Policy Space Response Oracles
A Survey

Bighashdel, A.; Wang, Yongzhao ; McAleer, Stephen ; Savani, Rahul; Oliehoek, F.A.

DOI
10.24963/ijcai.2024/880

Publication date
2024

Document Version
Final published version

Published in
Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence

Citation (APA)
Bighashdel, A., Wang, Y., McAleer, S., Savani, R., & Oliehoek, F. A. (2024). Policy Space Response Oracles: A Survey . In K. Larson (Ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence* (pp. 7951-7961). International Joint Conferences on Artifical Intelligence (IJCAI). https://doi.org/10.24963/ijcai.2024/880

Policy Space Response Oracles: A Survey

Ariyan Bighashdel[12] , **Yongzhao Wang**[3] , **Stephen McAleer**[4] , **Rahul Savani**[35]
and **Frans A. Oliehoek**[1]
[1]Delft University of Technology, NL
[2]Eindhoven University of Technology, NL
[3]The Alan Turing Institute, UK
[4]Carnegie Mellon University, USA
[5]University of Liverpool, UK
{a.bighashdel,f.a.oliehoek}@tudelft.nl, yongzhao.wang@turing.ac.uk, smcaleer@cs.cmu.edu,
rahul.savani@liverpool.ac.uk

Abstract

Game theory provides a mathematical way to study the interaction between multiple decision makers. However, classical game-theoretic analysis is limited in scalability due to the large number of strategies, precluding direct application to more complex scenarios. This survey provides a comprehensive overview of a framework for large games, known as Policy Space Response Oracles (PSRO), which holds promise to improve scalability by focusing attention on sufficient subsets of strategies. We first motivate PSRO and provide historical context. We then focus on the strategy exploration problem for PSRO: the challenge of assembling effective subsets of strategies that still represent the original game well with minimum computational cost. We survey current research directions for enhancing the efficiency of PSRO, and explore the applications of PSRO across various domains. We conclude by discussing open questions and future research.

1 Introduction

In recent decades, the exploration of multiagent systems has been a central focus in Artificial Intelligence (AI) research. A multiagent system, often referred to as a ***game***, comprises multiple decision-making agents that interact within a shared environment. To understand strategic behavior among these agents – where the optimal behavior of one agent depends on the behavior of others – game theory provides a mathematical framework that defines behavioral stability through solution concepts like the Nash equilibrium (NE). To identify such solutions, various ***equilibrium computation*** approaches have been developed [von Stengel, 2002; Savani and Turocy, 2024], either to enumerate all equilibria (see the work by Avis et al. [2010] for bimatrix games), or to find a single ***sample equilibrium*** (e.g., with the Lemke-Howson algorithm for a bimatrix game, or Linear Programming for a zero-sum matrix game [von Stengel, 2002]).

As the size of the game (i.e., the number of players and strategies) grows, the computational feasibility of enumeration diminishes, and one tends to focus on finding a sample equilibrium. In the special case of two-player zero-sum games (i.e., settings where two players strictly compete), a sample equilibrium already provides valuable insights: it reveals the game's unique *value*, which is the payoff that the first player can guarantee to obtain by playing a sufficiently strong (equilibrium) strategy *irrespective* of the strategy of the other player. However, even in zero-sum settings, many games that arise from practical applications are too large, and computing a sample equilibrium (even with polynomial-time methods for the resulting linear program) is infeasible. It is these huge games that are our primary focus in this survey.

As an alternative to traditional equilibrium computation methods, to reason about such huge games, a wide range of ***learning*** methods have been applied. Applying learning methods to games is known as ***multiagent learning*** [Shoham *et al.*, 2007], with one of the most prominent approaches being multiagent Reinforcement Learning (RL) [Albrecht *et al.*, 2024]. Compared to traditional methods, learning methods reduce the need to represent the entire game and create intelligent agents by exploring the game interactively. While learning approaches have significantly contributed to the development of intelligent agents, they face many inherent challenges in games. For example, independent learning across agents can render the environment ***non-stationary***, which is a challenge for convergence as each individual learner faces a potentially moving target [Tuyls and Weiss, 2012]. Another challenge is ***non-transitivity*** of a game, where there is not a clear notion of "better" strategy for an agent, and thus effective learning requires the learning scheme to maintain a population of strategies for each agent. Such non-transitivity exists ubiquitously in various games [Czarnecki *et al.*, 2020; Sanjaya *et al.*, 2022; Li *et al.*, 2023b].

Against this backdrop, the ***Policy Space Response Oracles*** (PSRO) framework [Lanctot *et al.*, 2017] emerged as a natural combination of traditional game-theoretic equilibrium computation with learning. In PSRO, a key concept is a ***restricted game*** with estimated payoffs[1], which acts as an

[1]To be precise, normally a restricted game constrains the strategy space (relative to the full game), but the payoffs are "exact", rather than being estimates. By contrast, in practice, PSRO imple-

approximation of the underlying full game. Restricted games are induced from simulations run over combinations of a particular set of strategies. This set of strategies is typically much smaller than the full game, and thus restricted games are feasible to analyze with traditional equilibrium computation methods. PSRO iteratively expands the restricted game by introducing new strategies generated via learning, based on the analysis of the current restricted game.

As a general solver for large-scale games, PSRO has been successfully applied to a wide range of game types and diverse application domains, from mechanism design for sequential auctions [Zhang *et al.*, 2023] to robust RL [Liang *et al.*, 2023]. Numerous PSRO variants have been developed, each tailored to leverage the specific characteristics of different underlying games. As a notable example of its success, algorithms inspired by PSRO have reached state-of-the-art performance in large-scale games such as Barrage Stratego [McAleer *et al.*, 2020], and in StarCraft [Vinyals *et al.*, 2019], where they have convincingly outperformed human experts and prior AI systems.

While PSRO and its variants have been covered to some extent in existing multiagent learning surveys (e.g., [Yang and Wang, 2020; Long *et al.*, 2023; Albrecht *et al.*, 2024]), a dedicated survey on PSRO like this one has been lacking. In this survey, we reflect on the historical development of PSRO, which arose from different research communities, and we place PSRO within the space of game solving approaches. We then present the latest developments on PSRO, highlighting both current and future research directions.

2 The PSRO Framework

The PSRO framework incorporates a synthesis of ideas originating from distinct research communities. Within the ***planning*** community, McMahan et al. [2003] laid the groundwork by formulating robust planning in Markov Decision Processes as a zero-sum game characterized by a vast strategy space. Drawing inspiration from Bender's decomposition in optimization [Benders, 1962], they introduced the ***Double Oracle*** (DO) algorithm. DO is an iterative algorithm for solving games with a finite number of strategies. DO maintains a restricted version of the full game and iteratively expands the restricted game by adding best responses to the current equilibrium. When DO terminates, no player can deviate unilaterally to gain extra payoff and therefore the equilibrium in the current restricted game is an NE of the full game. In finite games, DO is guaranteed to converge to an NE, though the restricted game could include all strategies of the full game in the worst case. Moreover, under proper tie-breaking assumptions, Zhang and Shandholm [2024] showed that DO takes exponentially many iterations to converge in partially-observable stochastic games and extensive-form games.

Concurrently, similar methods were being developed from a different perspective in the ***co-evolution*** research community. Co-evolutionary methods evolve multiple populations of (different) species in parallel. In this setting, there is no explicit fitness function, but the fitness of a population member depends on how well it interacts with members of other populations. Traditional challenges for co-evolution include the presence of intransitive cycles in traits, and the related problem of forgetting a trait that seems not useful at one point in the evolutionary process but later becomes useful again. The community explored ***memory mechanisms*** and game formulations where the co-evolutionary process was set up to discover NE as mixtures of traits [Angeline *et al.*, 1993; Popovici *et al.*, 2012]. To overcome the mentioned obstacles of intransitive cycles and forgetting traits, various archival structures were devised to facilitate monotonic convergence towards a range of game-theoretic solution concepts. For example, the "Nash memory" for symmetric games [Ficici and Pollack, 2003] identifies equilibrium strategies within a discovered restricted game. This was extended to asymmetric games via the Parallel Nash Memory (PNM) [Oliehoek *et al.*, 2006], broadening the standard DO framework to accommodate alternative oracle types beyond the best response.

The PSRO framework [Lanctot *et al.*, 2017] was introduced in a paper titled "A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning", bringing many of the mentioned ideas from the planning, multiagent RL, and co-evolution communities together in a unified framework, which also covers classical game-theoretic learning dynamics such as Fictitious Play [Brown, 1951]. Technically, PSRO generalizes DO in three ways. Firstly, PSRO introduces the concept of ***meta-strategy solver*** (MSS), which extracts a profile from the current restricted game as the next best-response target[2]. This enables the best-response target to extend beyond NE, transforming PSRO into a versatile framework capable of generalizing various classical and modern game-theoretic algorithms. Secondly, PSRO generalizes DO by allowing any form of (approximate) response oracles, including search, planning, and evolutionary algorithms etc (as used in PNM and other works [Oliehoek *et al.*, 2006; Li and Wellman, 2021]). This generalization enables best-response computation in environments with a large number of states and actions. Thirdly, compared to DO, the payoffs of profiles in PSRO are estimated through simulation. See Figure 1 for a depiction of the whole framework, contrasted with DO.

We note that the PSRO framework can be thought of as an instance of ***Empirical Game-Theoretic Analysis*** (EGTA) [Wellman, 2006], which includes a broad set of methods that build and analyze restricted games based on simulation. A comprehensive introduction to EGTA can be found in the survey by Wellman et al. [2024]. As a concrete example of the connection between EGTA and PSRO, in an early EGTA work, Schvartzman and Wellman [2009] deployed tabular RL as a best-response oracle (at a time when

mentations would typically estimate these payoffs through simulation (with stochasticity coming potentially from the environment). A restricted game with estimated payoffs would normally be termed an ***empirical game*** [Wellman, 2006] or ***meta-game*** [Lanctot *et al.*, 2017]. For simplicity, in this survey, we use the term "restricted game" throughout, regardless of whether payoffs are estimated or not.

[2] For simplicity, we often use the same term for the solution concept and the MSS that computes it. For example, we may say "Nash equilibrium" to mean the MSS that computes an NE.

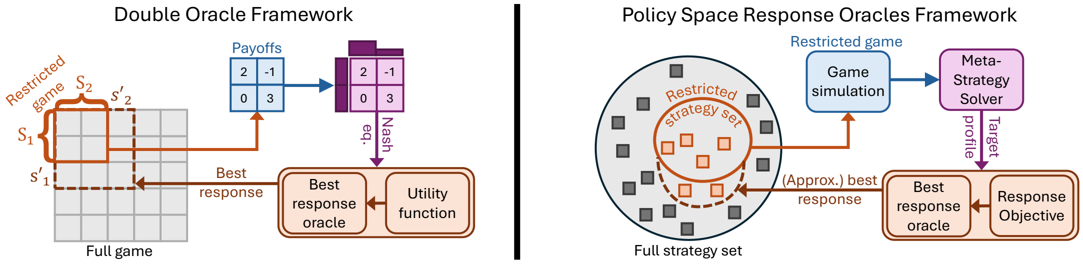

Figure 1: Illustration of the DO and PSRO frameworks respectively. The PSRO framework generalizes the DO framework by introducing MSSs, enabling best-response targets other than NE. Besides, PSRO accommodates various ROs and (approximate) best-response oracles.

deep RL did not exist yet) and NE of the restricted game as a best-response target for strategy generation (i.e. as the MSS).

2.1 The Framework

A normal-form (aka strategic-form) representation of the full game $\mathcal{G} = (N, (S_i), (u_i))$ is a tuple, where N is a finite set of players, each with a non-empty set of strategies S_i and a utility function $u_i : \Pi_{j \in N} S_j \to \mathbb{R}$. A restricted game $\hat{\mathcal{G}}_{S \downarrow X} = (N, (X_i), (\hat{u}_i))$ is a projection of the full game $\mathcal{G}$, with players choosing from restricted strategy sets $X_i \subseteq S_i$, allowing for utilities to be estimated via simulation.

Figure 1 shows the special case of DO applied to a bimatrix game on the left, and the general PSRO framework on the right. In PSRO, each player is initialized with a set of strategies X_i and the utilities for profiles in the profile space X are simulated, resulting in an initial restricted game $\hat{\mathcal{G}}_{S \downarrow X}$. At each iteration of PSRO, an MSS designates a profile $\sigma \in \Delta X$ from the current restricted game $\hat{\mathcal{G}}_{S \downarrow X}$ as the next best-response target, where Δ represents the probability simplex over a set. Then each player $i \in N$ independently computes (learns) a best response $s'_i \in S_i$ against its ***response objective*** (RO), which is a function of strategy profiles, denoted as $RO_i(\sigma)$. In standard PSRO, the RO can be written as $RO_i(\sigma) = u_i(s'_i, \sigma_{-i})$ and maximizing it over s'_i gives player i a best response against other players' strategies σ_{-i}. During this procedure, the other players' strategies σ_{-i} are fixed, which renders the environment stationary for the learning player to compute their response. Then the best response s'_i will be added to its strategy set X_i in the restricted game. This procedure repeats until a stopping criterion has been satisfied (e.g., a fixed number of iterations have been completed or the estimated regret of the restricted-game NE is below a threshold).

2.2 Strategy Exploration in PSRO

In essence, game-theoretic analysis in PSRO is performed by reasoning about restricted games. A restricted game is expected to contain an effective subset of strategies for representation tractability yet still represent the full game well strategically [Balduzzi *et al.*, 2018]. This challenge of restricted game construction with minimum computational cost (i.e., with the fewest strategies required) is described as the ***strategy exploration*** problem [Jordan *et al.*, 2010], which is the main research focus for developing PSRO methods. In PSRO, strategy exploration can be controlled by setting MSSs and ROs, which have a coupled impact; we refer to the joint choice as an MSS-RO combination.

The performance of strategy exploration given a specific MSS-RO combination is normally monitored through the concept of *regret*. The regret $\rho_i(\sigma)$ for a player i in a strategy profile σ is the difference between the player's payoff under σ and the payoff they could have achieved by employing their best-response strategy. Formally, it is defined as $\rho_i(\sigma) = \max_{s'_i \in S_i} u_i(s'_i, \sigma_{-i}) - u_i(\sigma_i, \sigma_{-i})$. This measure reflects the maximal expected gain of player i from unilaterally deviating from their current mixed strategy in σ to an alternative strategy in S_i. In an NE, each player's strategy is a best response to the strategies of the others, which implies that no player can gain by unilaterally changing their strategy. Consequently, a profile is an NE if and only if its regret is zero for all players. Moreover, the stability of a profile σ depends on the aggregation of regrets over players. There are basically two natural ways to aggregate regret over players: the max over regrets and the sum of regrets. Specifically, the sum of regrets of a strategy profile σ over players, denoted as $\rho(\sigma) = \sum_{i \in N} \rho_i(\sigma)$, is known as $NashConv(\sigma)$ [Lanctot *et al.*, 2017]. $NashConv(\sigma)$ measures how far the strategy profile is from NE. In the context of two-player zero-sum games, $NashConv(\sigma)$ is often referred to as *exploitability*, indicating the extent to which the strategy profile can be exploited by an adversary. The second way to aggregate is taking the max of regrets over players. The max of regrets is more standard than NashConv in game theory, as it directly corresponds to the definition of ϵ-NE (i.e., a profile within which no player can gain more than ϵ by unilateral deviation).

In practice, the computation of regrets requires an exact best-response oracle, which is achievable in small games through methods such as strategy enumeration or dynamic programming. However, in large games, computing an exact best response becomes impractical. In such cases, approximate best responses are employed, providing a lower bound on regrets. The accuracy of regret estimation improves with a higher-quality oracle. With limited computational resources, when we cannot find a "better" response for any player, Oliehoek et al. [2019] named the resulting profile a ***resource-bounded NE***, with the interpretation: "with these resources, we did not refute that this is an NE".

2.3 RL View of PSRO: Population-Based Training

In PSRO, a restricted game maintains a population of strategies (also known as policies, i.e., RL agents) and expands this population by introducing new strategies that respond to a specific distribution (mixture) of strategies within the existing population. Unlike self-play, which is a standard alternative approach in which a strategy is trained directly against itself, training a new strategy against a diverse set of strategies in the population, as in PSRO, can potentially enhance the robustness of the resulting strategy. Consequently, PSRO can be classified as a variant of ***population-based training***. The textbook by Albrecht et al. [2024] provides an excellent explanation of PSRO from this viewpoint.

2.4 Organization of the Survey

In this survey, we discuss PSRO from the perspective of game-theoretic analysis, that is, how to conduct effective strategy exploration given a specific goal. Although the performance of strategy exploration depends on the interplay between the chosen MSSs and ROs, existing literature predominantly focuses on setting either MSSs or ROs independently. Therefore, we organize our discussion on research directions and corresponding PSRO variants by first discussing setting MSSs and ROs independently (Sections 3 and 4 respectively). We then, in Section 5, discuss works that have investigated the joint choice of MSS-RO combination, before moving on to discuss in Section 6 how best to evaluate the effectiveness of such choices for strategy exploration. We then discuss research on improving the efficiency of PSRO (Section 7), explore applications of PSRO (Section 8), implementations of PSRO (Section 9), and conclude with open questions and future research directions (Section 10).

3 Strategy Exploration via MSS

In prior works, setting meta-strategy solvers was a primary way to control strategy exploration. In this section, we discuss these works, their motivations, and their efficacy for strategy exploration.

3.1 PSRO with Normal-Form Restricted Games

In the standard PSRO framework, a restricted game is represented in normal form. While the simulations typically unfold through sequential observations and decisions over time, the restricted game abstracts away this temporal structure.

Using Nash and its Variants as MSSs

In PSRO, the most common MSS target is NE, which can be computed by various game-theoretic methods based on the normal-form restricted game. PSRO with NE is essentially DO with deep RL for computing (approximate) best responses. Therefore, PSRO with NE inherits the convergence property of DO, given mild assumptions about the quality of the best responses: In finite games, as long as beneficial deviations can always be found with non-zero probability, PSRO with NE as MSS target will converge to an NE given enough iterations.

The Overfitting Problem. Despite its theoretical convergence guarantee, achieving exact convergence in large games is often unattainable due to constraints such as limited computational resources. Consequently, many prior works studying strategy exploration in large games revolve around the development of new algorithms that exhibit strong empirical performance (e.g., rapid convergence in terms of regret within a small number of PSRO iterations). These works found that a key problem that can prevent good empirical performance is *overfitting*. For strategy exploration, overfitting can arise in two distinct two ways. First, strategy exploration may overfit to the NE of the restricted game. Due to the limited information in the restricted game, the NE may not be an effective best-response target from a global view (i.e., using it, we may fail to generate non-trivial strategies for full-game play) while other best-response targets in the restricted game could be more effective. The second form of overfitting relates to only capturing a specific equilibrium in general games (e.g., games with more than two players or general-sum games) without sufficiently exploring the whole strategy space. Note that the overfitting can be a problem for any solution concept; we discussed it in the context of NE specifically due to NE's prominence as the MSS target in the literature.

Regularization to Prevent Overfitting. To address the overfitting problem, Lanctot et al. [2017] proposed an MSS, called Projected Replicator Dynamics (PRD), an adaptation of traditional replicator dynamics [Taylor and Jonker, 1978]. PRD ensures a probability lower bound for selecting each strategy in the restricted game, allowing the new best response to train against not only strategies in the equilibrium support but also those outside the support. PRD can be viewed as a form of ***regularization*** to prevent overfitting the response to a single exact NE of the restricted game. Due to the diverse training targets, PRD also improves the stability of the new strategy.

Building on the concept of regularization, subsequent research has focused on designing MSSs that prevent overfitting by effectively regularizing the best-response target. For instance, Wang et al. [2019] proposed an MSS that combines NE with a uniform distribution as the best-response target, enabling the best response to an NE strategy mixed with exploration elements. Wright et al. [2019] developed a history-aware approach with a best-response target mixed with previous targets. Online double oracle [Dinh *et al.*, 2022] integrated PSRO with online learning and used an online profile as the best-response target, which can be viewed as a form of regularization. Wang and Wellman [2023b] and Li et al. [2023c] employed quantal response equilibrium [McKelvey and Palfrey, 1995; Gemp *et al.*, 2022] as an MSS, regularizing with bounded rationality. Wang and Wellman [2023b] adopted an explicit view of regularization and introduced Regularized Replicator Dynamics (RRD), an MSS variant that truncates the NE search process in intermediate restricted games based on a regret criterion. Specifically, RRD computes the best-response target by running RD in the restricted game, stopping once the regret of the current profile with respect to the restricted game meets a specified regret threshold. The regret criterion enables RRD to support direct control of the degree of regularization and can be adjusted to fit a specific game.

MSSs Beyond Nash

Rectified Nash. Balduzzi et al. [2019] reformulated the strategy exploration problem as that of enlarging what they called the *gamescape*, which describes the payoff space covered by the restricted game. For symmetric two-player zero-sum games, they proposed ***rectified Nash*** as an MSS designed to expand the gamescape and enhance a diversity measure called effective diversity. In rectified Nash, best responses are only applied to opponent's equilibrium strategies that the learning player defeats or ties with.

Minimum-Regret Constrained Profile. One notable observation for PSRO with NE is that the full-game regret of the restricted-game NE, used as a measure for evaluating the performance of PSRO, does not decrease monotonically over PSRO iterations. In the worst case, the full-game regrets will increase until the last iteration, when a full-game NE is found (one example can be found in the work of McAleer et al. [2022b]). To address this issue, it was proposed to use ***minimum-regret constrained profiles*** (MRCP) [Jordan *et al.*, 2010; Wang *et al.*, 2022] as an MSS. An MRCP is the profile with minimum regret with respect to the full game[3]. With MRCP as the MSS, the resulting PSRO variant is known as *anytime* PSRO because regret monotonically decreases as the restricted game grows. Despite the difficulty of computing MRCP in general games, anytime PSRO leverages the properties of two-player zero-sum games and computes MRCP by regret minimization against a best response (RM-BR) [Johanson *et al.*, 2012]. In a further work [McAleer *et al.*, 2022a], anytime PSRO was extended by including not one but two strategies in the restricted game at each iteration, the first a best response to MRCP and the other a best response to the other player's latest strategy (i.e., the strategy added at the last PSRO iteration). This modification was observed to improve the performance of anytime PSRO. However, Wang et al. [2022] pointed out that MRCP regret will monotonically decrease for *any* MSS since the concept of MRCP is well-defined in any restricted game, which suggests that using MRCP as an MSS in anytime PSRO is not justified purely by the desire to monotonically decrease MRCP regret. We discuss the evaluation of strategy exploration further in Section 6.

Correlated Equilibrium. Apart from the issues discussed above, Marris et al. [2021] further argued that NE may not be an appropriate MSS for PSRO or even a solution concept in general-sum games due to its computational intractability. Therefore, they proposed utilizing correlated equilibrium (CE) and coarse correlated equilibrium (CCE) as MSSs, introducing a variant called Joint PSRO (JPSRO). Since multiple (C)CE exist in a restricted game, they select the unique (C)CE that maximizes the Gini impurity. Theoretical analysis demonstrated that JPSRO converges to a (C)CE. Zhao et al. [2023] combined a diversity measure with (C)CE and proposed a new MSS, called diverse (coarse) correlated equilibrium (DCCE). They showed the improved performance of PSRO with DCCE over JPSRO and many other PSRO variants. Relatedly, Team-PSRO finds a *team-maxmin equilibrium with coordination device* [Celli and Gatti, 2018], an NE defined on the team level, in two-team zero-sum games [McAleer *et al.*, 2023].

Game-Motivated MSSs. In addition to the above well-established solution concepts, the literature on PSRO has also investigated other solution concepts, which originate from specific games but can be generally applied to other game settings. One example is the Risk-Averse Equilibrium (RAE) introduced by Slumbers et al. [2023], aimed at managing risk in multiagent systems. Specifically, RAE minimizes potential variance in rewards by accounting for the strategies of other players. Another exampple is the work of Li et al. [2023c], which proposed the Nash Bargaining Solution (NBS), a concept originating from the bargaining games, as an MSS. NBS can be computed by maximizing the product of players' utilities in the restricted game.

Automated MSS Design. Distinct from prior works that design MSSs based on various solution concepts and heuristics, Feng et al. [2021] proposed Neural Auto-Curricula (NAC) based on meta-learning, which automates the design of MSSs in an end-to-end manner. Specifically, MSSs in NAC are parametrized by a neural network, which is trained by minimizing the regret of the meta-strategy in the resulting restricted games. These restricted games are generated through PSRO with the current MSS (i.e., the current neural network) in games sampled from a game distribution. With this training scheme, Feng et al. [2021] showed that NAC can learn an effective MSS for a family of games. Automated MSS design with Auto-Curricula was also discussed by Yang et al. [2021], who highlighted the significance of including behavioral diversity in auto-curricula and presenting several challenges in designing such auto-curricula for successful real-world applications. Another way to achieve automated MSS design is to select among existing MSSs to fit various games or different phases within a game adaptively. For example, Li et al. [2024b] applied hyperparameter optimization to learn weights among multiple MSSs and mixed the outputs of these MSSs based on the weights as the next best response target.

3.2 PSRO with Alternative Game Forms

Extensive-Form Games. Instead of employing the normal-form representation, some PSRO variants use an extensive-form representation for the restricted game, offering a richer way to encompass temporal patterns in actions and information for underlying sequential games. One such example is given by extensive-form DO (XDO) [McAleer *et al.*, 2021]. The extensive-form restricted game tree in XDO again only contains a subset of players' strategies. Similar to DO, NE is deployed as an MSS target, computed through Counterfactual Regret Minimization [Zinkevich *et al.*, 2007], and the best response computation will result in new actions at information states in the restricted game tree. Note that when a new action is added to an information state, multiple strategies will be added. So one iteration in XDO implicitly needs more simulation for profile evaluation than one DO iteration. Periodic DO [Tang *et al.*, 2023] extends XDO by improving the stopping threshold of the restricted game solver.

In a work by Konicki et al. [2022], the benefits of leveraging the extensive-form representation for the restricted game were further explored. They showed that with an extensive-form representation in PSRO, the true game can be approxi-

[3] MRCP was called the *least-exploitable restricted distribution* in the two-player zero-sum context by McAleer et al. [2022b].

mated more accurately than using a normal-form model constructed from the same amount of simulation data. This accuracy improvement stems from the fact that the simulation data for modeling a chance node in extensive form can be reused while the modeling needs to be re-simulated every time for evaluating a profile in normal form.

Mean-Field Games. Muller et al. [2022] and Wang and Wellman [2023a] adapted PSRO to mean-field games (MFGs). Since the utility function for MFGs is not generally linear in the mean field, the restricted MFG cannot be represented explicitly. Instead, Wang and Wellman [2023a] employed a game model learning approach [Sokota *et al.*, 2019], which is essentially a form of regression that learns a utility function over a restricted set of strategies and mean fields derived by these strategies. They proved the existence of NE in the restricted MFG and the convergence of PSRO to NE in MFGs.

4 Strategy Exploration via RO

Besides setting MSSs, establishing ROs to guide strategy exploration in PSRO has also been explored in the literature. One predominant way to design novel ROs is to include diversity measures in the standard RO, aimed at increasing the diversity of strategies in the restricted game. The importance of maintaining a diverse population of strategies has been demonstrated by many works in various domains [Czarnecki *et al.*, 2020; Zahavy *et al.*, 2023].

Specifically, Perez-Nieves et al. [2021] proposed diverse PSRO, which incorporates expected cardinality, a diversity measure defined through a determinantal point process, into the standard RO. Liu et al. [2021] define behavioral diversity (BD) and response diversity (RD), measuring diversity on different scales. BD is defined based on the distance between action-state coverages given by different strategies, while RD measures the distance between the payoff vector induced by the new strategy and the current restricted game. Subsequently, Liu et al. [2022c] proposed unified diversity measures to capture a variety of diversity metrics, which were later combined with the standard RO. Yao et al. [2023] observed that the current diversity measures for enlarging the gamescape fail to connect to the quality of NE approximation. They connected the RO to the quality of NE approximation by introducing population exploitability (PE), which is essentially the regret of MRCP [Wang *et al.*, 2022], to reflect the coverage of a policy hull (i.e., the convex combinations of strategies in the restricted game). They showed that a larger policy hull indicates lower PE. Therefore, they proposed an RO variant to enlarge the policy hull, aiming at promoting strategy exploration. We list the specifications of different methods for enhancing diversity in Table 1.

Aside from diversity, Li et al. [2023c] deployed Monte Carlo tree search (MCTS) as the best-response oracle using different values (e.g., social welfare) to update values of nodes along the sample path in the back-propagation step of MCTS. The employment of different back-propagation values can also be viewed as modifications of the RO.

5 Strategy Exploration via Joint MSS-RO

Generally speaking, MSSs and ROs have a coupled impact on strategy exploration. In this section, we discuss prior works that jointly vary MSSs and ROs.

α-Rank. Muller et al. [2020] proposed the adoption of α-Rank [Omidshafiei *et al.*, 2019] as the preferred solution concept due to its computational scalability and uniqueness in many-player general-sum games. To make PSRO with α-Rank as MSS converge, they introduced the ***preference-based best-response oracle***, which essentially returns a set of strategies that maximizes the probability mass under α-Rank from the set of better responses to the current strategy profile. With this MSS-RO combination, they proved that PSRO with α-Rank converges to something that they call a ***sink strongly-connected component***, which describes the distribution of strategies in long-term interactions. Yang et al. [2020] highlighted that as the number of players significantly increases, the Markov chain required in the computation of α-Rank becomes prohibitively large, yielding a scalability problem for α-Rank. To handle this challenge, they developed an efficient implementation of α-Rank based on DO and stochastic optimization.

Investigating the Joint Impact of MSSs and ROs. An empirical study by Wang and Wellman [2024] explicitly investigated the coupling impact of MSSs and ROs in strategy exploration. This research experimented with many MSS-RO combinations with unique characteristics. Their experimental results underscore the pivotal role of ROs in steering strategy exploration towards desired objectives, such as higher social welfare. Moreover, they showed that with a careful selection of MSS, the performance of strategy exploration can be further improved.

6 Evaluating Strategy Exploration

In addition to designing novel strategy exploration algorithms, a significant effort has also been put into investigating methodological considerations in evaluating strategy exploration, and proposing and justifying new evaluation methods. In PSRO, it may seem natural to employ the same MSS for both strategy generation and evaluation, as much of the prior works in PSRO exploration have done. For example, if NE is used as the MSS for strategy generation, then the regret of NE of intermediate restricted games will be used as the performance measure at each iteration of PSRO. Similarly, the regret of a uniform distribution over strategies will be the performance measure when the uniform distribution is employed as MSS, in which case PSRO reduces to fictitious play.

Wang and Wellman [2022] argued that this evaluation approach could yield a misleading conclusion about the performance of MSS-RO combinations. They highlighted that each MSS-RO combination essentially generates a distinct sequence of strategies, and thus the restricted game at any point reflects a distinct strategy space. The comparisons of different MSS-RO combinations are across different strategy spaces, which may not be faithfully represented by a simple summary such as an interim solution. Therefore, they proposed to use the regret of MRCP, where MRCP is the profile

Diversity Measure	Concept	Diversity Type		Enlargement Target		Compatible MSS	
		Behavioral	Response	Gamescape	Policy Hull	Nash	α-Rank
Effective Diversity [Balduzzi *et al.*, 2019]	Rectified Nash strategy		✓	✓		✓	
Expected Cardinality [Perez-Nieves *et al.*, 2021]	Determinantal point processes		✓	✓		✓	✓
Convex Hull Enlargement [Liu *et al.*, 2021]	Euclidean projection		✓	✓		✓	
Occupancy Measure Mismatching [Liu *et al.*, 2021]	f-divergence, occupancy measure	✓		✓		✓	
Unified Diversity Measure [Liu *et al.*, 2022c]	Strategy feature, diversity kernel		✓	✓		✓	✓
Policy Space Diversity [Yao *et al.*, 2023]	Bregman divergence, sequence-form	✓			✓	✓	

Table 1: Specifications of promoting-diversity methods.

closest to the full-game NE (in regret) in the restricted game, as the evaluation metric for evaluating the performance of multiple MSS-RO combinations. This means that the MSS employed for strategy generation is independent of the MSS (e.g., MRCP) used for evaluation, which should be fixed when comparing different restricted games, regardless of the MSS with which they are generated.

7 Improvements in Training Efficiency

There are two components in PSRO that can be computationally demanding: best response computation and payoff simulation in the restricted game. To improve the efficiency of PSRO, various methods have been developed, addressing issues related to these two aspects.

Parallelization. Leveraging parallelization, Lanctot et al. [2017] proposed the Deep Cognitive Hierarchy (DCH) model, which creates a training hierarchy where each player trains a best response strategy (with deep RL) against the NE of the restricted game with strategies at the same level or below it. This *warm-starts* best-response training, and speeds up PSRO compared to training best responses from scratch. Motivated by DCH, McAleer et al. [2020] proposed Pipeline PSRO (P2SRO). Similar to DCH, P2SRO initializes a bunch of strategies and assigns each strategy a level. Then P2SRO warm-starts training each strategy in parallel against the NE of the restricted game involving strategies with lower levels, which accelerates the overall training of PSRO.

Sample Efficiency. A distinctive characteristic of restricted games is that they are derived or estimated from simulation data. To improve the sample efficiency of PSRO, Smith and Wellman [2023] proposed to learn a full-game model about the game dynamics from the simulator concurrently with running PSRO, aimied at reducing the simulation cost by querying the full-game model. Zhou et al. [2022] developed an efficient PSRO (EPSRO) implementation for reducing the simulation cost of PSRO in two-player zero-sum games. The key insight is that the simulation for the restricted game is only used for computing best-response target profiles. So as long as best-response target profiles can be computed in other ways (e.g., a uniform MSS does not need the evaluation of a restricted game), there is no need to maintain the complete restricted game, avoiding unnecessary simulations.

Transfer learning. Transfer learning is a machine learning technique where a model trained on one task is repurposed for a different task. In PSRO, best-responding to different strategies can be viewed as such tasks and thus transfer learning can be applied to warm-start training new best responses. One example leveraging this idea is NeuPL [Liu *et al.*, 2022b], which represents all strategies in the strategy set via a shared neural network. NeuPL utilizes explicit parameter sharing for skill transfer, which was shown to effectively accelerate the adaptation to the opponent's meta-strategy. Liu et al. [2022a] further generalized NeuPL by optimizing best-responses against mixed-strategy profiles randomly sampled from the current restricted game, offering approximate optimality against any mixture over a diverse set of strategies at test time. Liu et al. [2024] combined NeuPL and JPSRO, enabling the transfer of knowledge in the computation of (C)CE with PSRO. Smith et al. [2023] also utilized transfer learning to reduce simulation costs in computing best responses. Their Mixed-Oracle method constructs a new best response by learning and maintaining best responses to the pure strategies of the opponent (represented as Q-value functions) and then mixing (Q-values) according to the meta-strategy.

8 Applications

Game-theoretic analysis in PSRO relies on restricted games, which abstract away the underlying game structures. This abstraction enables PSRO to solve a variety of games or address issues that can be formulated as a game, and yields numerous applications of PSRO in disparate domains. Specifically, PSRO has been applied to specialized games, including security games [Wang *et al.*, 2019; Wright *et al.*, 2019; Fang, 2019; Tong *et al.*, 2020; Xu *et al.*, 2021; Cui and Yang, 2023], bargaining games [Li *et al.*, 2023c; Wang and Wellman, 2024], Colonel Blotto games [An and Zhou, 2023], Google Research Football environments [Liu *et al.*, 2021; Song *et al.*, 2024], chess [Zahavy *et al.*, 2023; Li *et al.*, 2023b], Pursuit-Evasion games [Li *et al.*, 2023a; Li *et al.*, 2024a], auctions [Li and Wellman, 2021], and mechanism design for sequential auctions [Zhang *et al.*, 2023], which are among a class of the hardest extensive-form games to solve.

Moreover, PSRO applications over time have been extended to real-world domains including anti-jamming in satellite communication [Zou *et al.*, 2022], decision-making in beyond-visual-range air combat [Ma *et al.*, 2019], solving the power imbalance in power system resilience [Niu *et al.*, 2021], and tackling the traveling salesman problem in combinatorial optimization [Wang *et al.*, 2021]. Its utility is also evident in social network analysis for competitive influence maximization strategies [Ansari *et al.*, 2019] and in developing defense strategies for election safety [Yin *et al.*, 2018].

Besides these real-world applications, PSRO (including DO) has also been applied for designing novel algorithms in domains that can be modeled as a game. For example, the PSRO framework has facilitated developments in RL for robust policy discovery [Liang *et al.*, 2023] and policy generalization [Yang *et al.*, 2022], multiagent RL evaluation [Li and Wellman, 2024], value alignment in large language models [Ma *et al.*, 2023], public health services [Killian *et al.*, 2023], combinatorial optimization [Wang *et al.*, 2024], and the discovery of information in images [Giboulot *et al.*, 2023].

PSRO has also had an impact on enhancing computer vision and structured prediction methodologies, for example, via data augmentation for object detection [Behpour *et al.*, 2019a], active learning [Behpour *et al.*, 2019b], semi-supervised multi-label classification [Behpour, 2018], and video tracking [Fathony *et al.*, 2018]. Furthermore, PSRO-type methods have been adapted to enhance the training of Generative Adversarial Networks (GAN) [Oliehoek *et al.*, 2019; Aung *et al.*, 2022].

9 PSRO Implementations

Two software libraries that include PSRO implementations are OpenSpiel [Lanctot *et al.*, 2019] and MALib [Zhou *et al.*, 2023]. Both serve as comprehensive toolsets including various games and algorithms for exploring general reinforcement learning and search or planning in games.

10 Open Research Questions

Here, we briefly describe a few further research directions. A longer discussion of these directions can be found in the arXiv version[4].

Scalability in the Number of Players. In the standard PSRO framework, a restricted game is represented in normal form. As the number of players increases, the normal-form representation expands exponentially, leading to a significant increase in the cost of evaluating the restricted game. While this problem can be mitigated in certain special game types such as symmetric games, the practicality of applying PSRO diminishes when dealing with a very large number of players.

Multiple Equilibria. Another important question relates to the existence and computation of multiple equilibria, particularly within general-sum games. While current PSRO research primarily focuses on identifying a sample equilibrium, the capability of PSRO to compute multiple equilibria remains under-explored.

Combining PSRO with Subgame Solving or CFR. CFR-based and policy-gradient-based methods might perform better in games where mixing carefully at many decision nodes is crucial, providing more granular mixing at different game scales (e.g., information sets that define subgames). In contrast, PSRO usually mixes only at the root of the game, limited to distributions over strategies in the restricted game, and might require many policies to effectively mix at information sets [McAleer *et al.*, 2021]. Therefore, for game sections that do not require mixing such as complex control tasks, it would be more efficient to use the deep RL policies from PSRO or XDO, while decision nodes that require careful mixing should defer to a CFR-based or policy-gradient-based technique. Overall, a hierarchical PSRO structure that controls the granularity of the mixing might be needed.

Acknowledgements

We thank Marc Lanctot, Karl Tuyls, Michael P. Wellman, Yaodong Yang, and their students, for their invaluable advice on improving the survey.

Contribution Statement

Ariyan Bighashdel and Yongzhao Wang contributed equally.

References

[Albrecht *et al.*, 2024] S. V. Albrecht, F. Christianos, and L. Schäfer. *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches*. MIT Press, 2024.

[An and Zhou, 2023] Z. An and L. Zhou. Double Oracle Algorithm for Game-Theoretic Robot Allocation on Graphs. *arXiv:2312.11791*, 2023.

[Angeline *et al.*, 1993] P. J. Angeline, J. B. Pollack, and Others. Competitive Environments Evolve Better Solutions for Complex Tasks. In *ICGA*, 1993.

[Ansari *et al.*, 2019] A. Ansari, M. Dadgar, A. Hamzeh, J. Schlötterer, and M. Granitzer. Competitive influence maximization: integrating budget allocation and seed selection. *arXiv:1912.12283*, 2019.

[Aung *et al.*, 2022] A. P. P. Aung, X. Wang, R. Yu, B. An, S. Jayavelu, and X. Li. DO-GAN: A Double Oracle Framework for Generative Adversarial Networks. In *CVRP*, 2022.

[Avis *et al.*, 2010] D. Avis, G. D. Rosenberg, R. Savani, and B. Von Stengel. Enumeration of nash equilibria for two-player games. *Economic theory*, 42, 2010.

[Balduzzi *et al.*, 2018] D. Balduzzi, K. Tuyls, J. Pérolat, and T. Graepel. Re-evaluating evaluation. In *NIPS*, 2018.

[Balduzzi *et al.*, 2019] D. Balduzzi, M. Garnelo, Y. Bachrach, W. Czarnecki, J. Perolat, M. Jaderberg, and T. Graepel. Open-ended learning in symmetric zero-sum games. In *ICML*, 2019.

[Behpour *et al.*, 2019a] S. Behpour, K. M. Kitani, and B. D. Ziebart. ADA: Adversarial data augmentation for object detection. In *WACV*, 2019.

[Behpour *et al.*, 2019b] S. Behpour, A. Liu, and B. Ziebart. Active learning for probabilistic structured prediction of cuts and matchings. In *ICML*, 2019.

[Behpour, 2018] S. Behpour. Arc: Adversarial robust cuts for semi-supervised and multi-label classification. In *CVPR*, 2018.

[Benders, 1962] J. F. Benders. Partitioning procedures for solving mixed-variables programming problems. *Numerische Mathematik*, 4, 1962.

[4] The arXiv version: https://arxiv.org/pdf/2403.02227.

[Brown, 1951] G. W. Brown. Iterative solution of games by fictitious play. *Act. Anal. Prod Allocation*, 13, 1951.

[Celli and Gatti, 2018] A. Celli and N. Gatti. Computational results for extensive-form adversarial team games. In *AAAI*, 2018.

[Cui and Yang, 2023] J. Cui and X. Yang. Macta: A multi-agent reinforcement learning approach for cache timing attacks and detection. *ICLR*, 2023.

[Czarnecki *et al.*, 2020] W. M. Czarnecki, G. Gidel, B. Tracey, K. Tuyls, S. Omidshafiei, D. Balduzzi, and M. Jaderberg. Real world games look like spinning tops. In *NeurIPS*, 2020.

[Dinh *et al.*, 2022] L. C. Dinh, S. M. McAleer, Z. Tian, N. Perez-Nieves, O. Slumbers, D. H. Mguni, J. Wang, H. B. Ammar, and Y. Yang. Online double oracle. *TMLR*, 2022.

[Fang, 2019] F. Fang. Integrate learning with game theory for societal challenges. In *IJCAI*, 2019.

[Fathony *et al.*, 2018] R. Fathony, S. Behpour, X. Zhang, and B. Ziebart. Efficient and consistent adversarial bipartite matching. In *ICML*, 2018.

[Feng *et al.*, 2021] X. Feng, O. Slumbers, Z. Wan, B. Liu, S. McAleer, Y. Wen, J. Wang, and Y. Yang. Neural autocurricula in two-player zero-sum games. In *NeurIPS*, 2021.

[Ficici and Pollack, 2003] S. G. Ficici and J. B. Pollack. A game-theoretic memory mechanism for coevolution. In *GECCO*, 2003.

[Gemp *et al.*, 2022] I. Gemp, R. Savani, M. Lanctot, Y. Bachrach, T. W. Anthony, R. Everett, A. Tacchetti, T. Eccles, and J. Kramár. Sample-based approximation of nash in large many-player games via gradient descent. In *AAMAS*, 2022.

[Giboulot *et al.*, 2023] Q. Giboulot, T. Pevný, and A. D. Ker. The non-zero-sum game of steganography in heterogeneous environments. *IEEE Transactions on Information Forensics and Security*, 2023.

[Johanson *et al.*, 2012] M. Johanson, N. Bard, N. Burch, and M. Bowling. Finding optimal abstract strategies in extensive-form games. In *AAAI*, 2012.

[Jordan *et al.*, 2010] P. R. Jordan, L. J. Schvartzman, and M. P. Wellman. Strategy exploration in empirical games. In *AAMAS*, 2010.

[Killian *et al.*, 2023] J. A. Killian, A. Biswas, L. Xu, S. Verma, V. Nair, A. Taneja, A. Hegde, N. Madhiwalla, P. R. Diaz, S. Johnson-Yu, and Others. Robust planning over restless groups: engagement interventions for a large-scale maternal telehealth program. In *AAAI*, 2023.

[Konicki *et al.*, 2022] C. Konicki, M. Chakraborty, and M. P. Wellman. Exploiting extensive-form structure in empirical game-theoretic analysis. In *WINE*, 2022.

[Lanctot *et al.*, 2017] M. Lanctot, V. Zambaldi, A. Gruslys, A. Lazaridou, K. Tuyls, J. Pérolat, D. Silver, and T. Graepel. A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning. *NIPS*, 2017.

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

[Li and Wellman, 2021] Z. Li and M. P. Wellman. Evolution strategies for approximate solution of Bayesian games. In *AAAI*, 2021.

[Li and Wellman, 2024] Z. Li and M. P. Wellman. A metagame evaluation framework for deep multiagent reinforcement learning. In *IJCAI*, 2024.

[Li *et al.*, 2023a] S. Li, X. Wang, Y. Zhang, W. Xue, J. Černỳ, and B. An. Solving large-scale pursuit-evasion games using pre-trained strategies. In *AAAI*, 2023.

[Li *et al.*, 2023b] Y. Li, K. Xiong, Y. Zhang, J. Zhu, S. Mcaleer, W. Pan, J. Wang, Z. Dai, and Y. Yang. Jiangjun: Mastering Xiangqi by tackling non-transitivity in two-player zero-sum games. *TMLR*, 2023.

[Li *et al.*, 2023c] Z. Li, M. Lanctot, K. R. McKee, L. Marris, I. Gemp, D. Hennes, P. Muller, K. Larson, Y. Bachrach, and M. P. Wellman. Combining Tree-Search, Generative Models, and Nash Bargaining Concepts in Game-Theoretic Reinforcement Learning. *arXiv:2302.00797*, 2023.

[Li *et al.*, 2024a] P. Li, S. Li, X. Wang, J. Cerny, Y. Zhang, S. McAleer, H. Chan, and B. An. Grasper: A generalist pursuer for pursuit-evasion problems. In *AAMAS*, 2024.

[Li *et al.*, 2024b] P. Li, S. Li, C. Yang, X. Wang, X. Huang, H. Chan, and B. An. Self-adaptive psro: Towards an automatic population-based game solver. In *IJCAI*, 2024.

[Liang *et al.*, 2023] Y. Liang, Y. Sun, R. Zheng, X. Liu, T. Sandholm, F. Huang, and S. McAleer. Game-theoretic robust reinforcement learning handles temporally-coupled perturbations. *arXiv:2307.12062*, 2023.

[Liu *et al.*, 2021] X. Liu, H. Jia, Y. Wen, Y. Hu, Y. Chen, C. Fan, Z. Hu, and Y. Yang. Towards unifying behavioral and response diversity for open-ended learning in zero-sum games. In *NeurIPS*, 2021.

[Liu *et al.*, 2022a] S. Liu, M. Lanctot, L. Marris, and N. Heess. Simplex neural population learning: Any-mixture bayes-optimality in symmetric zero-sum games. In *ICML*, 2022.

[Liu *et al.*, 2022b] S. Liu, L. Marris, D. Hennes, J. Merel, N. Heess, and T. Graepel. NeuPL: Neural population learning. *arXiv:2202.07415*, 2022.

[Liu *et al.*, 2022c] Z. Liu, C. Yu, Y. Yang, Z. Wu, Y. Li, and Others. A Unified Diversity Measure for Multiagent Reinforcement Learning. In *NeurIPS*, 2022.

[Liu *et al.*, 2024] S. Liu, L. Marris, M. Lanctot, G. Piliouras, J. Z. Leibo, and N. Heess. Neural population learning beyond symmetric zero-sum games. In *AAMAS*, 2024.

[Long *et al.*, 2023] W. Long, T. Hou, X. Wei, S. Yan, P. Zhai, and L. Zhang. A Survey on Population-Based Deep Reinforcement Learning. *Mathematics*, 11, 2023.

[Ma *et al.*, 2019] Y. Ma, G. Wang, X. Hu, H. Luo, and X. Lei. Cooperative occupancy decision making of Multi-UAV in Beyond-Visual-Range air combat: A game theory approach. *IEEE Access*, 8, 2019.

[Ma *et al.*, 2023] C. Ma, Z. Yang, M. Gao, H. Ci, J. Gao, X. Pan, and Y. Yang. Red teaming game: A game-theoretic framework for red teaming language models. *arXiv:2310.00322*, 2023.

[Marris *et al.*, 2021] L. Marris, P. Muller, M. Lanctot, K. Tuyls, and T. Graepel. Multi-agent training beyond zero-sum with correlated equilibrium meta-solvers. In *ICML*, 2021.

[McAleer *et al.*, 2020] S. McAleer, J. B. Lanier, R. Fox, and P. Baldi. Pipeline PSRO: A scalable approach for finding approximate Nash equilibria in large games. In *NeurIPS*, 2020.

[McAleer *et al.*, 2021] S. McAleer, J. B. Lanier, K. A. Wang, P. Baldi, and R. Fox. XDO: A double oracle algorithm for extensive-form games. In *NeurIPS*, 2021.

[McAleer *et al.*, 2022a] S. McAleer, J. B. Lanier, K. Wang, P. Baldi, R. Fox, and T. Sandholm. Self-play psro: Toward optimal populations in two-player zero-sum games. *arXiv:2207.06541*, 2022.

[McAleer *et al.*, 2022b] S. McAleer, K. Wang, M. Lanctot, J. Lanier, P. Baldi, and R. Fox. Anytime optimal psro for two-player zero-sum games. *arXiv:2201.07700*, 2022.

[McAleer *et al.*, 2023] S. M. McAleer, G. Farina, G. Zhou, M. Wang, Y. Yang, and T. Sandholm. Team-PSRO for learning approximate TMECor in large team games via cooperative reinforcement learning. In *NeurIPS*, 2023.

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

[McMahan *et al.*, 2003] H. B. McMahan, G. J. Gordon, and A. Blum. Planning in the presence of cost functions controlled by an adversary. In *ICML*, 2003.

[Muller *et al.*, 2020] P. Muller, S. Omidshafiei, M. Rowland, K. Tuyls, J. Pérolat, S. Liu, D. Hennes, L. Marris, M. Lanctot, E. Hughes, and Others. A Generalized Training Approach for Multiagent Learning. In *ICLR*, 2020.

[Muller *et al.*, 2022] P. Muller, M. Rowland, R. Elie, G. Piliouras, J. Perolat, M. Lauriere, R. Marinier, O. Pietquin, and K. Tuyls. Learning Equilibria in Mean-Field Games: Introducing Mean-Field PSRO. In *AAMAS*, 2022.

[Niu *et al.*, 2021] L. Niu, D. Sahabandu, A. Clark, and R. Poovendran. A game-theoretic framework for controlled islanding in the presence of adversaries. In *GameSec*, 2021.

[Oliehoek *et al.*, 2006] F. A. Oliehoek, E. D. De Jong, and N. Vlassis. The parallel Nash memory for asymmetric games. In *GECCO*, 2006.

[Oliehoek *et al.*, 2019] F. A. Oliehoek, R. Savani, J. Gallego, E. van der Pol, and R. Groß. Beyond local Nash equilibria for adversarial networks. In *BNAIC*, 2019.

[Omidshafiei *et al.*, 2019] S. Omidshafiei, C. Papadimitriou, G. Piliouras, K. Tuyls, M. Rowland, J.-B. Lespiau, W. M. Czarnecki, M. Lanctot, J. Perolat, and R. Munos. α-rank: Multi-agent evaluation by evolution. *Scientific reports*, 9, 2019.

[Perez-Nieves *et al.*, 2021] N. Perez-Nieves, Y. Yang, O. Slumbers, D. H. Mguni, Y. Wen, and J. Wang. Modelling behavioural diversity for learning in open-ended games. In *ICML*, 2021.

[Popovici *et al.*, 2012] E. Popovici, A. Bucci, R. P. Wiegand, and E. D. De Jong. Coevolutionary Principles, 2012.

[Sanjaya *et al.*, 2022] R. Sanjaya, J. Wang, and Y. Yang. Measuring the non-transitivity in chess. *Algorithms*, 15(5):152, 2022.

[Savani and Turocy, 2024] R. Savani and T. L. Turocy. Gambit: The Package for doing Computation in Game Theory, version 16.2.0., 2024.

[Schvartzman and Wellman, 2009] L. J. Schvartzman and M. P. Wellman. Exploring large strategy spaces in empirical game modeling. In *AAMAS-AMEC Workshop*, 2009.

[Shoham *et al.*, 2007] Y. Shoham, R. Powers, and T. Grenager. If multi-agent learning is the answer, what is the question? *Artificial intelligence*, 171, 2007.

[Slumbers *et al.*, 2023] O. Slumbers, D. H. Mguni, S. B. Blumberg, S. M. Mcaleer, Y. Yang, and J. Wang. A game-theoretic framework for managing risk in multi-agent systems. In *ICML*, 2023.

[Smith and Wellman, 2023] M. O. Smith and M. P. Wellman. Co-Learning Empirical Games and World Models. *arXiv:2305.14223*, 2023.

[Smith *et al.*, 2023] M. O. Smith, T. Anthony, and M. P. Wellman. Strategic knowledge transfer. *JMLR*, 24, 2023.

[Sokota *et al.*, 2019] S. Sokota, C. Ho, and B. Wiedenbeck. Learning deviation payoffs in simulation-based games. In *AAAI*, 2019.

[Song *et al.*, 2024] Y. Song, H. Jiang, Z. Tian, H. Zhang, Y. Zhang, J. Zhu, Z. Dai, W. Zhang, and J. Wang. An empirical study on google research football multi-agent scenarios. *Machine Intelligence Research*, pages 1–22, 2024.

[Tang *et al.*, 2023] X. Tang, L. C. Dinh, S. M. Mcaleer, and Y. Yang. Regret-minimizing double oracle for extensive-form games. In *ICML*, 2023.

[Taylor and Jonker, 1978] P. D. Taylor and L. B. Jonker. Evolutionary stable strategies and game dynamics. *Mathematical biosciences*, 40, 1978.

[Tong *et al.*, 2020] L. Tong, A. Laszka, C. Yan, N. Zhang, and Y. Vorobeychik. Finding needles in a moving haystack: Prioritizing alerts with adversarial reinforcement learning. In *AAAI*, 2020.

[Tuyls and Weiss, 2012] K. Tuyls and G. Weiss. Multiagent learning: Basics, challenges, and prospects. *Ai Magazine*, 33(3):41–41, 2012.

[Vinyals *et al.*, 2019] O. Vinyals, I. Babuschkin, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575, 2019.

[von Stengel, 2002] B. von Stengel. Computing equilibria for two-person games. In *Handbook of game theory with economic applications*, volume 3. Elsevier, 2002.

[Wang and Wellman, 2023a] Y. Wang and M. P. Wellman. Empirical Game-Theoretic Analysis for Mean Field Games. In *AAMAS*, 2023.

[Wang and Wellman, 2023b] Y. Wang and M. P. Wellman. Regularization for Strategy Exploration in Empirical Game-Theoretic Analysis. *arXiv:2302.04928*, 2023.

[Wang and Wellman, 2024] Y. Wang and M. P. Wellman. Generalized response objectives for strategy exploration in empirical game-theoretic analysis. In *AAMAS*, 2024.

[Wang *et al.*, 2019] Y. Wang, Z. R. Shi, L. Yu, Y. Wu, R. Singh, L. Joppa, and F. Fang. Deep reinforcement learning for green security games with real-time information. In *AAAI*, 2019.

[Wang *et al.*, 2021] C. Wang, Y. Yang, O. Slumbers, C. Han, T. Guo, H. Zhang, and J. Wang. A game-theoretic approach for improving generalization ability of tsp solvers. *arXiv preprint arXiv:2110.15105*, 2021.

[Wang *et al.*, 2022] Y. Wang, Q. Ma, and M. P. Wellman. Evaluating strategy exploration in empirical game-theoretic analysis. In *AAMAS*, 2022.

[Wang *et al.*, 2024] C. Wang, Z. Yu, S. McAleer, T. Yu, and Y. Yang. Asp: Learn a universal neural solver! *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[Wellman *et al.*, 2024] M. P. Wellman, K. Tuyls, and A. Greenwald. Empirical game-theoretic analysis: A survey. *arXiv preprint arXiv:2403.04018*, 2024.

[Wellman, 2006] M. P. Wellman. Methods for empirical game-theoretic analysis. In *AAAI*, 2006.

[Wright *et al.*, 2019] M. Wright, Y. Wang, and M. P. Wellman. Iterated deep reinforcement learning in games: History-aware training for improved stability. In *ACM EC*, 2019.

[Xu *et al.*, 2021] L. Xu, A. Perrault, F. Fang, H. Chen, and M. Tambe. Robust reinforcement learning under minimax regret for green security. In *UAI*, 2021.

[Yang and Wang, 2020] Y. Yang and J. Wang. An overview of multi-agent reinforcement learning from game theoretical perspective. *arXiv:2011.00583*, 2020.

[Yang *et al.*, 2020] Y. Yang, R. Tutunov, P. Sakulwongtana, and H. B. Ammar. α^α-rank: Practically scaling α-rank through stochastic optimisation. In *AAMAS*, 2020.

[Yang *et al.*, 2021] Y. Yang, J. Luo, et al. Diverse autocurriculum is critical for successful real-world multiagent learning systems. *arXiv preprint arXiv:2102.07659*, 2021.

[Yang *et al.*, 2022] C. Yang, R. Wang, X. Wang, and Z. Wang. A game-theoretic perspective of generalization in reinforcement learning. *arXiv preprint arXiv:2208.03650*, 2022.

[Yao *et al.*, 2023] J. Yao, W. Liu, H. Fu, Y. Yang, S. McAleer, Q. Fu, and W. Yang. Policy space diversity for non-transitive games. *arXiv:2306.16884*, 2023.

[Yin *et al.*, 2018] Y. Yin, Y. Vorobeychik, B. An, and N. Hazon. Optimal defense against election control by deleting voter groups. *Artificial Intelligence*, 259, 2018.

[Zahavy *et al.*, 2023] T. Zahavy, V. Veeriah, S. Hou, K. Waugh, M. Lai, E. Leurent, N. Tomasev, L. Schut, D. Hassabis, and S. Singh. Diversifying AI: Towards creative chess with AlphaZero. *arXiv preprint arXiv:2308.09175*, 2023.

[Zhang and Sandholm, 2024] B. H. Zhang and T. Sandholm. Exponential lower bounds on the double oracle algorithm in zero-sum games. In *IJCAI*, 2024.

[Zhang *et al.*, 2023] B. Zhang, G. Farina, I. Anagnostides, F. Cacciamani, S. McAleer, A. Haupt, A. Celli, N. Gatti, V. Conitzer, and T. Sandholm. Computing optimal equilibria and mechanisms via learning in zero-sum extensive-form games. In *NeurIPS*, 2023.

[Zhao *et al.*, 2023] Z. Zhao, M. Wen, Y. Wen, and Y. Yang. Open-ended learning in general-sum games: The role of diversity in correlated equilibrium. 2023.

[Zhou *et al.*, 2022] M. Zhou, J. Chen, Y. Wen, W. Zhang, Y. Yang, Y. Yu, and J. Wang. Efficient Policy Space Response Oracles. *arXiv:2202.00633*, 2022.

[Zhou *et al.*, 2023] M. Zhou, Z. Wan, H. Wang, M. Wen, R. Wu, Y. Wen, Y. Yang, Y. Yu, J. Wang, and W. Zhang. Malib: A parallel framework for population-based multiagent reinforcement learning. *Journal of Machine Learning Research*, 24(150):1–12, 2023.

[Zinkevich *et al.*, 2007] M. Zinkevich, M. Johanson, M. Bowling, and C. Piccione. Regret minimization in games with incomplete information. *NIPS*, 2007.

[Zou *et al.*, 2022] M. Zou, J. Chen, J. Luo, Z. Hu, and S. Chen. Equilibrium Approximating and Online Learning for Anti-Jamming Game of Satellite Communication Power Allocation. *Electronics*, 11(21):3526, 2022.