Evaluating Strategy Exploration in Empirical Game-Theoretic Analysis

Yongzhao Wang
University of Michigan
Ann Arbor, USA
wangyzh@umich.edu

Qiurui Ma
Harvard University
Boston, USA
qiurui_ma@g.harvard.edu

Michael P. Wellman
University of Michigan
Ann Arbor, USA
wellman@umich.edu

ABSTRACT

In empirical game-theoretic analysis (EGTA), game models are extended iteratively through a process of generating new strategies based on 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 depend sensitively on criteria adopted and measures employed. We investigate some of the methodological considerations in evaluating strategy exploration, proposing and justifying new evaluation methods based on examples and experimental observations. In particular, we emphasize the fact that empirical games create a space of strategies and evaluation should reflect how well it covers the strategically relevant space. Based on this fact, we suggest that the *minimum regret constrained profile* (MRCP) provides a particularly robust basis for evaluating a space of strategies, and propose a local search method for computing MRCP. However, MRCP computation is not always feasible especially in large games. To evaluate strategy exploration in large games, we propose a new evaluation scheme that measures the strategic coverage of an empirical game. Specifically, we highlight consistency considerations for comparing across different approaches. We show that violation of the consistency considerations could yield misleading conclusions on the performance of different approaches. In accord with consistency considerations, we propose a profile-selection method, which effectively discovers the profile that can represent the strategic coverage of an empirical game through its regret information. We show that our evaluation scheme reveals the authentic learning performance of different approaches compared to previous evaluation methods.

KEYWORDS

Multi-agent Learning; Multi-agent Evaluation; Empirical Game-Theoretic Analysis

ACM Reference Format:
Yongzhao Wang, Qiurui Ma, and Michael P. Wellman. 2022. Evaluating Strategy Exploration in Empirical Game-Theoretic Analysis. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 9 pages.

1 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 [16, 17] 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 generate new strategies. For example, DeepMind's milestone achievement in the complex video strategy game StarCraft II entailed an elaborate population-based search approach [21] informed by game-theoretic concepts.

Many recent works [1, 6, 9] 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 other agents to play strategies from previous iterations. A general formulation of this approach is the *Policy Space Response Oracle* (PSRO) algorithm [6]. PSRO can be viewed as a form of *empirical game-theoretic analysis* (EGTA) [20, 24], 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 [4] 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 *et al.* [8] for two-player games and called the *double oracle* (DO) method.

PSRO defines 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 §2, as well as alternative approaches to strategy exploration outside the PSRO framework (e.g., not involving RL or varying from best-response).

In practical terms, the proof of a method is whether it produces a superior solution (e.g., a champion Go program). However, we also seek to understand the relative effectiveness of strategy exploration methods across problem settings, and this remains an

open problem for EGTA methodology. Comparing the performance of alternative methods is subtle because each generates a distinct sequence of strategies, and thus the empirical game model at any point reflects a distinct strategy space. The relevant comparisons are across different strategy spaces, which may not be faithfully represented by a simple summary such as an interim solution. This fact has tended to be neglected in studies proposing and evaluating new ideas on strategy exploration [6, 9], and as we demonstrate below, this can lead to misleading conclusions on the performance of different approaches.

The present study illuminates several methodological considerations for strategy exploration. First, we identify a key distinction between PSRO and other learning dynamics, which is that the empirical game model evolves through extending a space of strategies and hence its evaluation should reflect how well it covers the strategically relevant space.

Second, we seek a principled evaluation metric for empirical games, and suggest the proposal by Jordan *et al.* [4] that the regret of the *minimum-regret constrained-profile* (MRCP) can serve this purpose. We show the effectiveness and advantages of using MRCP as a metric through examples. To find MRCP more accurately, we propose a variant of the amoeba method [10] that outperforms previous approaches in matrix games. MRCP calculation is not always computationally feasible, so we identify some desiderata for alternative evaluation metrics. Per the first point above we highlight the importance of evaluating the whole space of strategies in an empirical game. Further, we propose 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. Based on these considerations, we propose a new evaluation solver selection scheme for EGTA, which leads to a sensible comparison across MSSs. We demonstrate the significance of our considerations and approaches in both synthetic and real-world games.

Finally, we consider the problem of regret-based evaluation in situations where calculating exact best response is infeasible thus accurate regret is not available. One alternative that is widely applied is using generated strategies for evaluation purpose where regret calculation for different MSSs only considers deviations within the generated strategies.

Contributions of this study include:

(1) Recognition that empirical games create a space of strategies and evaluation should reflect how well they cover the strategically relevant space. To serve this purpose, we suggest MRCP as evaluation metric and present evidence that MRCP provides a particularly robust basis for evaluation.
(2) The notion of *solver-based regret* for evaluation, with focus on consistency considerations for comparing MSSs. We demonstrate the potential for misleading results when consistency is violated as in the prior literature.
(3) A new evaluation solver selection scheme which leads to a sensible comparison across MSSs;
(4) A variant of the amoeba method that outperforms previous approaches in matrix games, plus some insight on MRCP approximation in games wherein regret calculation is restricted.

2 RELATED WORK ON STRATEGY EXPLORATION

In the first instance of automated strategy generation in EGTA, Phelps *et al.* [13] employed genetic search over a parametric strategy space, optimizing performance against an equilibrium of the empirical game. Schvartzman & Wellman [15] combined RL with EGTA in an analogous manner. Questioning whether 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 [14]. They identified situations where adding a best response to equilibrium would perform poorly, and proposed some alternative approaches. Jordan *et al.* [4] 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 [6]. 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 [8]. Lanctot *et al.* [6] 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) employs an RD search for equilibrium [18, 19], but truncates the replicator updates 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.* [9] of a recently proposed evolutionary-based concept, α-rank [11], 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 [2], 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.* [23] 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 *et al.* [26] 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 *et al.* [4] 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.* [1] 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 proposed 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.* [12] proposed using spectral analysis of the α-rank best response graph, and Czarnecki *et al.* [3] visualize the strategic topography of real-world games as a spinning top wherein layers are transitive and strategies within a layer are cyclic.

3 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 $\mathcal{G}$ 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:

$$\rho^{\mathcal{G}}(\sigma) = \sum_{i \in N} \rho_i^{\mathcal{G}}(\sigma). \tag{1}$$

Some treatments employ max instead of sum for this; when necessary to disambiguate we refer to (1) as *sum-regret*. Both are relevant measures of distance from equilibrium, and we appeal to the *max-regret* variant in our approach to approximating MRCP in §5.3.

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.[1] 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.

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 §2) 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).

[1] Because payoffs are estimated through simulation, $\hat{u}$ is also subject to sampling error. This presents additional statistical issues [20, 22, 25]; here we ignore those and focus on the issues that arise from strategy set restriction.

4 EVALUATING STRATEGY EXPLORATION

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.

4.1 Evaluating an Empirical Game Model

From the perspective of strategy exploration, the key feature of an empirical game model is what strategies it incorporates.[2] In EGTA, 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 $\mathcal{G}$.

The profile in the empirical game closest to being a solution of the full game is the MRCP, as described above. Formally, $\bar{\sigma}$ is an MRCP of $\mathcal{G}_{S \downarrow X}$ iff:

$$\bar{\sigma} = \underset{\sigma \in \Delta(X)}{\operatorname{argmin}} \sum_{i \in N} \rho_i^{\mathcal{G}}(\sigma) \tag{2}$$

The regret of MRCP thus provides a natural measure of how well X covers the strategically relevant space. In prior literature, MRCP was studied in games with fixed strategy sets rather than a setting where strategy sets are iteratively built. We extend the study of its properties to our strategy exploration setting. We first note 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 of $\mathcal{G}$ is contained in the empirical game, that is, X covers the support of the NE. We claim both properties of MRCP are important and desirable for evaluation purposes.

Unfortunately, direct use of MRCP as a means for evaluating strategy exploration can be computationally challenging. Calculating regret of a profile, the quantity we are minimizing, generally requires a best-response oracle for the full game, which itself can be quite computationally expensive (which is why we often find RL the best available method). And even given an effective way to calculate regret, the search for MRCP is a non-convex optimization problem over the profile space of the empirical game.

4.2 Solver-Based Regret

Given the general 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 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.

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

An MSS that is commonly employed 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. Whereas any MSS is eligible to play the role of solver, 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.

4.3 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. Indeed, much prior work in PSRO exploration has done exactly this [6, 7, 9].[3]

As we demonstrate below, however, evaluating alternative MSSs M and M' for exploration using their respective MSSs as solvers can produce misleading comparisons, caused by neglecting the principle of evaluating the empirical game as a whole. Instead, we argue, one should apply the same solver-based regret measure to evaluate results under M and M'. 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.

To illustrate the necessity of solver consistency, we offer two examples to demonstrate how a violation of our consistency criterion could lead to a misleading conclusion.

Example 1. Consider the symmetric zero-sum matrix game of 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 are presented in Table 2. Due to symmetry, the two players' strategy sets and MSS-proposed strategies are identical.

	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: A symmetric zero-sum game (Example 1).

Fig. 1a presents regret curves for both MSSs using NE-based regret, as well as the uniform-based regret curve for FP. If we violate the consistency criterion and compare uniform-based regret of FP with the NE-based regret of DO (i.e., green versus blue curves in Fig. 1a), we would conclude FP converges faster than DO in the first two iterations. However, FP cannot actually be better at strategy exploration, as the strategies introduced, a^1 and a^3, are identical under two MSSs. Moreover, at the third iteration, FP fails to add any new strategy, and so the improvement shown is not attributable to the exploration process.

Iter#	Strategy Sets	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)

Iter#	Strategy Sets	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	$(a_1^1, a_1^2, a_1^3),(a_2^1, a_2^2, a_2^3)$	$(\frac{1}{5}, \frac{2}{5}, \frac{2}{5}), (\frac{1}{5}, \frac{2}{5}, \frac{2}{5})$

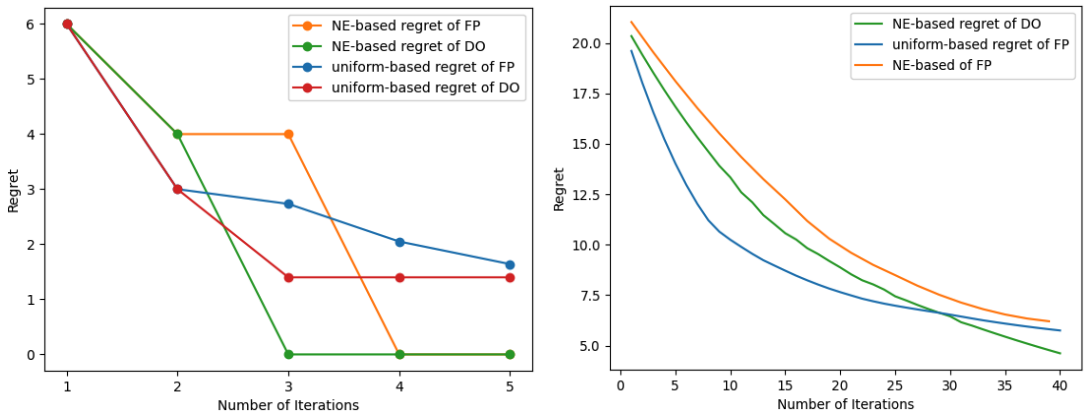
Table 2: PSRO process for DO and Fictitious Play.

Comparing the two MSSs under NE-based regret (i.e., green versus orange regret curves), we see that where FP and DO generate identical empirical games their evaluations coincide. Thus, following the rule of consistency avoids reaching a misleading conclusion about exploration. Note that we would reach the same conclusion if the two MSSs are evaluated under uniform-based regret (i.e., red versus blue curves). However, we observe that not all MSSs are equally effective for evaluation. In this example, although uniform-based regret consistently evaluates equivalent empirical games, its low weight on newly added strategies fails to adequately reflect exploration achievements. For example, the uniform-based regret curve remains well above zero even after the full-game NE has been covered in the empirical game. In Section 4.4, we provide a detailed discussion of this phenomenon and propose a scheme for evaluation solver selection.

Of course, if the goal is just to evaluate DO and FP as online algorithms, then the green versus blue comparison is appropriate. A key virtue of the PSRO framework, however, is that it highlights exploration as a distinct issue and provides the MSS abstraction for addressing it. Within an iterative EGTA approach, the choice of solver to employ for decision making at any stage is completely orthogonal to the method used to extend the game model, and so focusing attention on algorithms that couple these in particular ways (e.g., using the same MSS for solving and exploration) is unnecessarily limiting.

Example 2. We further verify our observations in a synthetic zero-sum game with 100 strategies per player. Resulting regret curves averaged over 10 random starts are shown in Fig. 1b. As for the previous example, comparing uniform-based regret of FP against NE-based regret of DO—breaking our consistency criterion—would lead us astray. First, we see that FP performs best initially, but is ultimately overtaken by DO. More importantly, as we demonstrate in §5.1 below, even the assessment that FP's strategy exploration is more effective than DO's over the first thirty iterations is invalid. Indeed, the blue-versus-green comparison up to iteration 30 shows that the uniform strategy profile in the empirical game of FP is

[3] Although Li and Wellman [7] is not focused on strategy exploration, it does present some plots (Figs. 2 and 3) with multiple curves using different MSSs for evaluating regret. For other works, we verified this by examining the published code and through our own efforts to reproduce the results in these papers. Specifically, we found the code published as part of OpenSpiel [5] evaluates progress in exploration by regret of the MSS employed for exploration. We also reproduced the learning performance of PSRO with different MSSs and inferred that the MSS used for evaluation is the same as the one for strategy exploration, which is often apparent by examination of regret curves. For example, the NE-based regret curve of fictitious play oscillates dramatically while its uniform-based regret curve is much more smooth. So it is easy to identify which MSS was used for evaluation.



(a) 3×3 **game (Example 1).** **(b)** 100×100 **game (Example 2).**

Figure 1: Regret curves evaluating NE and uniform as MSSs strategy exploration, under different solvers.

more stable (has lower regret) than NE in the empirical game of DO. But as in the prior example, this is an artifact of selecting the uniform rather than the NE profile for evaluation. Moreover, as illustrated below in Fig. 3, we should generally expect there to exist non-NE profiles in the empirical game of DO with significantly lower regret in the true game.

This example demonstrates mixed use of evaluation metrics may result in improper comparison among the performance of MSSs. Indeed, we have found that this phenomenon is quite common in prior work, leading in particular to misleading evaluations of FP as a strategy exploration approach. In formulating the general consistency criterion, we emphasize that improper comparisons could be made with any two MSSs; the issue is not limited to FP or any specific MSSs employed in these examples.

4.4 Evaluation Solver Selection

We further examine the consistency criterion in simplified poker games, specifically two-player Kuhn poker and Leduc poker. These poker games have been commonly employed in prior work within the PSRO framework, facilitating comparison of experimental results. Specifically, we evaluate FP, PRD, and NE as MSSs. Moreover, to select an effective solver to implement the consistency criterion, we propose a new evaluation solver selection scheme, designed to reveal the authentic performance of MSSs for strategy exploration.

4.4.1 Solver Consistency with FP. For Leduc poker, Fig. 2a indicates DO performs better than FP under NE-based regret. However, the uniform-based regret is quite misleading as a measure of exploration performance of DO. It actually increases over much of the range, which would seem to suggest that adding strategies makes the game model worse, which intuitively makes little sense.

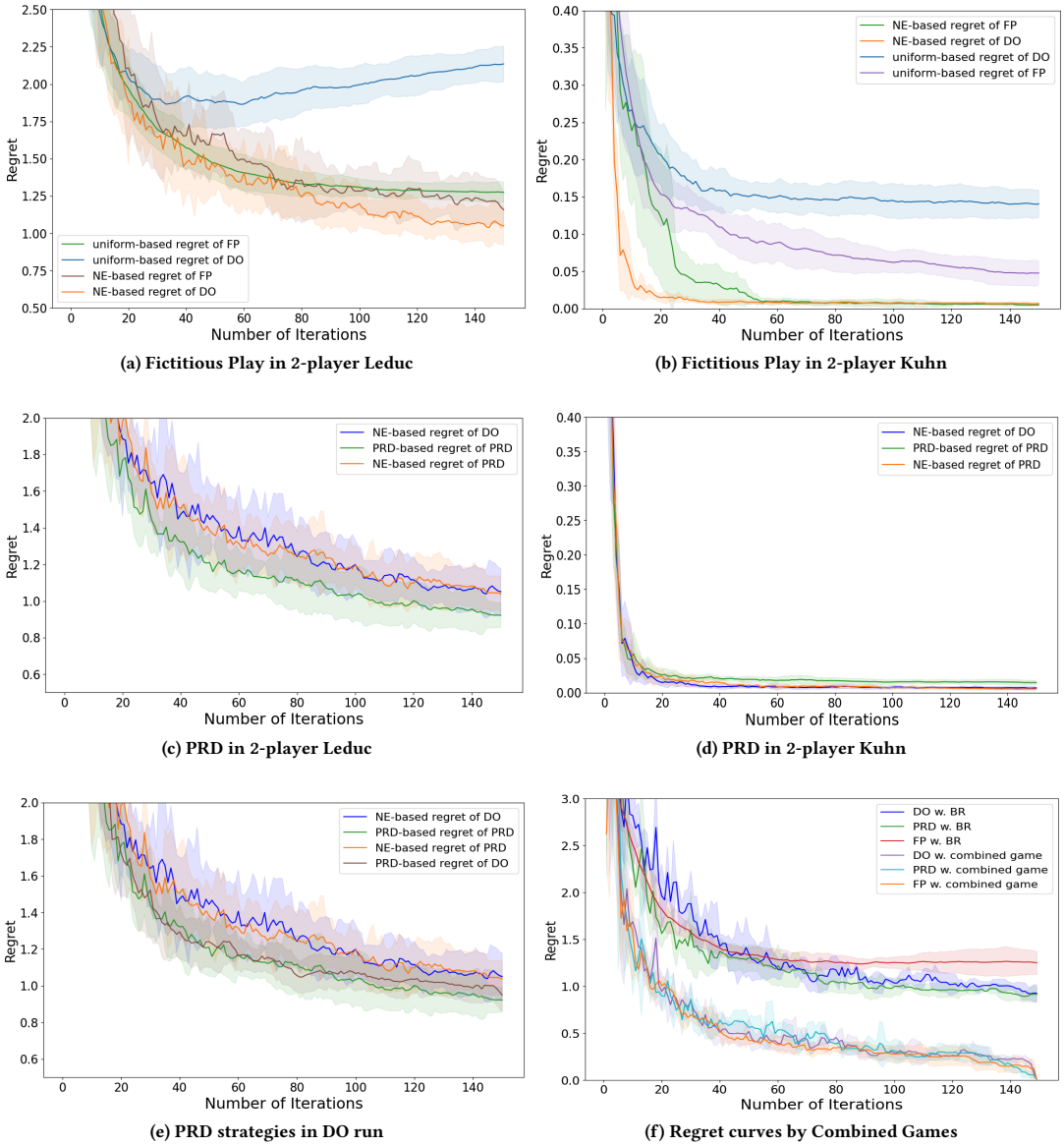
In Kuhn poker (Fig. 2b), DO again outperforms FP under NE-based regret. Uniform-based regret of DO is misleading for Kuhn as it is for Leduc poker.[4] FP shows much faster convergence under NE-based rather than uniform-based regret after twenty iterations or so. Indeed, the uniform-based regret is far from zero even at a hundred iterations. As we saw in the examples above, uniform-based evaluation may misleadingly show smooth improvement where there is none. Here we see again that it can also leave the impression of slow progress even when the empirical game actually contains the key strategies needed for accurate solution.

4.4.2 Solver Consistency with PRD. We show experimental results of PSRO with PRD in Leduc poker in Fig. 2c. We first note that following the rule of consistency, there is little performance gap between PRD and DO (i.e., the blue and orange curves). If we violate consistency and compare PRD-based regret of PRD against NE-based regret of DO (green versus blue curves), however, we would be prone to conclude that PRD clearly and significantly outperforms DO. For Kuhn poker (Fig. 2d) we would conclude there is little difference, but looking closely and ignoring consistency might lead us to conclude that PRD is slightly worse in the limit. In both cases, we see that the choice of evaluation solvers can drive assessments about exploration performance.

The above examples have shown that not all MSSs are equally suited for evaluation, even if used in compliance with the consistency criterion. Consistency is important for achieving meaningful comparisons, but not sufficient. Conclusions about exploration performance are also sensitive to the selection among MSSs as evaluation solvers.

4.4.3 An Evaluation Solver Selection Scheme. Recall that MRCP is the MSS minimizing solver-based regret and thus the regret of the MRCP of an empirical game measures how well the empirical game covers the strategically relevant space. If we could feasibly compute the MRCP or an approximation, that would be a natural choice for solver-based regret. Though this is infeasible in general, we can capture the spirit of MRCP by attempting to minimize solver-based regret. Toward this end, we propose a heuristic evaluation solver

[4]Our conjecture is that the new poker strategies introduced by DO after a point are very good at exploiting vulnerabilities in the current equilibrium, but quite poor as poker players overall. These strategies are quite important to include in the empirical game, to prevent exploitable solutions, even though they should not be part of the solutions themselves. This is a common game-reasoning phenomenon, providing another explanation for why uniform is a poor choice of solver for evaluation.



(a) Fictitious Play in 2-player Leduc

(b) Fictitious Play in 2-player Kuhn

(c) PRD in 2-player Leduc

(d) PRD in 2-player Kuhn

(e) PRD strategies in DO run

(f) Regret curves by Combined Games

Figure 2: Experimental regret curves for poker games.

selection scheme that chooses the solver with lowest-regret curve among running solvers. We demonstrate the significance of our scheme for evaluating different MSSs by checking the previous PRD example.

In the example, if we merely adhere to solver consistency with NE-based regret (i.e., comparing blue versus orange regret curves in Fig. 2c), we would not distinguish the performance difference between PRD and DO. In this case, NE in the empirical game exhibits relatively high regret with respect to the true game. We know it is far from MRCP, as the green curve in this plot demonstrates the existence of lower-regret profiles in the same empirical games. Although we cannot tell exactly where the MRCP lies, the PRD solver in this example clearly provides a better approximation than does the NE solver. Considering PRD as the solver for evaluation and following solver consistency, we can likewise evaluate DO using PRD-based regret. The result is shown in the purple curve of Fig. 2e (other regret curves are as in Fig. 2c). PRD-based regret of DO is indeed lower than NE-based regret of DO (purple versus blue curves), and thus PRD as an evaluation solver successfully identifies the profiles with lower regret in the empirical games across DO iterations. This achieves our purpose of identifying profiles closer to MRCP as the basis for evaluation.

By comparing the PRD-based regret curves of DO and PRD, we observe that they exhibit similar improvement rates through early

iterations, but eventually PRD shows a small consistent advantage. This we regard as the best available evidence from these experiments on the authentic relationship between PRD and DO. Had we ignored solver consistency and compared the green and blue curves, we would have correctly concluded PRD's superiority but grossly overestimated the performance gap.

To state our proposal more explicitly: we argue for selecting the solver that minimizes regret in the given context. Specifically, fix a set of MSSs $\mathcal{M}$, typically the same set of MSSs being evaluated for strategy exploration. Let $\mathcal{R}$ be a set of PSRO runs employed to select the evaluation solver. At each iteration t of each run $r \in \mathcal{R}$, we have an empirical game over strategy set X_t^r. For each X_t^r and solver $M \in \mathcal{M}$, we evaluate regret in the full game of the empirical game solution under M. We then designate as evaluation solver M^* the MSS that performs the best over these runs:

$$M^* = \operatorname*{argmin}_{M \in \mathcal{M}} \sum_{r \in \mathcal{R}} \sum_{t} \rho^{\mathcal{G}}(M(\mathcal{G}_{S \downarrow X_t^r})).$$

Alternatively, we can accommodate the possibility that which solver minimizes true-game regret may vary over the course of the strategy exploration process. We propose a *pointwise* selection scheme, which designates an evaluation solver M_t^* for each iteration t:

$$M_t^* = \operatorname*{argmin}_{M \in \mathcal{M}} \sum_{r \in \mathcal{R}} \rho^{\mathcal{G}}(M(\mathcal{G}_{S \downarrow X_t^r})).$$

Note that the pointwise scheme, like that for selecting a single solver, accords with our consistency criterion. Variations that combine regrets across runs and time in some way other than summation are also admissible.

5 PERFORMANCE OF MRCP AND CALCULATION REFINEMENT

5.1 Evaluation Performance of MRCP

Though computation of MRCP in large games is generally infeasible, for experimental purposes we can evaluate it in a feasible context. Here we present such an evaluation on matrix games of fixed and modest size. Fig. 3 displays averaged regret curves of PSRO runs on the same synthetic matrix game of Example 2, with FP and DO evaluated by MRCP-based regret. We observe that the MRCP-based regret by definition is lower than its NE-based regret counterpart. In this instance, the comparison using MRCP-based regret validates the qualitative comparison using NE-based regret. Notice that the gap between NE-based regret and MRCP-based regret diminishes as DO and FP gradually converge to a 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. MRCP is monotone by definition, the steady performance improvement reflects more accurately the progress in quality of empirical game model achieved by strategy exploration.

5.2 MRCP Calculation Refinement

In matrix games, MRCP can be approximated by solving an optimization problem, for example, using the amoeba method [10]. When applying the amoeba method to the MRCP optimization problem, we have to reconcile the fact that the optimization problem

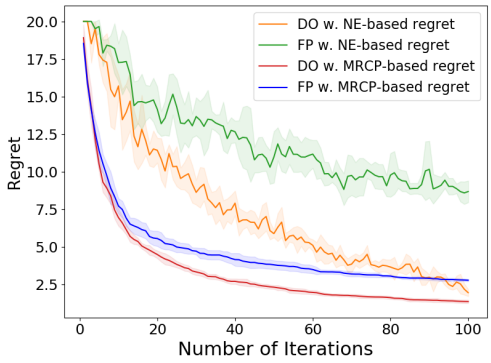


Figure 3: MRCP-based Regret vs NE-based regret.

is constrained while the amoeba method is an unconstrained optimization technique. To handle this issue, Jordan *et al.* [4] propose a binary search (BS) to select the maximum feasible reflection and expansion scaling parameters (step sizes), respectively. However, this approach handles infeasibility by compromising the quality of the reflected and expanded points since the optimal solution points, which are high-dimensional vectors, may not be reached given fixed scaling parameters. We fix this problem instead by projecting an infeasible point onto the unit simplex to maintain feasibility. Our algorithm is specified in detail in Appendix A.

5.3 MRCP Approximation in Large Games

Calculating MRCP in large games can be infeasible since it demands a large number of regret queries each entailing an expensive best-response calculation. We therefore seek an affordable way to approximate MRCP in large games. We start by deriving an upper bound for the regret of a mixed-strategy profile through the deviation payoff of a finite set of pure-strategy profiles. We then approximate MRCP by minimizing the upper regret bound. This approach allows us to focus on pure-strategy deviations which is a more manageable space compared to the search over mixed-strategy profiles.

We derive the upper regret bound as follows:

$$\begin{aligned} \rho_i^{\mathcal{G}}(\sigma) &= \max_{s_i' \in S_i} u_i(s_i', \sigma_{-i}) - u_i(\sigma_i, \sigma_{-i}) \\ &= \max_{s_i' \in S_i} \sum_{s_{-i} \in S_{-i}} \sigma(s_{-i}) u_i(s_i', s_{-i}) - \sum_{s_i \in S_i} \sum_{s_{-i} \in S_{-i}} \sigma(s_i)\sigma(s_{-i}) u_i(s_i, s_{-i}) \\ &\leq \sum_{s_{-i} \in S_{-i}} \sigma(s_{-i}) \max_{s_i' \in S_i} u_i(s_i', s_{-i}) - \sum_{s_i \in S_i} \sum_{s_{-i} \in S_{-i}} \sigma(s_i)\sigma(s_{-i}) u_i(s_i, s_{-i}). \end{aligned} \tag{3}$$

Note that the utility structure of a game may affect the quality of our regret bound. For example, in two-player zero-sum games, since the sum of players' utilities is zero for every profile, the term $\sum_{s_i \in S_i} \sum_{s_{-i} \in S_{-i}} \sigma(s_i)\sigma(s_{-i}) u_i(s_i, s_{-i})$ (i.e., expected utility of playing σ for player i) is canceled when we sum the regret bound over players. As a result, minimizing the summation of upper bounds always produces a pure strategy profile, which could result in a large estimation error.

	Size = 3					Size = 5					Size = 7				
Index	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
$\rho(\bar{\sigma})$	359	262	232	428	305	176	124	487	364	75	95	228	627	103	322
$\rho(\tilde{\sigma})$	505	275	265	532	353	253	144	727	365	106	575	397	794	183	322
$\rho(\sigma^*)$	615	275	242	554	949	535	144	806	737	377	491	514	973	172	507

	Size = 9					Size = 11					Size = 13				
Index	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
$\rho(\bar{\sigma})$	160	121	180	181	17	247	250	243	68	108	324	60	209	103	204
$\rho(\tilde{\sigma})$	249	156	205	230	21	263	405	378	165	108	435	60	318	134	483
$\rho(\sigma^*)$	236	314	759	330	420	388	596	446	152	646	705	216	327	479	204

Table 3: MRCP quality with approximation in symmetric-zero sum games.

We handle this issue by replacing sum-regret (1) with the maximal regret over players. Our approximate MRCP $\tilde{\sigma}$ employs the max-regret variant:

$$\tilde{\sigma}^X = \operatorname*{argmin}_{\sigma \in \Delta(X)} \max_{i \in N} \rho_i^{\mathcal{G}}(\sigma) \tag{4}$$

This modification prevents the expected utility term from being canceled, leading to a more effective result from minimizing the regret bound.

To verify that using max-regret does not unduly distort results, we can evaluate the sum-regret of the profile produced by minimizing either version. Let $\bar{\sigma}^X$ be the profile minimizing sum-regret with respect to strategy set X, and $\tilde{\sigma}^X$ the corresponding MRCP using max-regret (4). Note that for any X, $\rho(\bar{\sigma}^X) \leq \rho(\tilde{\sigma}^X)$. Table 5 in Appendix compares the two MRCP definitions in five instances of Kuhn poker, for each of three sizes of two-player Kuhn poker. As we see, the MRCP calculated using max-regret is quite close to the actual sum-regret MRCP in minimizing sum-regret.

We now measure the quality of our approximation using the upper regret bound (3) with the max-regret version of MRCP (4). Our experiment employs a synthetic two-player zero-sum game with 200 strategies and utilities uniformly sampled from $[-R, R]$, $R = 1000$. Table 3 compares the regrets of exact MRCP $\bar{\sigma}$, *approximated* MRCP $\tilde{\sigma}$ (we overload the notation for convenience), and NE σ^* (i.e., a benchmark). We observe that in some sampled empirical games, the approximation gives profiles with very similar regret as that of the MRCP.

6 EVALUATION WITHOUT EXACT BEST RESPONSE

As noted above, calculating profile regret for purposes of evaluating MSSs generally requires identifying a best-response strategy. However, computing the exact best response may not be feasible in complex games. A particular approach is to collect the strategies generated across a set of PSRO runs, and evaluate regret with respect to that set. We refer to the game with all generated strategies as the *combined game*. In general, regret with respect to the combined game is a lower bound on regret with respect to the true full game. Since the combined game has been used in practice as a heuristic approach to evaluate strategy exploration, it is important to examine its effectiveness.

To test the effectiveness of this approach, we compare results for evaluation with respect to a combined game with that of exact best response (i.e., the ground truth in our context), for some games where calculating exact best responses is feasible. Results are shown in Fig. 2f. We observe that high-regret profiles in the true game may exhibit quite low regret in the combined game. Most concerning is that the slack in the regret bound may vary across MSSs being evaluated, thus producing misleading comparisons. Specifically in Fig. 2f, despite the apparent higher regret of FP profiles in the true game, FP profiles exhibit lower regret in the combined game. Our explanation for the phenomenon is that when one MSS can explore certain strategy to which strategies generated by other MSSs can deviate largely but not vice versa, the combined game fails to identify the correct ordering of MSSs. Details of our combined-game analysis are provided in Appendix B.

7 CONCLUSION

The primary contributions of this study are methodological considerations for evaluating strategy exploration in EGTA, within the PSRO framework. Our observations address nuances that have not been observed before, and may have led to misleading conclusions about the effectiveness of proposed methods. In particular, we propose an evaluation scheme with a consistency condition, dictating that progress in strategy exploration under different MSSs be evaluated with respect to the same solver. This condition, while seemingly obvious, has not always been followed, perhaps because it is natural in online learning settings to evaluate a method at any point based on its own solution criterion. In the context of strategy exploration, in contrast, what is important is not what the latest strategy is, but how it affects the solution of the model it is being added to.

ACKNOWLEDGMENTS

This work was supported in part by the US Army Research Office under MURI W911NF-18-1-0208.

REFERENCES

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

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

[3] Wojciech Marian Czarnecki, Gauthier Gidel, Brendan Tracey, Karl Tuyls, Shayegan Omidshafiei, David Balduzzi, and Max Jaderberg. 2020. Real World Games Look Like Spinning Tops. *34th Annual Conference on Neural Information Processing Systems* (2020).

[4] Patrick R. Jordan, L. Julian Schvartzman, and Michael P. Wellman. 2010. Strategy Exploration in Empirical Games. In *9th International Conference on Autonomous Agents and Multi-Agent Systems* (Toronto). 1131–1138.

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

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

[7] Zun Li and Michael P. Wellman. 2021. Evolution Strategies for Approximate Solution of Bayesian Games. In *35th AAAI Conference on Artificial Intelligence*.

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

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

[10] John A Nelder and Roger Mead. 1965. A simplex method for function minimization. *Comput. J.* 7, 4 (1965), 308–313.

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

[12] Shayegan Omidshafiei, Karl Tuyls, Wojciech M. Czarnecki, Francisco C. Santos, Mark Rowland, Jerome Connor, Daniel Hennes, Paul Muller, Julien Perolat, Bart De Vylder, Audrunas Gruslys, and Rémi Munos. 2020. Navigating the Landscape of Multiplayer Games. *Nature Communications* 11, 5603 (2020).

[13] S. Phelps, M. Marcinkiewicz, S. Parsons, and P. McBurney. 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* (Hakodate). 705–712.

[14] L. Julian Schvartzman and Michael P. Wellman. 2009. Exploring Large Strategy Spaces in Empirical Game Modeling. In *AAMAS-09 Workshop on Agent-Mediated Electronic Commerce*. Budapest.

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

[16] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 362, 6419 (2018), 1140–1144.

[17] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of Go without human knowledge. *Nature* 550, 7676 (2017), 354–359.

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

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

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

[21] Oriol Vinyals, Igor Babuschkin, and Wojciech M. Czarnecki *et al.* 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575 (2019), 350–354.

[22] Yevgeniy Vorobeychik. 2010. Probabilistic Analysis of Simulation-Based Games. *ACM Transactions on Modeling and Computer Simulation* 20, 3 (2010), 16:1–25.

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

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

[25] Bryce Wiedenbeck, Ben-Alexander Cassell, and Michael P. Wellman. 2014. Bootstrap techniques for empirical games. In *13th International Conference on Autonomous Agents and Multi-Agent Systems* (Paris). 597–604.

[26] Mason Wright, Yongzhao Wang, and Michael P. Wellman. 2019. Iterated Deep Reinforcement Learning in Games: History-Aware Training for Improved Stability. In *20th ACM Conference on Economics and Computation* (Phoenix). 617–636.