Evaluating Strategy Exploration in Empirical Game-Theoretic Analysis

Yongzhao Wang*, **Qiuri Ma***, **Michael P. Wellman**
University of Michigan, Ann Arbor
{wangyzh, qmaai, wellman}@umich.edu

Abstract

In empirical game-theoretic analysis (EGTA), game models are extended iteratively through a process of generating new strategies based on learning from experience with prior strategies. The *strategy exploration* problem in EGTA is how to direct this process so to construct effective models with minimal iteration. A variety of approaches have been proposed in the literature, including methods based on classic techniques and novel concepts. Comparing the performance of these alternatives can be surprisingly subtle, depending sensitively on criteria adopted and measures employed. We investigate some of the methodological considerations in evaluating strategy exploration, defining key distinctions and identifying a few general principles based on examples and experimental observations. In particular, we highlight consistency considerations for comparing across different approaches, and the potential for misleading results depending on how intermediate models are evaluated. We present evidence that the *minimum regret constrained profile* (MRCP) provides a particularly robust basis for evaluation, and propose MRCP as a promising concept for heuristic strategy generation.

Introduction

Recent years have witnessed dramatic advances in developing game-playing strategies through iterative application of (deep) reinforcement learning (RL). DeepMind's breakthroughs in Go and other two-player strategy games (Silver et al. 2017, 2018) demonstrated the power of learning through self-play. In self-play the learner generates an improved strategy assuming its opponent plays the current strategy. For many games, iterating best-response in this manner will cycle or otherwise fail to converge, which has led to consideration of alternative approaches to search for improved strategies. For example, DeepMind's milestone achievement in the complex video strategy game StarCraft II entailed an elaborate population-based search approach (Vinyals et al. 2019) informed by game-theoretic concepts.[1]

Many recent works (Lanctot et al. 2017; Balduzzi et al. 2019; Muller et al. 2020) have likewise appealed to game-theoretic methods to direct iterative RL for complex games. At each iteration, a new strategy is generated for one agent through RL, fixing the other agents to play strategies from previous iterations. A general formulation of this approach was presented by Lanctot et al. (2017) as the *Policy Space Response Oracle* (PSRO) algorithm. PSRO can be viewed as a form of *empirical game-theoretic analysis* (EGTA) (Wellman 2006), a general name for the study of building and reasoning about game models based on simulation. In EGTA, game models are induced from simulations run over combinations of a particular set of strategies. The *strategy exploration* problem in EGTA considers how to extend the considered strategy set, based on the current empirical game model. For example, one natural approach is to compute a Nash equilibrium (NE) of the current model, and generate a new strategy that optimizes payoff when other agents play that equilibrium. This approach of iteratively extending strategy sets by best-response to equilibrium was introduced by McMahan, Gordon, and Blum (2003) for two-player games and called the *double oracle* (DO) method.

The authors of PSRO defined an abstract operation on empirical games, termed *meta-strategy solver* (MSS), that extracts an opponent profile from the current empirical game as target for the next best-response calculation. In this framework, choosing an MSS determines the strategy exploration method. For example with NE-calculation as MSS in a two-player game, PSRO reduces to DO. An MSS that simply selects the most recently added strategy corresponds to self-play (SP). A variety of MSSs have been proposed and assessed in the literature on iterative RL-based approaches to games. We survey some of these ideas in the Related Work section, as well as alternative approaches to strategy exploration outside the PSRO framework (e.g., not involving RL or varying from best-response).

Understanding the relative effectiveness of strategy exploration methods across problem settings remains an open problem. We offer some new suggestions on exploration technique in this paper, but the main focus of our study is on how to evaluate alternative methods. For some purposes, the proof of a method is whether it produces a superior solution (e.g., a champion Go program). But for game reasoning more broadly, especially once we go beyond two-player zero-sum games, there is no fixed benchmark to measure a solution against. Moreover, in an iterative EGTA setting, we generally expect to produce a series of increasingly refined

*The two authors contributed equally to this work.

[1] https://deepmind.com/blog/article/AlphaStar-Grandmaster-level-in-StarCraft-II-using-multi-agent-reinforcement-learning

and accurate game models, but may not reasonably expect a definitive solution in any finite amount of time. Therefore, the question is how to evaluate intermediate game models, and how to assess the progress over time of alternative ways to direct the iterative exploration of strategy space.

We delve into several methodological considerations for strategy exploration. First, we seek a theoretically justifiable evaluation metric for empirical games, and recall the proposal by Jordan, Schvartzman, and Wellman (2010) that the regret of the *minimum-regret constrained-profile* (MRCP) can serve this purpose. Calculation of MRCP is not always computationally feasible, so we identify some desiderata for alternative evaluation metrics. Specifically we argue for the importance of focusing on the space of strategies in an empirical game, and highlight some consistency considerations for comparing across different MSSs. We point out the MSS used for evaluation is not necessarily the same as the MSS in strategy exploration and define *solver-based regret* for evaluation purposes. We demonstrate the significance of our considerations in both synthetic and real-world games. Based on the experiments, we hypothesize a connection between learning performance and the regret of profile target proposed by the MSS. We conduct preliminary experiments using MRCP as MSS, which indicate that MRCP is a promising concept for strategy exploration.

Finally, we consider the problem of regret-based evaluation in situations where calculating exact best response is infeasible. We define an encompassing empirical game, called *the combined game*, whose strategy set is a union of all strategies generated across experimental runs. For evaluation purposes, we can calculate the regret of a strategy profile with respect to this combined game. We test this approach in games where exact best responses are available and thus the ground truth performance of different MSSs is known. We find that high-regret profiles in the true game may exhibit low regret in the combined game, thus casting doubt on the accuracy of this approach.

Contributions of this study include:

1. We recall the MRCP as evaluation metric and present evidence that MRCP provides a particularly robust basis for evaluation. We identify some desiderata for alternative evaluation metrics when calculation of MRCP is computationally infeasible;
2. We define solver-based regret for evaluation and highlight some consistency considerations for comparing MSSs, pointing out that MSS used for evaluation is not necessarily the same as the MSS in strategy exploration. We demonstrate the potential for misleading results when consistency is violated in the prior literature;
3. We investigate the properties of MRCP and propose MRCP as a promising concept for strategy exploration;
4. We assess the approach of evaluating strategy exploration using the combined game when calculating exact best response is infeasible.

Related Work on Strategy Exploration

The first instance of automated strategy generation in EGTA was by Phelps et al. (2006), who employed genetic search over a parametric strategy space, optimizing performance against an equilibrium of the empirical game. Schvartzman and Wellman (2009b) combined RL with EGTA in an analogous manner. Questioning whether searching for a best response to equilibrium is an ideal way to add strategies, these same authors framed and investigated the general problem of *strategy exploration* in EGTA (Schvartzman and Wellman 2009a). They identified situations where adding a best response to equilibrium would perform poorly, and proposed some alternative approaches. Jordan, Schvartzman, and Wellman (2010) extended this line of work by proposing exploration of strategies that maximize the gain to deviating from a rational closure of the empirical game.

Investigation of strategy exploration was furthered significantly by introduction of the PSRO framework (Lanctot et al. 2017). PSRO entails adding strategies that are best responses to *some* designated other-agent profile, where that profile is determined by the meta-strategy solver (MSS) applied to the current empirical game. The prior EGTA approaches cited above effectively employed NE as MSS as in the DO algorithm (McMahan, Gordon, and Blum 2003). Lanctot et al. (2017) argued that with DO the new strategy may overfit to the current equilibrium, and accordingly proposed and evaluated several alternative MSSs, demonstrating their advantages in particular games. For example, their *projected replicator dynamics* (PRD) starts with an RD search for equilibrium (Taylor and Jonker 1978; Smith and Price 1973), but then adjusts the result to ensure a lower bound on probability of playing each pure strategy. Any solution concept for games could in principle be employed as MSS, as for example the adoption by Muller et al. (2020) of a recently proposed evolutionary-based concept, α-rank (Omidshafiei et al. 2019), within the PSRO framework.

The MSS abstraction also connects strategy exploration to iterative game-solving methods in general, whether or not based on EGTA. Using a uniform distribution over current strategies as MSS essentially reproduces the classic *fictitious play* (FP) algorithm (Brown 1951), and as noted above, an MSS that just selects the most recent strategy equates to self-play (SP). Note that these two MSS instances do not really make substantive use of the empirical game, as they derive from the strategy sets alone.

Wang et al. (2019) illustrated the possibility of combining MSSs, employing a mixture of NE and uniform which essentially averages DO and FP. Motivated by the same aversion to overfitting the current equilibrium, Wright, Wang, and Wellman (2019) proposed an approach that starts with DO, but then fine-tunes the generated response by further training against a mix of previously encountered strategies.

In the literature, a profile's fitness as solution candidate is measured by its regret in the true game. Jordan, Schvartzman, and Wellman (2010) defined *MRCP* (minimum-regret constrained-profile) as the profile in the empirical game with minimal regret relative to the full game. Regret of the MRCP provides a measure of accuracy of an empirical game, but we may also wish to consider the coverage of a strategy set in terms of diversity. Balduzzi et al. (2019) introduced the term *Gamescape* to refer to the scope of joint strategies covered by the exploration process to a given point. They employed

this concept to characterize the effective diversity of an empirical game state, and propose a new MSS called *rectified Nash* designed to increase diversity of the Gamescape. Finally, we take note of a couple of recent works that characterize Gamescapes in terms of topological features. Omidshafiei et al. (2020) proposed using spectral analysis of the α-rank best response graph, and Czarnecki et al. (2020) visualize the strategic topography of real-world games as a spinning top wherein layers are transitive and strategies within a layer are cyclic.

Preliminaries

A normal-form game $\mathcal{G} = (N, (S_i), (u_i))$ consists 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 any strategy yielding maximum payoff for i, holding the other players' strategies constant:

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

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 $\mathcal{G} = (N, (S_i), (u_i))$ 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 a player can gain in expectation by unilaterally deviating from its mixed strategy in σ to an alternative strategy in $\mathcal{G}$. An NE strategy profile has zero regret for each player. A profile is said to be a 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:[2]

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

An *empirical game* $\mathcal{G}_{S \downarrow X}$ is an approximation of the true game $\mathcal{G}$, in which each player $i \in N$ chooses a strategy s_i from a restricted strategy set $X_i \subseteq S_i$. That is, $\mathcal{G} \downarrow \mathcal{X} = (N, (X_i), (\hat{u}_i))$, where $\hat{u}$ is a projection of u onto the strategy space X. Payoffs are estimated through simulation. We use the notation $\rho^{\mathcal{G}_{S \downarrow X}}$ to make clear when we are referring to regret with respect to an empirical game as opposed to the full game.

[2]Some treatments employ max instead of sum for this; we adopt summed regret in this paper to align with NashConv

A meta-strategy solver (MSS), denoted by $h \in H$, is a function mapping from an empirical game to a strategy profile σ within the empirical game. Examples of MSS (introduced in Related Work) include NE, PRD, uniform, etc. PSRO employing a given MSS may have an established name (e.g., PSRO with NE is DO, with uniform is FP); otherwise we may simply refer to the overall algorithm by the MSS label (e.g., PRD may denote the MSS or PSRO with this MSS).

Evaluating Strategy Exploration

In EGTA, the purpose of evaluating strategy exploration is to understand the relative effectiveness of different exploration methods (e.g., MSSs) across different problem settings. We achieve this purpose through analyzing the intermediate empirical game models they generate during exploration.

Evaluating an Empirical Game Model

From the perspective of strategy exploration, the key feature of an empirical game model is what strategies it incorporates.[3] In the EGTA framework, the restricted strategy set X is typically a small slice of the set of all strategies S, so the question is how well X covers the strategically relevant space. There may be several ways to interpret "strategically relevant", but one natural criterion is whether the empirical game $\mathcal{G}_{S \downarrow X}$ covers solutions or approximate solutions to the true game $\mathcal{G}$.

This criterion is directly captured by the profile in the empirical game with minimal regret relative to the full game: that is, the MRCP as described above. Formally, the MRCP $\bar{\sigma}$ is defined as follows.

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

First notice that the regret of MRCP decreases monotonically as the empirical game model is being extended, since adding strategies can only increase the scope of minimization. Moreover, MRCP tracks convergence in that the regret of MRCP reaches zero exactly when an NE is contained in the empirical game.

Despite these appealing characteristics of MRCP, it is not always practically feasible to find it for the following reasons: (a) regret calculation generally requires a best-response oracle, which could be computationally expensive (e.g., we often find RL the best available method). Moreover, multiples invocations of this oracle are required to find one MRCP; (b) The search for MRCP in an empirical game is a non-differentiable optimization problem and standard optimization techniques that calculate the gradient of the Lagrangian do not apply.

Solver-Based Regret

Given the difficulty of computing MRCP, studies often employ some other method to select a profile from the empirical game to evaluate. Any such method can be viewed as

[3]The accuracy of the estimated payoff functions over these strategies is also relevant, but mainly orthogonal to exploration and outside the scope considered here.

a meta-strategy solver, and so we use the term *solver-based regret* to denote regret in the true game of a strategy profile selected by an MSS from the empirical game. In symbols, the solver-based regret using a particular MSS is given by $\rho^{\mathcal{G}}(MSS(\mathcal{G}_{S\downarrow X}))$. By definition, MRCP is the MSS that minimizes solver-based regret.

A common choice of MSS for solver-based regret is NE. NE-based regret measures the stability in the true game of a profile that is perfectly stable in the empirical game. As such, it can be justified as measuring the quality of using the empirical-game result in the true game, which is what one might naturally do if strategy exploration were halted at this point. If an alternative solution concept is preferred, that could likewise be employed in solver-based regret.

Indeed any MSS is eligible, however not all are well-suited for evaluating strategy exploration. For example, SP simply selects the last strategy added, and is completely oblivious to the rest of the strategy set X. This clearly fails to measure how well X as a whole captures the strategically relevant part of S, which is the main requirement of an evaluation measure as described above.

Solver Consistency for Evaluation

Our framework as described to this point employs MSSs in two distinct ways: to direct a strategy exploration process, and to evaluate intermediate results in strategy exploration. It may seem natural to evaluate exploration that employs MSS M in terms of solver-based regret with M as solver, but we argue this can produce misleading comparisons. If we aim to compare exploration with two alternative MSSs M and M', then we should apply the same solver-based regret measure to evaluate results under these MSSs. In other words, the MSS employed in solver-based regret should be fixed and independent of the MSSs employed for exploration. We term this the *consistency* criterion, and observe that some prior literature violates this condition.

To illustrate the necessity of solver consistency, we offer two examples based on synthetic games.

Example 1. Consider the symmetric zero-sum matrix game in Table 1. Starting from the first strategy of each player, we perform PSRO with uniform and NE as MSSs respectively. The first few iterations of PSRO is demonstrated in Table 2 and Table 3. Due to the symmetry, 2 players' strategy sets and MSS proposed strategies are identical. We plot the uniform-based regret and NE-based regret curves of FP run as well as the NE-based regret curve of DO run in Figure 1a.

If we violate the consistency criterion and compare uniform-based regret of FP with the NE-based regret of DO, we would conclude FP converges faster than DO in the first two iterations. However, FP does not introduce any new strategic interaction than DO because both empirical games, including a^1 and a^3, are identical under two MSSs. Moreover, at the third iteration, FP adds the third action again thus introducing redundancy to the empirical game without changing it substantially. In this scenario, uniform-based regret changes due to the redundancy and may be mistakenly regarded as a change in the empirical game. In this study, we loosen our requirement and permit metrics not to be invariant to redundancy as long as they correctly compare MSSs.

Following the rule of consistency, when the same MSS is applied to evaluation, the equivalence of two identical empirical games is apparent. At iterations 2, 4, and 5, identical empirical games of FP and DO yield the same NE-based regret. Notice that comparing MRCP-based regret between two empirical games also follows the consistency criteria and identifies the equivalence.

	a_2^1	a_2^2	a_2^3
a_1^1	(0, 0)	(-0.1, 0.1)	(-3, 3)
a_1^2	(0.1, -0.1)	(0, 0)	(2, -2)
a_1^3	(3, -3)	(-2, 2)	(0, 0)

Table 1 Symmetric Zero-Sum Game.

#Iter	Strategy Set	DO proposed strategy
1	(a_1^1), (a_2^1)	$(1), (1)$
2	(a_1^1, a_1^3), (a_2^1, a_2^3)	$(0, 1), (0, 1)$
3	(a_1^1, a_1^2, a_1^3), (a_2^1, a_2^2, a_2^3)	$(0, 1, 0), (0, 1, 0)$

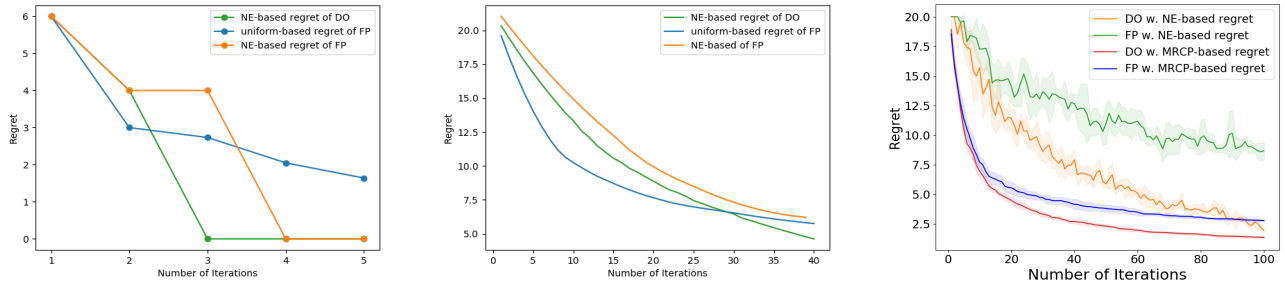
Table 2 **PSRO process for DO.** Items are arranged according to the index of players.

#Iter	Strategy Set	FP proposed strategy
1	(a_1^1),(a_2^1)	$(1), (1)$
2	(a_1^1, a_1^3),(a_2^1, a_2^3)	$(\frac{1}{2}, \frac{1}{2}), (\frac{1}{2}, \frac{1}{2})$
3	(a_1^1, a_1^3),(a_2^1, a_2^3)	$(\frac{1}{3}, \frac{2}{3}), (\frac{1}{3}, \frac{2}{3})$
4	(a_1^1, a_1^2, a_1^3),(a_2^1, a_2^2, a_2^3)	$(\frac{1}{4}, \frac{1}{4}, \frac{1}{2}), (\frac{1}{4}, \frac{1}{4}, \frac{1}{2})$
5	(a_1^1, a_1^2, a_1^3),(a_2^1, a_2^2, a_2^3)	$(\frac{1}{5}, \frac{2}{5}, \frac{2}{5}), (\frac{1}{5}, \frac{2}{5}, \frac{2}{5})$

Table 3 **PSRO process for Fictitious Play.** Items are arranged according to the index of players.

Example 2. We further verify our observations in a synthetic zero-sum game with 100 strategies per player. The result is shown in Figure 1b where curves are averaged over 10 random starts. Similar to the previous example, comparing uniform-based rerget of FP against NE-based rerget of DO will lead us astray into believing FP performs better than DO in the first thirty iterations. However, the conclusion is made breaking the consistency criteria and is in fact invalid.

Before 30 iterations, the uniform-based regret curve indicates the existence of a strategy profile in the empirical game of FP more stable, i.e. with lower regret, than NE in the empirical game of DO. Nevertheless, it is very much likely there exists another profile in the empirical game of DO that has far lower regret in the true game, which we shall demonstrate in the next section. Therefore, we conclude the mixed use of evaluation metrics may result in improper comparison among the performance of MSSs.



(a) Regret curves in synthetic game. (b) Regret curves in matrix game. (c) MRCP-based Regret vs NE-based regret.

Figure 1 Evaluation Results for Strategy Population

In matrix games, it is possible to calculate the MRCP of an empirical game using direct search. Following Jordan, Schvartzman, and Wellman (2010), we apply the amoeba method (Nelder and Mead 1965) to search for MRCP. Figure 1c presents the results of averaged PSRO runs with FP and DO on a two-player zero-sum game with 200 strategies per player. As shown in Figure 1c, the MRCP-based regret by definition is lower than its NE-based regret counterpart. Moreover, the NE-based regret reports the relative performance of theoretical ground truth. Notice that the gap between NE-based regret and MRCP-based regret dwindles as DO and FP gradually converge to true game NE, i.e., all regrets approach zero. We also observe that the MRCP-based regret curves are much smoother than the NE-based regret curves and indicates a steady performance improvement on the population of strategies.

Evaluation Consistency in Poker Games

We further examine the consistency criterion in poker games. Specifically, PSRO is implemented with uniform, PRD and NE in 2-player Kuhn's poker and Leduc poker. We benchmark FP and PRD's performance against NE-based regret curve of DO in our experiments. Our implementation is based on *OpenSpiel* (Lanctot et al. 2019) with Deep Q-Network (Mnih et al. 2013) as the approximate best response oracle. Each curve is averaged over 10 random seeds. Experimental details are included in Appendix.

Fictitious Play For Leduc poker, Figure 2a indicates DO performs better than FP, no matter with NE-based regret or uniform-based regret and there is little difference between the two regret curves. In Kuhn's poker shown in Figure 2b, however, the NE-based regret of FP converges at around iteration 60 while uniform-based regret is far from zero even at a hundred iterations, leaving an impression that FP converges much slower. The evaluation inconsistency artificially enlarges gap between the performance of DO and FP in this example.

Projected Replicator Dynamics Next, we reproduce the PSRO experiments with PRD in Leduc poker in Figure 2c. Observing the evaluation consistency, the NE-regret curves of DO and PRD reveals that the PRD is slightly better than DO in terms of lower regret and variance. If the consistency is violated, however, and PRD-based regret of PRD is compared against NE-based regret of DO, then the conclusion that PRD dramatically outperforms DO will be drawn, resulting in two distinct conclusions.

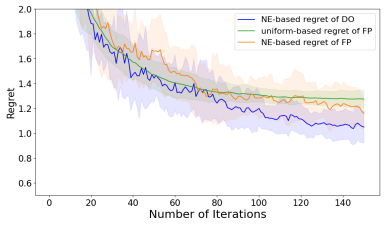
To further understand this phenomenon, we measure the PRD-based regret of DO in Figure 3. Although the strategies are generated by DO, the PRD strategies in the empirical game of DO exhibit similar convergence rate and regret scale as the ones in the empirical game of PRD. The relative position of two PRD-based regret curves supports the claim that the PRD slightly outperforms DO and verifies the importance of evaluation consistency. The second observation is that there could exist more stable strategy profiles than the ones proposed by the MSS in an empirical game, as the PRD strategy profiles being more stable than NE profiles in DO. This again highlights the importance of focusing on evaluating the performance of a strategy population rather than a single profile in EGTA. For Kuhn's poker shown in Figure 2d, the difference between regret curves are minimal. For comparison, we unify all NE-based regret curves with different MSSs in Figure 2e and Figure 2f.

MRCP as a Meta-Strategy Solver

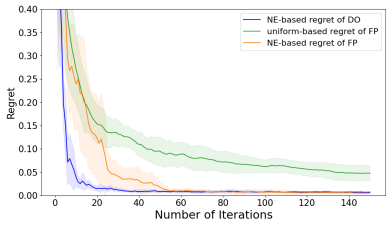
We have observed the existence of strategy profiles with lower regret than NE in the empirical game. One natural question to ask is whether training against these stable profile targets benefits strategy exploration in real-world games, e.g., using MRCP as MSS. We hypothesize the existence of a connection between the learning performance and the regret of target profile, based on the observation that training against the stable profile target leads to better performance in our poker-game experiments.

To test the hypothesis, we compare the performance of MRCP as MSS against DO and FP in the matrix-form two-player Kuhn's poker and a synthetic two-player zero-sum game. In Kuhn's poker, we randomly select 4 starting points and implement PSRO. Figure 4a-4d show that with 3 out of 4 starting points, MRCP converges faster than DO. For the matrix game, Figure 4e and 4f show the benefits of applying MRCP but the performance varies across different starting points.

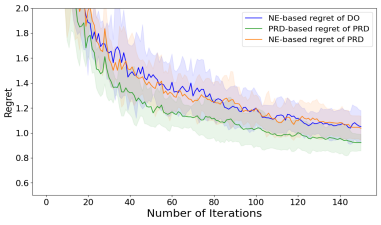
Theoretically, multiple MRCPs could exist in an empirical game and MRCP is not necessarily a pure strategy profile



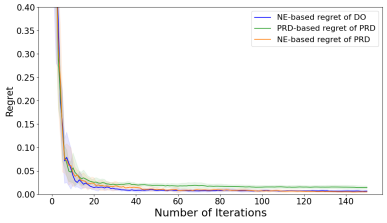
(a) Fictitious Play in 2-player Kuhn



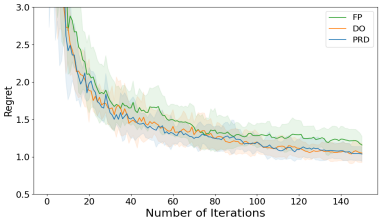
(b) Fictitious Play in 2-player Kuhn



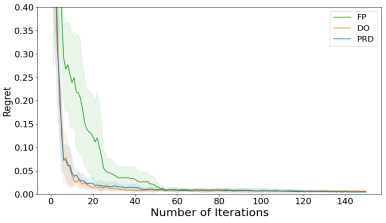
(c) PRD in 2-player Leduc



(d) PRD in 2-player Leduc



(e) Full Leduc



(f) Full Kuhn

Figure 2 Experimental Results in Poker Games

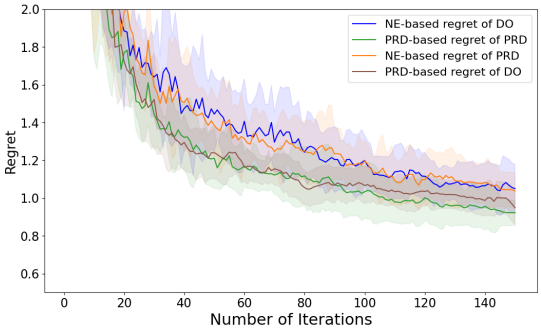


Figure 3 PRD strategies in DO run.

	a_2^1	a_2^2	a_2^3
a_1^1	(0, 0) [2]	(-1, 1) [6]	(-0.5, 0.5)
a_1^2	(1, -1) [6]	(0, 0) [10]	(-5, 5)
a_1^3	(0.5, -0.5)	(5, -5)	(0, 0)

Table 4 **Symmetric Zero-Sum Game for MRCP.** Regret of profiles is shown in the square parenthesis.

in general. Moreover, purely using MRCP as a MSS does not guarantee convergence to NE since the best-responding strategy could already exist in the empirical game. We define this property of MRCP as follows.

Definition. An empirical game with strategy space $X \subset S$ is *closed* with respect to MRCP $\bar{\sigma}$ if

$$\forall i \in N, s_i = \underset{s_i' \in S_i}{\operatorname{argmax}}\, u_i(s_i', \sigma_{-i}) \in X_i.$$

To illustrate this concept, consider the symmetric zero-sum matrix game in Table 4. Starting from the first strategy of each player and implementing PSRO with MRCP, we have the empirical game including a^1 and a^2. Since the (a_1^1, a_2^1) is a MRCP (considered all pure and mixed strategy profiles) and best responding to the profile gives a^2 again, the empirical game is *closed* and never extends to the full game wherein the true NE is (a_1^3, a_2^3).

In our experiments, we deal with this issue by only introducing new strategy with highest deviation payoff outside the empirical game and thus guarantee convergence. An alternative is to switch between DO and MRCP whenever this issue happens and the convergence is guaranteed due to the convergence property of DO.

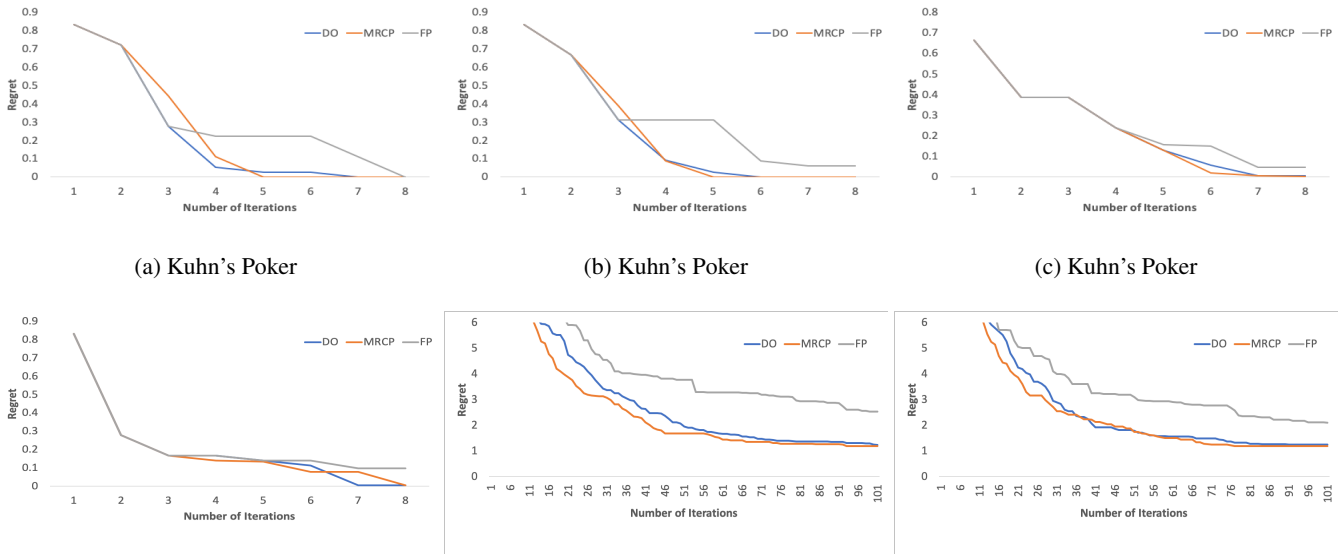
Based on our preliminary results, we believe that the MRCP can be a promising concept for heuristic strategy generation. The performance of MRCP can be further improved by measures like properly handling the "*closed*" issue. Moreover, further study on approximating MRCP in large games remains open to research.

Evaluation without Exact Best Response

Evaluation with Combined Game

The exact best response oracle provides a regret calculation tool for evaluating MSSs. Despite its availability in the poker games the we experiment with, in many complex games, calculating for the exact best response is not feasible due to the large game trees. An alternative is applying the concept of the combined game as a heuristic evaluation approach. In this section, we verify its effectiveness from a game-theoretic perspective.

A combined game is an encompassing empirical game whose strategy set is a union over all strategy sets of the empirical games that are generated across different MSSs or random seeds. The missing payoff entries are further simu-



(a) Kuhn's Poker (b) Kuhn's Poker (c) Kuhn's Poker

(d) Kuhn's Poker (e) Zero-Sum Game (f) Zero-Sum Game

Figure 4 **MRCP as a MSS.** Y axis depicts MRCP-based regret.

lated. The evaluation is conducted by viewing the combined game as the true game and regret calculation only considers deviations within the combined game.

We consider three MSS PRD, NE and uniform in 2-player Leduc poker. For each MSS, we perform 3 differently seeded PSRO runs with 150 strategies, thus each player has 1350 strategies in the combined game. We compare results of evaluation conducted with the combined game and the exact best response. In Figure 5, each curve is an average of three runs with different seeds. The three averaged NE-based regret curves on the top are generated with the exact best response oracle, i.e., evaluating with respect to the true game, while the bottom ones are created by only considering deviations within the combined game. The stratification is caused by the fact that the regret of a profile in the true game is lower bounded by the empirical game.

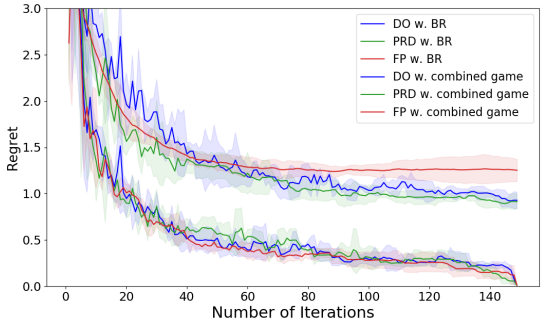


Figure 5 **Regret curves by Combined Games.** Curves of same MSS colored in similar color, with combined game calculated regret in the lower stratum.

We observe that the regret curves based on the combined game does not truthfully reflect the order of performance with exact best response. In Figure 5, despite the apparent higher regret of FP profiles in the true game, FP profiles exhibit low regret in the combined game. The disparity might stem from the strategy generation process of FP: new strategies are trained against a uniform opponent strategy, which makes it harder to be exploited by the counterpart of DO and PRD. We further speculate the advantage that PRD and DO holds over FP in true game only manifest in certain game states. When the exact best response is absent and strategies from the three MSSs compete head to head, the specific game states are never reached. Based on our results, we cast doubt on the accuracy of this evaluation approach.

The combined game essentially attempts to utilize partial strategy space to approximate the relative performance in the global, which does not hold in general. But in certain types of game, for example, in transitive games, local relative performance does indicate the global performance.

Conclusion

The primary contribution of this study is the methodological considerations for evaluating strategy exploration in EGTA. We demonstrate the importance of these considerations in both synthetic games and real-world games and observe that the violation of the evaluation consistency could result in inaccurate conclusions about the performance of different MSSs. We present evidence that the MRCP provides a theoretically robust basis for evaluation. We propose MRCP as a promising concept for strategy exploration based on our observations in poker games and properties of MRCP are investigated. We examine the effectiveness of the combined game for evaluating strategy exploration when calculating the exact best response is computational infeasible. Based on our results, we cast doubt on the accuracy of this evaluation approach.

References

Balduzzi, D.; Garnelo, M.; Bachrach, Y.; Czarnecki, W. M.; Perolat, J.; Jaderberg, M.; and Graepel, T. 2019. Open-ended learning in symmetric zero-sum games. In *36th International Conference on Machine Learning*.

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

Czarnecki, W. M.; Gidel, G.; Tracey, B.; Tuyls, K.; Omidshafiei, S.; Balduzzi, D.; and Jaderberg, M. 2020. Real World Games Look Like Spinning Tops. *arXiv preprint arXiv:2004.09468* .

Jordan, P. R.; Schvartzman, L. J.; and Wellman, M. P. 2010. Strategy Exploration in Empirical Games. In *Ninth International Conference on Autonomous Agents and Multi-Agent Systems*, 1131–1138.

Lanctot, M.; Lockhart, E.; Lespiau, J.-B.; Zambaldi, V.; Upadhyay, S.; Pérolat, J.; Srinivasan, S.; Timbers, F.; Tuyls, K.; Omidshafiei, S.; et al. 2019. OpenSpiel: A framework for reinforcement learning in games. *arXiv preprint arXiv:1908.09453* .

Lanctot, M.; Zambaldi, V.; Gruslys, A.; Lazaridou, A.; Tuyls, K.; Pérolat, J.; Silver, D.; and Graepel, T. 2017. A unified game-theoretic approach to multiagent reinforcement learning. In *31st Annual Conference on Neural Information Processing Systems*, 4190–4203.

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

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; and Riedmiller, M. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* .

Muller, P.; Omidshafiei, S.; Rowland, M.; Tuyls, K.; Perolat, J.; Liu, S.; Hennes, D.; Marris, L.; Lanctot, M.; Hughes, E.; et al. 2020. A generalized training approach for multiagent learning. *ICLR 2020* .

Nelder, J. A.; and Mead, R. 1965. A simplex method for function minimization. *The Computer Journal* 7(4): 308–313.

Omidshafiei, S.; Papadimitriou, C.; Piliouras, G.; Tuyls, K.; Rowland, M.; Lespiau, J.-B.; Czarnecki, W. M.; Lanctot, M.; Perolat, J.; and Munos, R. 2019. α-rank: Multi-agent evaluation by evolution. *Scientific Reports* 9(1): 1–29.

Omidshafiei, S.; Tuyls, K.; Czarnecki, W. M.; Santos, F. C.; Rowland, M.; Connor, J.; Hennes, D.; Muller, P.; Perolat, J.; De Vylder, B.; et al. 2020. Navigating the Landscape of Games. *arXiv preprint arXiv:2005.01642* .

Phelps, S.; Marcinkiewicz, M.; Parsons, S.; and McBurney, P. 2006. 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*, 705–712.

Schvartzman, L. J.; and Wellman, M. P. 2009a. Exploring Large Strategy Spaces in Empirical Game Modeling. In *AAMAS-09 Workshop on Agent-Mediated Electronic Commerce*. Budapest.

Schvartzman, L. J.; and Wellman, M. P. 2009b. Stronger CDA strategies through empirical game-theoretic analysis and reinforcement learning. In *Eighth International Conference on Autonomous Agents and Multi-Agent Systems*, 249–256. Budapest.

Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; Lillicrap, T.; Simonyan, K.; and Hassabis, D. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 362(6419): 1140–1144.

Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. 2017. Mastering the game of Go without human knowledge. *Nature* 550(7676): 354–359.

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

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

Vinyals, O.; Babuschkin, I.; Czarnecki, W. M.; Mathieu, M.; Dudzik, A.; Chung, J.; Choi, D. H.; Powell, R.; Ewalds, T.; Georgiev, P.; Oh, J.; Horgan, D.; Kroiss, M.; Danihelka, I.; Huang, A.; Sifre, L.; Cai, T.; Agapiou, J. P.; Jaderberg, M.; Vezhnevets, A. S.; Leblond, R.; Pohlen, T.; Dalibard, V.; Budden, D.; Sulsky, Y.; Molloy, J.; Paine, T. L.; Gulcehre, C.; Wang, Z.; Pfaff, T.; Wu, Y.; Ring, R.; Yogatama, D.; Wünsch, D.; McKinney, K.; Smith, O.; Schaul, T.; Lillicrap, T.; Kavukcuoglu, K.; Hassabis, D.; Apps, C.; and Silver, D. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575: 350–354.

Wang, Y.; Shi, Z. R.; Yu, L.; Wu, Y.; Singh, R.; Joppa, L.; and Fang, F. 2019. Deep reinforcement learning for green security games with real-time information. In *33rd AAAI Conference on Artificial Intelligence*, 1401–1408.

Wellman, M. P. 2006. Methods for Empirical Game-Theoretic Analysis (Extended Abstract). In *Twenty-First National Conference on Artificial Intelligence*, 1552–1555. Boston.

Wright, M.; Wang, Y.; and Wellman, M. P. 2019. Iterated Deep Reinforcement Learning in Games: History-Aware Training for Improved Stability. In *20th ACM Conference on Economics and Computation*, 617–636.

Appendices

The Combined Game Example

Here we give an example of the idea of combined game. Suppose we run three MSSs (R, B, G) in a 2-player game. For each MSS, we take 3 runs whose strategies are assigned different darkness. So the payoff matrix of the combined game is shown in Figure 6.

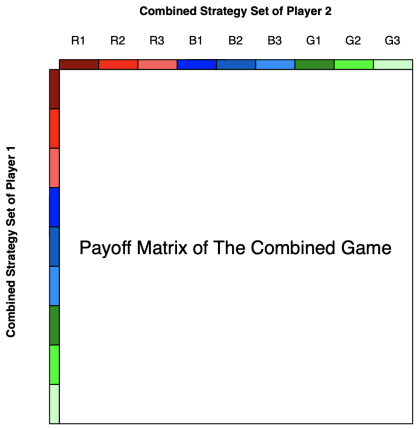


Figure 6 Combined Game Example

Choice of Solution Concepts in Empirical Game

In EGTA framework, the empirical game is used to approximate the true game and NE of the empirical game is picked as an approximate NE to the true game. In some scenarios, e.g., games with large strategy space, limited number of strategies are available to the empirical game due to constraints on computational resources or the difficulty of handcrafting strategies. When the true game may not be well-approximated, we pose a following question: is NE of the empirical game truly the best solution concept to adopt?

Through our experiments, we notice that the NE could have far larger regret in the true game compared with other profiles in the empirical game. Even worse, it is possible that the NE possesses the largest regret among all possible profiles in some extreme cases. Consider matrix game Table 5 with empirical game (a_1^2, a_2^2) for both players. The NE (a_1^2, a_2^2) is the most unstable pure strategy profile in the empirical game. Table 6 list the regret of each pure strategy profile.

	a_2^1	a_2^2	a_2^3
a_1^1	(0, 0)	(-1, 1)	(-2, 2)
a_1^2	(1, -1)	(0, 0)	(-5, 5)
a_1^3	(2, -2)	(5, -5)	(0, 0)

Table 5 Handcrafted Symmetric Zero-Sum Game.

The example inspires us to reconsider the solution concept to apply to real world games: NE strategies of the empirical game could be far from optimum in a global perspective. Is playing NE truly the best option remains a question.

Profiles	**regret**
(a_1^1, a_2^1)	4
(a_1^2, a_2^1)	7
(a_1^1, a_2^2)	7
(a_1^2, a_2^2)	10

Table 6 Regrets of Pure Strategy Profiles.

Experiment Parameters

In the experiments, all synthetic 2-player zero-sum matrix games, regardless of 100 or 200 strategies per player, are generated with uniformly sampled integer payoffs from -10 to 10, both inclusive. We use *Openspiel* default parameter sets for experiments on Leduc and Kuhn's poker: each payoff entry in an empirical game is an average of 1000 repeated simulations; DQN is adopted as a best response oracle, its parameters in Table 7.

Projected Replicator Dynamics is implemented with lower bound for strategy probability 1e-10, maximum number of steps 1e5 and step size 1e-3.

Parameter	**Value**
learning rate	1e-2
Batch Size	32
Replay Buffer Size	1e4
Episodes	1e4
optimizer	adam
layer size	256
number of layer	4
Epsilon Start	1
Epsilon End	0.1
Exploration Decay Duration	3e6
discount factor	0.999
network update step	10
target network update steps	500

Table 7 DQN parameter

Detailed Responses to Prelim Questions

Regret Function

Question In the slides, the regret function is shown as

$$\rho_i^{\mathcal{G}}(\sigma) = \max_{s_i' \in S_i} u_i(s_i', \sigma_{-i}) - u_i(\sigma_i, \sigma_{-i}). \quad (1)$$

The first question is whether the regret can be negative and the second question is why the maximizing operation is not over the probability simplex of mixed strategies.

Answer to Question 1 The key to these two questions is that the utility function is the expected utility function. Let me explain what this means. Let's slightly change the notation and denote $u_i(s_i, s_{-i})$ as the utility if player i plays s_i while others play s_{-i}. Denote $U_i(\sigma_i, \sigma_{-i})$ as the expected utility function for player i. Mathematically,

$$U_i(\sigma_i, \sigma_{-i}) = \sum_{s_i \in S_i} \sum_{s_{-i} \in S_{-i}} \sigma_i(s_i)\sigma_{-i}(s_{-i})u_i(s_i, s_{-i})$$

where $\sigma_i(s_i)$ is the probability of player s_i in σ_i and $\sigma_{-i}(s_{-i})$ is defined similarly.

In equation 1, the u_i is exactly the U_i. Now we rewrite equation 1.

$$\begin{aligned}
\rho_i^{\mathcal{G}}(\sigma) &= \max_{s_i' \in S_i} U_i(s_i', \sigma_{-i}) - U_i(\sigma_i, \sigma_{-i}) \\
&= \max_{s_i' \in S_i} \sum_{s_{-i} \in S_{-i}} \sigma_{-i}(s_{-i})u_i(s_i, s_{-i}) - \sum_{s_i \in S_i} \sum_{s_{-i} \in S_{-i}} \sigma_i(s_i)\sigma_{-i}(s_{-i})u_i(s_i, s_{-i}) \\
&= \max_{s_i' \in S_i} \sum_{s_{-i} \in S_{-i}} \sigma_{-i}(s_{-i})u_i(s_i, s_{-i}) - \sum_{s_i \in S_i} \sigma_i(s_i) \sum_{s_{-i} \in S_{-i}} \sigma_{-i}(s_{-i})u_i(s_i, s_{-i})
\end{aligned}$$

Now we can see that why the regret is non-negative. Notice that $\sum_{s_i \in S_i} \sigma_i(s_i) = 1$. Denote

$$\bar{s} = \operatorname*{argmax}_{s_i' \in S_i} \sum_{s_{-i} \in S_{-i}} \sigma_{-i}(s_{-i})u_i(s_i', s_{-i})$$

$$v(s_i) = \sum_{s_{-i} \in S_{-i}} \sigma_{-i}(s_{-i})u_i(s_i, s_{-i})$$

If $\forall s_i \in S_i$ such that $\sigma_i(s_i) > 0$ and $v(s_i) = v(\bar{s})$, then the regret $\rho_i^{\mathcal{G}}(\sigma) = 0$. Otherwise, there exists $s_i \in S_i$ such that $\sigma_i(s_i) > 0$ and $v(s_i) < v(\bar{s})$ so $\rho_i^{\mathcal{G}}(\sigma) > 0$. To sum up, the regret is always non-negative.

Answer to Question 2 We prove the following two optimization problems are equivalent.

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

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

Since the maximizing operation is only on the first term, we only need to pay attention to

$$\max_{s_i' \in S_i} U_i(s_i', \sigma_{-i}) = \max_{s_i' \in S_i} \sum_{s_{-i} \in S_{-i}} \sigma_{-i}(s_{-i})u_i(s_i', s_{-i})$$

and

$$\max_{\sigma_i' \in \Delta(S_i)} U_i(\sigma_i', \sigma_{-i}) = \max_{\sigma_i' \in \Delta(S_i)} \sum_{s_i \in S_i} \sigma_i'(s_i) \sum_{s_{-i} \in S_{-i}} \sigma_{-i}(s_{-i})u_i(s_i, s_{-i})$$

Using the same logic as in Answer to Question 1. Let

$$\sigma_i^* = \operatorname*{argmax}_{\sigma_i' \in \Delta(S_i)} U_i(\sigma_i', \sigma_{-i})$$

$\forall s_i \in S_i$ such that $\sigma_i^*(s_i) > 0$, s_i maximizes

$$\sum_{s_{-i} \in S_{-i}} \sigma_{-i}(s_{-i})u_i(s_i, s_{-i})$$

(Otherwise, $\sigma_i^*(s_i) = 0$.) All such s_i gives the same maximum (otherwise σ_i^* does not give the highest value of U_i), which is equal to the value of our first optimization problem. So these two optimization problem are equivalent.

The core intuition for both questions here is that we have a linear programming problem and the optimum happens at the extreme point of the simplex.

Evaluation Consistency

Question The consistency criterion is obvious. Why do people still violate it?

Answer First I think it is very interesting to find something that is obvious but no one notices it. People studying game theory used to applying learning algorithms and only caring about the performance of the newest strategy. When moving to the PSRO framework, they take their previous evaluation method for granted ignoring the key feature of PSRO that we have a population of strategies. Therefore, the real cause of the inconsistency issue is the lack of understanding of the PSRO framework although being consistency for evaluation seems obvious in machine learning literature.

(Next Page continues)

Complex Game Learning

Question What makes the learning in complex game like StarCraft II difficult?

Answer

1. First, such game can be cyclic. In certain complex game, there could be millions of strategies in a cycle. So continually exploring such cycle is expensive.
2. The imperfect information of these games make the learning hard. The imperfect information means we don't know what exact actions our opponent take so we don't know what is the true state of the game, similar to partial observability. For example, in Starcraft II, we have the Fog of War and I can only see the enemy once we encounter.
3. The large strategy space and state space make the learning of the optimal strategies very hard since we cannot enumerate all of the combinations. So we need to search or do sampling, which usually leads to a sub-optimal solution. This is why we usually don't care the convergence to NE in these games but whether we exceed the human performance.
4. One play of these games usually takes at least half an hour. So we can see we have a long-time planning issue here. Together with the large strategy space and state space, learning can be very difficult.