs can be viewed as perturbed or approximated versions of Nash equilibrium, we adopt an explicit regularization perspective to the specification and analysis of MSSs. We propose a novel MSS called *regularized replicator dynamics* (RRD), which simply truncates the process based on a regret criterion. We show that the regularization approach exhibits desired properties for strategy exploration and RRD outperforms existing MSSs in various games. We extend our study to three-player games, for which the payoff matrix is cubic in the number of strategies and so exhaustively evaluating profiles may not be feasible. We propose a profile search method that can identify solutions from incomplete models, and combine this with iterative model construction using a regularized MSS. Finally, we suggest an explanation for the effectiveness of regularization demonstrated in our experiments.

1 INTRODUCTION

The term *empirical game-theoretic analysis* (EGTA) (Tuyls et al., 2020; Wellman, 2016) describes a broad set of methods for game reasoning with models based on simulation data. Many of multiagent systems of interest are not easily expressed or tackled analytically, and EGTA offers an alternative approach whereby a space of strategies is examined through simulation, combined with game model induction and inference. The number of strategies that can be explicitly incorporated in game models is significantly limited by computational constraints, hence the selection of strategies to include is pivotally important. For accurate analysis results, we require that the included strategies are high-performing and cover the key strategic issues (Balduzzi et al., 2019). The challenge of efficiently assembling an effective portfolio of strategies for EGTA is called the *strategy exploration* problem (Jordan et al., 2010).

Strategy exploration in EGTA is most clearly formulated within an iterative procedure, whereby generation of new strategies is interleaved with game model estimation and analysis. The *Policy Space Response Oracle* (PSRO) algorithm of Lanctot et al. (2017) provides a flexible framework for iterative EGTA, where at each iteration, new strategies are generated through reinforcement learning (RL). The learning player trains in an environment where other players are fixed in a profile (pure or mixed) comprising strategies from previous iterations. The key design question is how to set the other-player profile to be employed as a training target. In PSRO, the component that derives this target is called a *meta-strategy solver* (MSS), as it takes an empirical game model as input and "solves" it to produce the target profile. The learning agent then employs RL to search for a strategy best-responding to the MSS target. In effect, specifying an MSS defines the strategy exploration method for PSRO.

An obvious choice for MSS is the solution concept employed as the objective game analysis, often Nash equilibrium (NE). Incrementally adding strategies that are best-responses to NE of the current

strategy set is known as the *double oracle* (DO) algorithm (McMahan et al., 2003), and PSRO with NE as MSS is essentially DO with RL for computing (approximate) best-response. Though DO is often effective, there is ample evidence that best-response to NE is not always the best approach to strategy exploration. Schvartzman & Wellman (2009b) observed cases where it would approach true equilibrium extremely slowly, such that even adding random strategies could provide substantial speedups. More generally, Lanctot et al. (2017) argued that best-responding to Nash overfits to the current equilibrium strategies, and thus tends to produce results that do not generalize to the overall space. This was indeed their major motivation for defining a generalized MSS concept for strategy exploration. For example, as an alternative MSS Lanctot et al. (2017) proposed *projected replicator dynamics* (PRD), which employs a replicator dynamics (RD) search for equilibrium (Taylor & Jonker, 1978; Smith & Price, 1973), truncating the replicator updates to ensure a lower bound on probability of playing each pure strategy.

In this study, we take a further step in this direction and adopt an explicit regularization perspective to the specification and analysis of MSSs. We propose a novel MSS called **Regularized Replicator Dynamics** (RRD), which truncates the NE search process in intermediate game models based on a regret criterion. Specifically, at each iteration of PSRO, we update players' strategy profile with RD in the empirical game, stopping if the regret of current profile with respect to the empirical game meets certain regret threshold. We show that RRD exhibits many desired proprieties for strategy exploration compared to previous MSSs (e.g., adjust exploration based on the potential of strategies). In terms of convergence, we prove that PSRO with RRD converges to a game model containing profiles with regret bound related to the selected threshold, and show that RRD finds profiles with regret much lower than the theoretical regret bound in practice, which we claim is one typical advantage of exploring strategies with empirical game models. We demonstrate the performance of RRD in various games and show that RRD outperforms several existing MSSs in terms of convergence rate and quality of intermediate empirical game models. Moreover, we investigate several features of learning with RRD and find that the superior learning performance obtained by RRD is not sensitive to the precision of regret threshold but highly relies on the specific search procedure dictated by RRD, that is, using other profiles with the same regret bound as best-response targets would yield much worse learning performance compared to using profiles proposed by RRD.

When the number of players increases, the cost of analyzing a game model substantially ascends in PSRO since the payoff matrix grows exponentially in the number of players. To mitigate this issue, we propose a PSRO-compatible profile search method, called **Backward Profile Search** (BPS), which finds solution concepts without simulating the whole payoff matrix. We combine RRD with BPS by only applying RRD to the subgame proposed by BPS, and we demonstrate the effectiveness of RRD in a 3-player game.

Finally, we investigate the cause of the effectiveness of regularization through experiments and find that regularization could significantly reduce the regret of best-response targets with respect to the full game, which we claim largely accounts for the improved performance.

Contributions of this study include:

1. A novel MSS RRD that truncates the NE search process in intermediate game models based on a regret criterion. Our MSS exhibits many desired proprieties for strategy exploration compared to previous MSSs. We demonstrate that our MSS outperforms several existing ones in various games under a convincing evaluation criterion for EGTA.
2. A comprehensive analysis on learning with RRD, including theoretical convergence guarantee as well as various features observed through experiments.
3. A new profile search method compatible with the iterative model construction, which finds solution concepts without simulating the whole payoff matrix. We combine RRD with the search method and show its effectiveness for learning in a 3-player game.
4. A novel explanation of the effectiveness of regularization in PSRO.

2 Related Work on Strategy Exploration

The first instance of automated strategy generation in EGTA was a genetic search over a parametric strategy space, optimizing performance against an equilibrium of the empirical game (Phelps et al.,

2006). Schvartzman & Wellman (2009a) deployed tabular RL as a best-response oracle in EGTA for strategy generation. These same authors framed the general problem of *strategy exploration* in EGTA and investigated whether better options exist beyond best-responding to an equilibrium (Schvartzman & Wellman, 2009b). Jordan et al. (2010) further extended this line of work by adding strategies that maximize the deviation gain from an empirical rational closure.

Investigation of strategy exploration was advanced significantly by introduction of the PSRO framework (Lanctot et al., 2017). PSRO applied deep RL as an approximate best-response oracle to certain designated other-agent profile selected by the MSS. When employing NE as MSS, PSRO reduces to the DO algorithm (McMahan et al., 2003). To generate strategy effectively, Lanctot et al. (2017) balanced between overfitting to NE and generalizing to the strategy space outside the empirical game, and proposed *projected replicator dynamics* (PRD), which employs an RD search for equilibrium (Taylor & Jonker, 1978; Smith & Price, 1973) and ensures a lower bound on probability of playing each pure strategy. For notation simplicity, we refer to a MSS as PSRO with that MSS when the context is unambiguous. For example, we refer to PRD as PSRO with PRD.

Following the line of PSRO, many works present MSS instances that *regularize* the target profile to prevent from best-responding to an exact equilibrium. Specifically, Wang et al. (2019) employed a mixture of NE and uniform, which essentially samples whether to apply DO or FP for every PSRO iteration, thus illustrating the possibility of combining MSSs, Wright et al. (2019) added a fine-tuning step to DO, which fine-tunes the generated policy network against a mix of previous equilibrium strategies. Balduzzi et al. (2019) introduced a new MSS, called *rectified Nash*, designed to increase diversity of empirical strategy space. Dinh et al. (2021) proposed a MSS for two-player zero-sum game that applies online learning to the empirical game and outputs the online profile as a best-response target. Beyond selecting NE as a solution concept, Muller et al. (2020) proposed a new MSS based on an evolutionary-based concept, α-rank Omidshafiei et al. (2019), within the PSRO framework.

PSRO also generalizes many classic learning dynamics in game theory. For example, selecting a uniform distribution over current strategies as MSS essentially reproduces the classic *fictitious play* (FP) algorithm Brown (1951). Moreover, an MSS that iteratively best-responds to the most recent strategy duplicates the iterated best response algorithm. Note that those algorithms do not substantively rely on the empirical game since they derive from the strategy sets directly.

3 PRELIMINARIES

A normal-form game $\mathcal{G} = (N, (S_i), (u_i))$ is a tuple of a finite set of players N indexed by i; a non-empty set of strategies S_i for player $i \in N$; and a utility function $u_i : \prod_{j \in N} S_j \to \mathbb{R}$ for player $i \in N$, where $\prod$ is the Cartesian product.

A mixed strategy σ_i is a probability distribution over strategies in S_i, with $\sigma_i(s_i)$ denoting the probability player i plays strategy s_i. We adopt conventional notation for the other-agent profile: $\sigma_{-i} = \prod_{j \neq i} \sigma_j$. Let $\Delta(\cdot)$ represent the probability simplex over a set. The mixed strategy space for player i is given by $\Delta(S_i)$. Similarly, $\Delta(S) = \prod_{i \in N} \Delta(S_i)$ is the mixed profile space.

Player i's *best response* to profile σ is the subset of strategies yielding maximum payoff for i, fixing the other players' strategies:

$$br_i(\sigma_{-i}) = \underset{\sigma_i' \in \Delta(S_i)}{\operatorname{argmax}} u_i(\sigma_i', \sigma_{-i}).$$

Let $br(\sigma) = \prod_{i \in N} br_i(\sigma_{-i})$ be the overall best-response correspondence for a profile σ. A Nash equilibrium (NE) is a profile σ^* such that $\sigma^* \in br(\sigma^*)$.

Player i's *regret* in profile σ in game $\mathcal{G}$ is given by

$$\rho_i^{\mathcal{G}}(\sigma) = \max_{s_i' \in S_i} u_i(s_i', \sigma_{-i}) - u_i(\sigma_i, \sigma_{-i}).$$

Regret captures the maximum player i can gain in expectation by unilaterally deviating from its mixed strategy in σ to an alternative strategy in S_i. An NE strategy profile has zero regret for each player. A profile is said to be an ϵ-Nash equilibrium (ϵ-NE) if no player can gain more than ϵ by

unilateral deviation. The regret of a strategy profile σ is defined as the sum over player regrets:

$$\rho^{\mathcal{G}}(\sigma) = \sum_{i \in N} \rho_i^{\mathcal{G}}(\sigma).$$

An *empirical game* $\mathcal{G}_{S \downarrow X}$ is an approximation of the true game $\mathcal{G}$, in which players choose from restricted strategy sets $X_i \subseteq S_i$, and payoffs are estimated through simulation. That is, $\mathcal{G}_{S \downarrow X} = (N, (X_i), (\hat{u}_i))$, where $\hat{u}$ is a projection of u onto the strategy space X. We use the notation $\rho^{\mathcal{G}_{S \downarrow X}}$ to refer to the regret with respect to an empirical game as opposed to the full game.

Solver-based regret (Wang et al., 2021) is the regret (with respect to the full game) of the profile selected by a particular MSS based on an empirical game. For example, NE-based regret is the regret of the NE of an empirical game. When the MSS used in the solver-based regret is same as the MSS used for strategy exploration, we refer to the solver-based regret as *online regret*.

The profile in the empirical game closest to being a solution of the full game is called *minimum-regret constrained-profile* (MRCP) (Jordan et al., 2010). Formally, $\bar{\sigma}$ is an MRCP iff:

$$\bar{\sigma} = \operatorname*{argmin}_{\sigma \in \Delta(X)} \sum_{i \in N} \rho_i^{\mathcal{G}}(\sigma)$$

The regret of MRCP thus provides a natural measure of how well X covers the strategically relevant space (Wang et al., 2021).

We use the term *exploratory probability* to represent the probability assigned to strategies for exploration purpose (e.g., the probability lower bound in PRD and the probability of strategies outside NE support when mixing uniform and NE (Wang et al., 2019)).

Replicator Dynamics (RD) describes the evolution of players' strategy profiles and provides insights into the dynamical characteristics of games. For 2-player games with empirical payoff matrix $\boldsymbol{U} = (\boldsymbol{U}_1, \boldsymbol{U}_2)$, the evolution of probability on the k-th strategy of player 1 and 2 is given by

$$\frac{d\sigma_1^k}{dt} = \sigma_1^k[(\boldsymbol{U}_1\sigma_2)^k - \sigma_1^T\boldsymbol{U}_1\sigma_2], \quad \frac{d\sigma_2^k}{dt} = \sigma_2^k[(\sigma_1^T\boldsymbol{U}_2)^k - \sigma_1^T\boldsymbol{U}_1\sigma_2]$$

where σ_1 and σ_2 are column vectors of strategy probabilities of two players, respectively.

4 Regularization for Strategy Exploration

4.1 Regularization in Regret Space

In previous work (discussed in §2), many MSSs can be viewed as perturbed or approximated versions of NE. The perturbation or approximation are usually achieved by heuristically assigning strategies with fixed minimal exploratory probabilities. Despite the simplicity of this method, the quality of exploration is questionable. For example, consider a perturbation scheme, as employed in PRD, which guarantees each strategy to be sampled with a fixed minimal probability $\delta > 0$. Note that the scheme simply treats all exploratory strategies, i.e., strategies that take the minimal probability δ, strategically equivalent. However, those strategies are not equally important for strategy exploration. Best responding against some exploratory strategies may not approach NE strategies, so they are obviously negligible than those potentially yielding NE strategies in strategy exploration. Therefore, such homogeneous exploration fails to take into account the relative importance of strategies during strategy exploration and may hinder the learning progress. We refer to this as *homogeneous exploration* issue.

Moreover, the selection of the exploratory probability threshold δ is non-trivial in PSRO. Note that the total probability on the exploratory strategies is accumulating as the game model expands, and could become extremely large, thus hindering the learning. Specifically, a large δ could decelerate the learning since a large portion of probabilities are assigned to exploratory strategies, leading NE search astray. On the other hand, if we select a small δ to prevent the total probability from growing fast, the effect of regularization would be not apparent. Therefore, the selection of a fixed exploratory probability is arduous and a dynamic selection scheme is required. We refer to this as *threshold selection* issue.

To mitigate these issues, we propose a new MSS, called *Regularized Replicator Dynamics*, which truncates NE search within the empirical game based on regret information. In brief, at each iteration of PSRO, we update players' strategy profile with RD in the empirical game, stopping if the regret (with respect to the empirical game) of current profile meets certain regret threshold λ. The procedure of PSRO with RRD is shown in Algorithm 1 (PSRO) and Algorithm 2 (RRD), where the latter finds the best-response target for the former. In Algorithm 2, we first initialize RD with a uniform distribution over strategies. Then we apply RD update until the regret (with respect to the empirical game) of current profile in RD is below our regret threshold λ. For example, in 2-player games, we update players' mixed strategies at each RD iteration by $\sigma_1 \leftarrow P(\sigma_1 + \alpha\frac{d\sigma_1}{dt})$ and $\sigma_2 \leftarrow P(\sigma_2 + \alpha\frac{d\sigma_2}{dt})$, where α is a step size for RD and P is a projection operator that maintains the feasibility of mixed strategies, namely $P(\sigma_i) = \text{argmin}_{\sigma_i' \in \Delta} ||\sigma_i' - \sigma_i||$. Since RD may not converge to an equilibrium in certain games, the minimum reachable regret by RD, denoted by λ_{min}, could be non-zero. Therefore, if the selected λ is smaller than λ_{min}, RD would never stop. To handle this issue, we select λ by first running RD for a sufficient large number of iterations T' to obtain λ_{min} and only choosing $\lambda \geq \lambda_{min}$. With this approach, the regret threshold λ is always reachable by RD[1]. Finally, the output profile returns to PSRO as a best response target.

Algorithm 1 PSRO with RRD

Input: initial strategy sets for all players X
Simulate expected utility for $\sigma \in X$
Initialize a meta-strategy σ_i = Uniform(X_i)
while PSRO iteration $i = 1, 2, ..., \mathcal{T}$ **do**
  for player $i \in N$ **do**
    for many RL training episodes **do**
      Sample a profile $s_{-i} \in \sigma_{-i}$
      Train BR oracle s_i' against s_{-i}
    end for
    Add the new strategy to the strategy set
  end for
  Simulate missing utilities for profiles in X
  Compute a meta-strategy $\sigma = \text{RRD}(\mathcal{G}_{S\downarrow X})$
end while
Return An empirical game $\mathcal{G}_{S\downarrow X}$

Algorithm 2 RRD

Input: an empirical game $\mathcal{G}_{S\downarrow X}$
Parameters: regret threshold λ and step size α for RD update
Initialize RD with σ_i = Uniform(X_i)
while $\rho^{\mathcal{G}_{S\downarrow X}}(\sigma) > \lambda$ **do**
  for player $i \in N$ **do**
    $\sigma_i \leftarrow P(\sigma_i + \alpha\frac{d\sigma_i}{dt})$
  end for
end while
Return σ

The desired properties of RRD resolve both the homogeneous exploration issue and the threshold selection issue. First, RRD can be viewed as an exploration approach wherein the exploratory probability on each strategy is directed by RD rather than a fixed value as in prior literature. In other words, RRD controls exploration through the interactions among players based on their utility information, thus taking into account the relative importance of strategies. We give a simple illustration of this property in Appendix G.

Second, RRD adjusts regularization probabilities automatically based on the information collected over PSRO iterations. Since RRD does not rely on any fixed exploratory probability parameters, it would not suffer from the total exploratory probability explosion as PSRO proceeds. Despite RRD requires a regret threshold λ as the stopping criterion for RD, we show that the selection of this threshold turns to be much easier since a wide range of such parameters leads to a superior learning performance (details in Section 5).

4.2 Convergence of Regularization

One natural question to ask is that whether the quality of solution is guaranteed with RRD. To answer this question, we seek a theoretical bound for the regret of solution given by RRD and prove that RRD with a selected regret threshold λ would end up with an empirical game containing profiles that have λ regret or better with respect to the full game, i.e., a λ-NE.

[1]We found that a typical λ_{min} is approximately 1e-6 in our experiments, which is sufficiently small for a claim of convergence. Therefore, the selection of λ is almost non-restrictive. But in other games, RD may reach a profile with a relative large regret.

Theorem 1. Policy Space Response Oracle with Regularized Replicator Dynamics associated with a selected regret threshold λ converges to an empirical game containing at least one λ-NE.

We prove Theorem 1 in Appendix A. The purpose of Theorem 1 is to reveal the authentic convergence property of RRD rather than pursuing a strong NE convergence guarantee. Despite Theorem 1 indicates that the convergence of RRD is not as strong as the convergence of DO in the limit, which converges to exact NE, we focus more on their practical performance under computational budget constraints where the limit (i.e., all strategies are explored) would never be reached.

4.3 Regularized Strategy Exploration in Multi-Player Games

One major obstacle of extending PSRO to games with more than two players is that the size of the empirical game grows exponentially in the number of players and analyzing it becomes difficult. Even in games with few players, an exhaustive simulation of the payoff matrix is not affordable. To mitigate this issue, we propose a PSRO-compatible profile search scheme, called *backward profile search* (BPS), which finds solution concepts without simulating the whole payoff matrix. Our scheme BPS is based on the idea by Brinkman & Wellman (2016) and modified for PSRO compatibility.

BPS searches for an empirical NE starting from the profile constituted by the newest added strategies at current PSRO iteration, denoted as the initial *subgame*, and the NE of the subgame is proposed as a candidate NE. Rather than simulating the whole payoff matrix, BPS only simulates possible deviation profiles to the candidate NE in the empirical game. If no beneficial deviation can be found, the candidate NE is *confirmed* as the NE of the empirical game. Otherwise, the deviation profile is added to the previous subgame to form a new subgame. Then BPS loops from calculating a candidate NE of the new subgame and stops until an empirical NE is confirmed. For strategy exploration, we only apply RRD to the subgame containing a confirm NE, enabling a moderate number of simulation savings as well as a superior learning. We show our BPS algorithm and give details of how RRD is applied in Appendix C.

5 Experimental Results

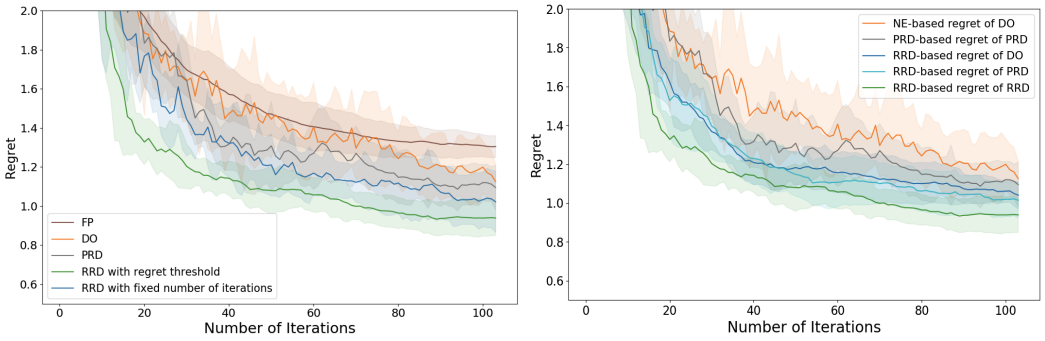
5.1 Performance of Regularized MSSs

We evaluate the performance of RRD in terms of online regret as well as the evaluation criterion specific for PSRO by Wang et al. (2021) in 2-player Leduc poker and real-world games defined by Czarnecki et al. (2020). The online regret metric was widely employed in prior work for comparing online performance of algorithms while the one by Wang et al. (2021) evaluates the learning performance of algorithms in terms of the strategic coverage of the empirical game. We show that RRD outperforms other MSSs with either evaluation metric.

5.1.1 2-player Leduc Poker

In Figure 1a, we show the online regret (with respect to the full game) of DO, PRD, RRD with fixed number of RD iterations and RRD with regret threshold. We first observe that RRD yields a rapid convergence to a low-regret value compared to others. To show the benefits of using a regret threshold as a stopping criterion compared to using a fixed number of RD iterations, we plot the regret curve of RD using a fixed number of iterations. We observe that despite its learning performance outperforms DO and PRD, it is worse than the performance with a regret threshold. This is because the number of RD iterations needed for convergence to a superior profile varies across PSRO iterations. In particular, using an underestimated number of RD iterations may generate a profile with poor performance for strategy exploration while using an overestimated one pushes learning towards DO, thus losing regularization. Moreover, as shown in Appendix B.2, we observe that using a fixed number of iterations could lead to a profile with arbitrarily high regret, which affects the generation of effective strategies.

In Figure 1b, we follow the rule of consistency (Wang et al., 2021), a critical evaluation consideration for EGTA, and verify the learning performance of RRD. According to the consistency criterion, we compare MSSs with the same RRD-based regret and authenticate the faster convergence of RRD over DO and PRD in terms of the strategic coverage of the empirical game.

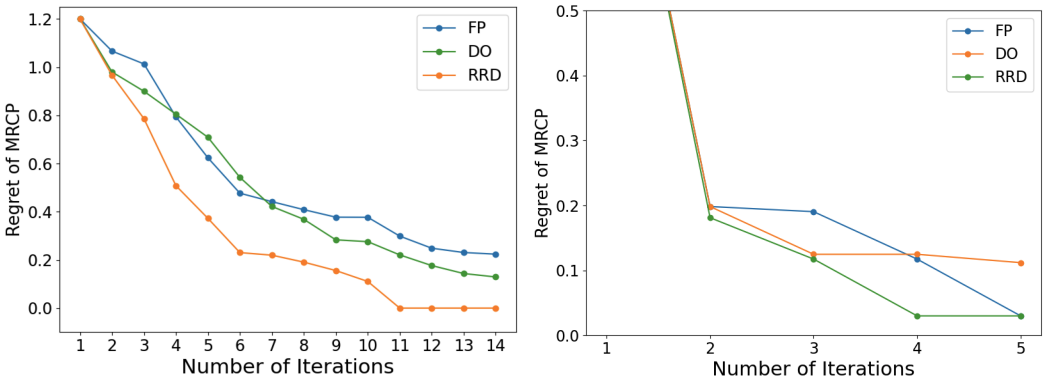


(a) Online regret curves. (b) Regret curves with rule of consistency.

Figure 1: Results for 2-player Leduc poker.

5.1.2 Real-World Games

To further evaluate our algorithm, we select Hex, a two-player board game, in which players attempt to connect opposite sides of a hexagonal board, and a random game of skill (RGS) defined by Czarnecki et al. (2020), whose geometry of strategy space assembles the shape of a spinning top. We evaluate MSSs in terms of the regret of MRCP (averaged over 5 different initial strategies for hex and starting from the uniform strategy for RGS), which is theoretically justified evaluation metric for PSRO when its computation is feasible (Wang et al., 2021). We observe that RRD exhibits faster convergence in both games. We report experimental details (e.g., standard deviation) in Appendix B.3.



(a) Hex. (b) Random game of skill.

Figure 2: Results for real-world games.

5.1.3 Multi-Player Games

We combine BPS with RRD and apply the combination to 3-player Leduc poker. As shown in Figure 3, although RRD is only applied to the subgame of the empirical game, we still gain benefits from regularization. Meanwhile, we list the average of number of simulations required to confirm a NE at different PSRO iterations in Table 1. As shown in Table 1, BPS saves a reasonable number of simulations compared to using the whole payoff matrix. Moreover, the benefit of BPS becomes more apparent as the number of iteration increases.

5.2 Analysis of Learning with RRD

To investigate the stability of learning performance with respect to regret threshold λ, we select a wide range of λs for RRD and compare their learning performance with DO in Figure 4a (i.e., averaged

regret with standard deviation at iteration 100). We find that the range of regret thresholds, yielding a better learning performance than DO, is wide, which is a desirable property for hyperparameter tuning. In addition, we observe that as regret threshold λ increases, the learning performance first improves and then becomes worse. This means that both excessive and inadequate regularization would damage the learning performance.

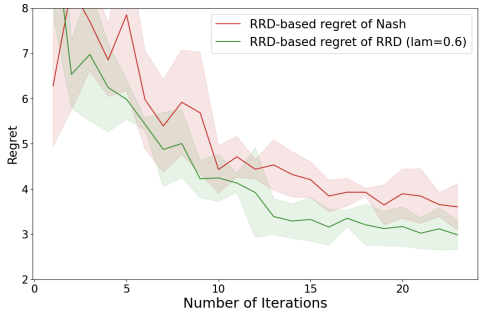


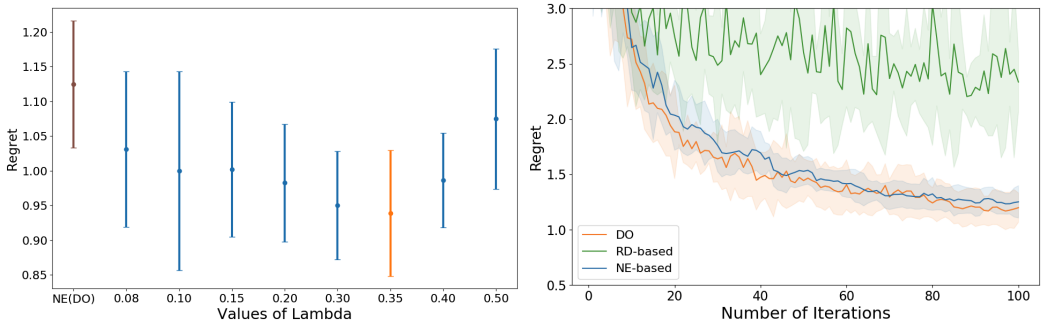
Figure 3: RRD for 3-player Leduc poker

Iter#	#Sims w. BPS	#Sims w/o BPS	Diff.
10	880	1000	120
20	7100	8000	900
25	14200	15625	1425
30	24667	27000	2333
35	39112	42875	3763
40	58400	64000	5600
45	82234	91125	8891
50	112290	125000	12710

Table 1: RRD for 3-player Leduc poker

Besides, we observe that the regret of the RRD profile after convergence verifies our Theorem 1 empirically. Since we set regret threshold $\lambda = 0.35$ for both games, Theorem 1 implies that there must exist a profile with regret lower than or equal to 0.35. In practice, we observe that the regret of MRCP is much lower than $\lambda = 0.35$. This phenomenon indicates a tremendously desirable property of learning in games with EGTA. The property is that distinct from online game learning wherein the online profile is expected to converge to certain solution concept, learning with EGTA succeeds whenever the solution falls into the empirical game, which does not require the convergence of online profiles given by MSSs, and thus reducing the difficulty of learning. This explains why learning with regularization leads to convergence in contrast with being cyclic in online settings (Mertikopoulos et al., 2018).

Finally, motivated by the fact that there exists infinite number of λ-NE in an empirical game, we question whether the superior performance is biased towards the one determined by RRD. We answer this question by assigning RD with a different initialization. Recall that we initialize RRD with a uniform distribution over strategies in the empirical game. We attempt to start RD with a randomly selected profile from the empirical game and stop it with the same regret threshold λ. As shown in Figure 4b, despite outputting a λ-NE, this RD turns to output profiles with terrible online regret. Even following the rule of consistency, its learning performance is worse than DO and hence worse than RRD with uniform initialization. This indicates that the superior performance is biased towards the one determined by RRD and that not all λ-NE exhibit the same capability for being an effective best-response target.



(a) Range of regret threshold.

(b) Non-uniform initial strategies of RD.

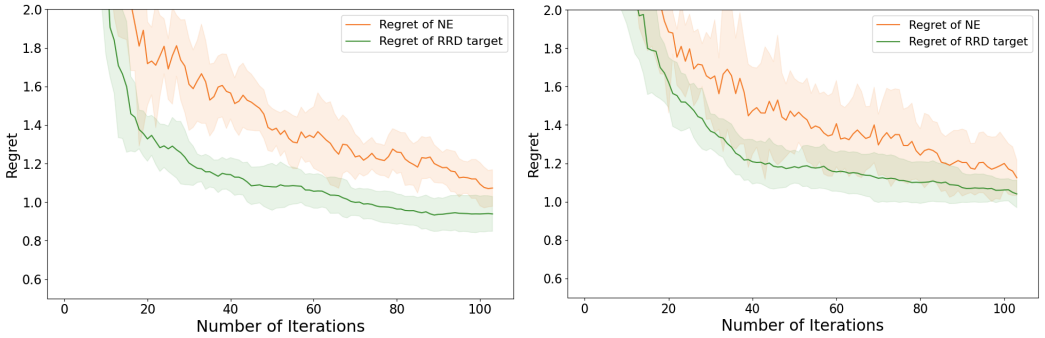
Figure 4: Properties of RRD.

6 WHY AVOID OVERFITTING TO NE?

As discussed in Section 2, many MSSs, which apply regularization to NE, have been proposed and showed benefits for accelerating learning. However, there are very few explanations on why regularization works. We survey some previous efforts that attempt to explain the effectiveness of regularization in Appendix D, and we propose a novel insight based on our experimental observations.

As an iterative learning approach, PSRO predicts players' behavior in an empirical game at each iteration and trains best-response against these strategies to improve the performance of strategies progressively. Since the NE of an empirical game represent players' perfect rationality in the empirical game, it is very natural to select the empirical NE as a training target assuming players would play it at the current iteration. However, we find that in many games empirical NE possess higher regret with respect to the full game than other profiles, which is caused by the limited game information in an empirical game. Our key experimental observation is that the stability of profiles, i.e., regret w.r.t the full game, improves by applying regularization since we not only best respond to strategies in the NE support but also exploratory strategies, thus preventing players from being heavily exploited.

To observe this phenomenon, in Figure 5a and Figure 5b, we consider two PSRO runs with MSS RRD and NE, respectively, and plot the regret of RRD profile and the regret of NE based on the same empirical games in each run. In both plots, we observe that given an intermediate empirical game, the regret (with respect to the full game) of RRD profile is much smaller than that of the empirical NE, which is perfectly stable in the empirical game. Therefore, regularization in fact decreases the regret of target profiles and improves the stability of solutions. We believe that the decrease in regret is the major factor for improved learning performance since the low-regret profile is closer to a true NE and training against such a low-regret profile may speed up overall convergence.



(a) PSRO with RRD.

(b) PSRO with Nash.

Figure 5: Analysis for regularization.

Note that this observation does not mean that the lower-regret profile we select, the better overall learning performance we obtain because the pursue of selecting an extremely stable profile as the training target may result in a slow update of the training target, yielding similar strategies to be added over PSRO iterations. We give an example for illustration and show the performance of using MRCP as a MSS in Appendix E.

7 CONCLUSION

In this work, we propose RRD as a novel MSS for PSRO, which applies regularization in the regret space and prevents overfitting to an empirical NE. Theoretically, we prove the performance guarantee of RRD. In our experiments, we show that RRD outperforms several existing MSSs in various games and investigate many properties of learning with RRD. We extend our study to 3-player games and propose BPS, a PSRO-compatible profile search method that reduces simulation cost, and show the benefit of regularization when combining BPS with RRD in 3-player Leduc poker. Finally, we study the mechanism of regularization behind the learning performance and claim that regularization could significantly change the stabilization of profile targets, thus contributing to a faster learning.

REFERENCES

David Balduzzi, Marta Garnelo, Yoram Bachrach, Wojciech M Czarnecki, Julien Perolat, Max Jaderberg, and Thore Graepel. Open-ended learning in symmetric zero-sum games. In *36th International Conference on Machine Learning*, 2019.

Erik Brinkman and Michael P. Wellman. Shading and efficiency in limit-order markets. In *IJCAI-16 Workshop on Algorithmic Game Theory*, 2016.

George W Brown. Iterative solution of games by fictitious play. *Activity analysis of production and allocation*, 13(1):374–376, 1951.

Wojciech Marian Czarnecki, Gauthier Gidel, Brendan Tracey, Karl Tuyls, Shayegan Omidshafiei, David Balduzzi, and Max Jaderberg. Real world games look like spinning tops. *arXiv preprint arXiv:2004.09468*, 2020.

Le Cong Dinh, Yaodong Yang, Zheng Tian, Nicolas Perez Nieves, Oliver Slumbers, David Henry Mguni, and Jun Wang. Online double oracle. *arXiv preprint arXiv:2103.07780*, 2021.

Patrick R. Jordan, L. Julian Schvartzman, and Michael P. Wellman. Strategy exploration in empirical games. In *9th International Conference on Autonomous Agents and Multi-Agent Systems*, pp. 1131–1138, 2010.

Marc Lanctot, Vinicius Zambaldi, Audrūnas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. A unified game-theoretic approach to multiagent reinforcement learning. In *31st Annual Conference on Neural Information Processing Systems*, pp. 4190–4203, 2017.

Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, et al. OpenSpiel: A framework for reinforcement learning in games. *arXiv preprint arXiv:1908.09453*, 2019.

Stephen McAleer, John Lanier, Pierre Baldi, and Roy Fox. CFR-DO: A double oracle algorithm for extensive-form games. In *AAAI-21 Workshop on Reinforcement Learning in Games*, 2021.

Richard D. McKelvey and Thomas R. Palfrey. Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38, 1995.

Richard D. McKelvey and Thomas R. Palfrey. Quantal response equilibria for extensive form games. *Experimental Economics*, 1(1):9–41, 1998.

Richard D McKelvey, Andrew M McLennan, and Theodore L Turocy. Gambit: Software tools for game theory. 2006.

H. Brendan McMahan, Geoffrey J. Gordon, and Avrim Blum. Planning in the presence of cost functions controlled by an adversary. In *20th International Conference on Machine Learning*, pp. 536–543, 2003.

Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *29th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 2703–2717, 2018.

Paul Muller, Shayegan Omidshafiei, Mark Rowland, Karl Tuyls, Julien Perolat, Siqi Liu, Daniel Hennes, Luke Marris, Marc Lanctot, Edward Hughes, et al. A generalized training approach for multiagent learning. In *8th International Conference on Learning Representations*, 2020.

Eugene Nudelman, Jennifer Wortman, Yoav Shoham, and Kevin Leyton-Brown. Run the GAMUT: A comprehensive approach to evaluating game-theoretic algorithms. In *3rd International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pp. 880–887, 2004.

Shayegan Omidshafiei, Christos Papadimitriou, Georgios Piliouras, Karl Tuyls, Mark Rowland, Jean-Baptiste Lespiau, Wojciech M Czarnecki, Marc Lanctot, Julien Perolat, and Remi Munos. α-rank: Multi-agent evaluation by evolution. *Scientific Reports*, 9(1):1–29, 2019.

S. Phelps, M. Marcinkiewicz, S. Parsons, and P. McBurney. A novel method for automatic strategy acquisition in n-player non-zero-sum games. In *5th International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pp. 705–712, 2006.

L. Julian Schvartzman and Michael P. Wellman. Stronger CDA strategies through empirical game-theoretic analysis and reinforcement learning. In *8th International Conference on Autonomous Agents and Multi-Agent Systems*, pp. 249–256, Budapest, 2009a.

L. Julian Schvartzman and Michael P. Wellman. Exploring large strategy spaces in empirical game modeling. In *AAMAS-09 Workshop on Agent-Mediated Electronic Commerce*, Budapest, 2009b.

J. Maynard Smith and George R. Price. The logic of animal conflict. *Nature*, 246(5427):15–18, 1973.

Peter D. Taylor and Leo B. Jonker. Evolutionary stable strategies and game dynamics. *Mathematical biosciences*, 40(1-2):145–156, 1978.

Karl Tuyls, Julien Pérolat, Marc Lanctot, Edward Hughes, Richard Everett, Joel Z. Leibo, Csaba Szepesvári, and Thore Graepel. Bounds and dynamics for empirical game theoretic analysis. *Autonomous Agents and Multi-Agent Systems*, 34:7, 2020.

Yongzhao Wang, Qiurui Ma, and Michael P. Wellman. Evaluating strategy exploration in empirical game-theoretic analysis. *arXiv preprint arXiv:2105.10423*, 2021.

Yufei Wang, Zheyuan Ryan Shi, Lantao Yu, Yi Wu, Rohit Singh, Lucas Joppa, and Fei Fang. Deep reinforcement learning for green security games with real-time information. In *33rd AAAI Conference on Artificial Intelligence*, pp. 1401–1408, 2019.

Michael P. Wellman. Putting the agent in agent-based modeling. *Autonomous Agents and Multi-Agent Systems*, 30:1175–1189, 2016.

Mason Wright, Yongzhao Wang, and Michael P. Wellman. Iterated deep reinforcement learning in games: History-aware training for improved stability. In *20th ACM Conference on Economics and Computation*, pp. 617–636, 2019.