Leveraging Offline Data for Large-Scale Multi-Agent Reinforcement Learning

JB Lanier[*]
University of California Irvine
jblanier@uci.edu

Nathan Monette[*†]
University of Oxford
nathan.monette@cs.ox.ac.uk

Roy Fox
University of California Irvine
royf@uci.edu

Abstract

Finding approximate equilibria for large-scale imperfect-information competitive games such as StarCraft, Dota, and CounterStrike remains computationally challenging due to sparse rewards and extensive exploration over long horizons. In this paper, we propose a multi-agent starting-state sampling strategy designed to substantially accelerate online exploration in regularized policy-gradient game methods. Motivated by an assumption that offline demonstrations from skilled humans provide good coverage of high-level strategies relevant to equilibrium play, we propose to initialize reinforcement learning data collection at intermediate states sampled from offline data to facilitate exploration of strategically relevant subgames. Referring to this method as Data-Augmented Game Starts (DAGS), we perform experiments using synthetic datasets and analytically tractable, long-horizon control variants of two-player Kuhn and Leduc poker. When combined with policy regularization towards dataset behavior, DAGS significantly extends the complexity and scale of games solvable by regularized policy gradient methods within fixed computational budgets. We also release a new set of benchmark environments that drastically increase exploration challenges and state counts in existing OpenSpiel games while keeping exploitability measurements analytically tractable.

1 Introduction

Deep reinforcement learning (RL) has achieved remarkable success in large-scale imperfect-information competitive games, surpassing human professionals in StarCraft [1], Dota [2], Go [3], and Poker [4]. However, successes like these require massive computational resources and time, rendering large-scale game solving impractical for most practitioners. One key factor underlying this high computational cost is the difficulty of exploration in environments with sparse rewards, long decision horizons, and complex hidden-state dynamics. Especially in video games, traditional methods may spend enormous amounts of computation repeatedly exploring long sequences of mechanically intricate actions, even though strategically interesting decisions may often occur only at specific junctures.

A widely-adopted approach to mitigate exploration difficulty is initializing RL policies with behavioral cloning from human demonstrations or other offline data [5, 6]. While effective for bootstrapping

[*]Equal Contribution
[†]Work partially done while at the University of California Irvine.

Preprint.

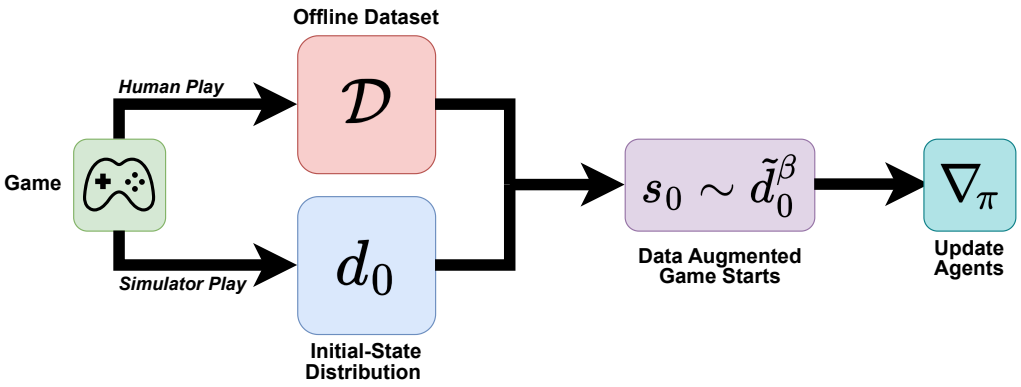

Figure 1: Given an imperfect information two-player zero-sum game, we assume access to its initial-state distribution d_0 as well as an an offline dataset of states $\mathcal{D}$, representative of human-level play. Using a mixture parameter β, we then produce the *augmented* initial-state distribution $\tilde{d}_0^{\beta}$ which also includes intermediate states from trajectories in $\mathcal{D}$. Using data collected with agents π on such a distribution, we perform accelerated training with standard policy gradient methods.

basic competence, significant exploration remains necessary to reach equilibrium play. Single-agent RL has made substantial gains in sample efficiency by leveraging offline demonstrations more actively through intermediate-state resets, beginning online episodes from informative, intermediate states recorded in demonstrations rather than always restarting from initial conditions [7]. Motivated by this success, we propose extending intermediate-state resets to multi-agent competitive settings for the purpose of accelerating equilibrium convergence in large-scale, imperfect-information games.

We focus on games that decompose into a small number of recurring high-level strategic choices separated by long, mechanically demanding control segments. In such games (e.g. fighting games [8] and soccer [9]), the horizon and state spaces are enormous, but the effective set of distinct coarse strategic options that players employ is comparatively small. This makes it plausible to collect human gameplay that covers a broad range of these high-level strategies, so that an RL policy mixing over them can already be close to equilibrium.

Our proposed method, Data-Augmented Game Starts (DAGS), incorporates multi-agent intermediate-state resets into online RL training. DAGS randomly initializes episodes from global states sampled along offline trajectories generated by mechanically proficient play that doesn't need to be strategically optimal (such as typical human gameplay). Our main data requirement is that these offline states provide broad coverage of strategically relevant subgames. By resetting into such states, training can concentrate exploration on resolving high-level strategic interactions rather than discovering and repeatedly relearning the long mechanical action sequences needed to reach them from the initial state. In effect, DAGS amortizes much of the mechanical exploration into the offline data collection phase and reallocates online samples toward refining responses across a richer distribution of strategically relevant subgames, leading to faster convergence toward low-exploitability play.

Evaluating equilibrium convergence in large-scale games is difficult due to the computational infeasibility of exact exploitability calculations. To measure learning performance in larger games, we also introduce a set of analytically tractable benchmarks that extend standard OpenSpiel two-player zero-sum games, such as Kuhn and Leduc Poker, with complex gridworld control subtasks. In these benchmarks, agents must navigate to specific locations to execute their intended inner-game actions, significantly elongating episode lengths and amplifying the exploration challenge. Crucially, policies learned in these gridworld-extended benchmarks can be analytically reduced to equivalent policies in their simpler inner-game counterparts (Kuhn, Leduc, etc), thus enabling exact exploitability calculations despite large state spaces and episode lengths.

Using these novel benchmarks and synthetic offline data, we empirically demonstrate that DAGS significantly extends the scalability and complexity of games solvable within fixed computational budgets using proximal policy optimization (PPO) self-play with strong regularization [10]. Additionally, we show that guiding exploration through KL-regularization toward offline-dataset behaviour

further amplifies these improvements, enabling equilibrium convergence in even larger and more challenging game variants.

However, modifying the starting-state distribution in imperfect-information games can, in principle, distort agents' beliefs over private information and furthermore change a game's equilibria. We provide preliminary analysis and simple counterexamples illustrating this risk, and identify quantifying and mitigating such bias as an important directions for ongoing work.

In summary, our contributions include: (i) a multi-agent intermediate-state reset method (DAGS) leveraging offline demonstrations to accelerate equilibrium convergence, (ii) novel analytically reducible benchmarks for evaluating exploitability convergence in large games, and (iii) empirical validation demonstrating substantial scalability improvements from DAGS across challenging long-horizon competitive tasks. (iv, ongoing work) an empirical study of and mitigation techniques for the bias induced by starting-state augmentation in imperfect-information games.

2 Preliminaries

We model our environments as two-player zero-sum partially observable stochastic games (POSGs) with finite horizon. A POSG is a tuple

$$\mathcal{G} = (N, S, A, O, T, \Omega, r, \gamma, d_0),$$

where $N = \{1, 2\}$ is the set of players, S is the state space, $A = A_1 \times A_2$ is the joint action space, $O = O_1 \times O_2$ is the joint observation space, T is the transition kernel, Ω is the observation function, $r = (r_1, r_2)$ are the per-player rewards, $\gamma \in [0, 1)$ is the discount factor, and d_0 is the initial-state distribution. Each player i follows a stochastic policy $\pi_i(a \mid o)$ mapping observations to action distributions, and a joint policy $\pi = (\pi_1, \pi_2)$ together with d_0 and T induces a distribution over trajectories and returns. We assume $r_1 = -r_2$ and evaluate learned policies using *exploitability*, defined as the average value gain a worst-case best-response opponent can obtain against the current joint policy. Exploitability is zero if and only if the joint policy is a Nash equilibrium.

Our learning setup builds directly on the regularized PPO self-play framework of Rudolph et al. [10]. We represent the policy with a neural network π_θ and use a shared parameterization for both players in symmetric self-play, with the observation encoding player identity. [10] show that PPO with strong entropy regularization closely matches magnetic mirror descent [11] in policy space and can achieve low exploitability in large imperfect-information games. Our method is implemented on top of this regularized PPO self-play baseline.

3 Data-Augmented Game Starts (DAGS)

Our proposed method, Data-Augmented Game Starts (DAGS), augments self-play training with intermediate-state resets drawn from offline trajectories. We first define the offline state dataset and the augmented initial-state distribution, then we describe how DAGS is integrated into regularized PPO self-play.

3.1 Offline state dataset

We assume access to an offline dataset of global states $\mathcal{D} = \{s_m\}_{m=1}^{M}$ collected from trajectories generated by a fixed behavioral policy $\hat{\pi} = (\hat{\pi}_1, \hat{\pi}_2)$. Formally, these states are sampled from the state visitation distribution of $\hat{\pi}$ in the underlying POSG, $s_m \sim d_{\hat{\pi}}$, where $d_{\hat{\pi}}$ denotes the discounted occupancy measure induced by $(d_0, \hat{\pi})$. In practice, $\hat{\pi}$ is chosen to be mechanically competent in the control aspects of the game (e.g., reliably navigating to action locations in our control benchmarks), but not necessarily strategically near equilibrium. This reflects the realistic setting where we may have access to human or scripted gameplay that exhibits high control skill but with suboptimal high-level strategy.

Each entry s_m in $\mathcal{D}$ is a full environment state in the POSG, including all hidden information and simulator variables. When we reset to s_m during training, each player i observes only their own observation $o^i = \Omega(s_m, i)$, and the environment continues to evolve according to the original transition dynamics T. Thus DAGS modifies only the starting-state distribution while preserving the game's information structure and dynamics.

Algorithm 1 Data-Augmented Game Starts (DAGS) with PPO self-play

Require: POSG $\mathcal{G}$ with initial distribution d_0, offline dataset $\mathcal{D}$, mixture parameter $\beta \in [0, 1]$, joint policy $\pi_\theta = (\pi_{\theta_1}, \pi_{\theta_2})$, PPO hyperparameters

while not converged **do**
  Initialize empty batch $\mathcal{B}$
  for episode $= 1, \ldots, K$ **do**
    Sample $u \sim \mathrm{Uniform}(0, 1)$
    if $u < \beta$ **then**
      Sample initial state s_0 from $\mathcal{D}$ ▷ DAGS reset
    else
      Sample initial state $s_0 \sim d_0$ ▷ root start
    Roll out self-play episode from s_0 using joint policy π_θ
    Add the resulting trajectory data to $\mathcal{B}$
  $\theta \leftarrow \mathrm{PPOUPDATES}(\theta, \mathcal{B})$

3.2 Augmented initial-state distribution

Let d_0 denote the original initial state distribution of the game. DAGS defines a new initial-state distribution $\tilde{d}_0$ that places probability mass on both the root states and offline states. We write

$$\tilde{d}_0^\beta = (1 - \beta)\, d_0 + \beta\, d_{\mathcal{D}}, \qquad \beta \in [0, 1], \tag{1}$$

where $d_{\mathcal{D}}$ is the empirical distribution over $\mathcal{D}$, for example $d_{\mathcal{D}}(s) = \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}\{s = s_m\}$. The mixture coefficient β controls the fraction of episodes that begin from offline states. The special case $\beta = 0$ recovers standard root-start self-play, while $\beta = 1$ corresponds to always starting from offline states. In our main experiments we use a simple choice of $d_{\mathcal{D}}$ (uniform over $\mathcal{D}$), leaving more sophisticated sampling schemes for future work.

Training with DAGS can be viewed as solving an *augmented game* in which the only change is that episodes start from $\tilde{d}_0^\beta$ instead of d_0. From a reinforcement-learning perspective, this reweights states according to their frequency in the offline data and the mixture parameter β, thereby directing exploration toward strategically relevant subgames that are hard to reach by naive exploration from the root.

3.3 Self-play training with DAGS

We integrate DAGS into PPO self-play by changing only how initial states are sampled. This process is formalized in Algorithm 1.

In large strategy spaces with sparse rewards, we optionally add a simple imitation-learning regularizer that biases exploration toward offline behavior. Concretely, we fit an imitation policy $\hat{\pi}$ to $\mathcal{D}$ via behavioral cloning, and augment the PPO loss with a small KL penalty $\mathcal{L}_{\mathrm{IL}}$,

$$\mathcal{L}_{\mathrm{IL}} = \lambda_{\mathrm{IL}} \cdot \mathbb{E}_t[\mathrm{KL}\big(\pi_\theta(\cdot \mid o_t) \,\|\, \hat{\pi}(\cdot \mid o_t)\big)]. \tag{2}$$

This encourages π_θ to stay near human-like behavior early in training while self-play still determines the final policy through the PPO objective.

DAGS amortizes much of the exploration required to discover long mechanical control sequences into the offline data collection phase. Online self-play no longer needs to repeatedly rediscover how to navigate to strategically important regions of state space from the root; instead, it can start directly from those regions and allocate its sample budget to refining high-level strategic behavior. In our control Kuhn and Leduc benchmarks, this leads to substantially improved exploitability and scalability compared to root-start PPO self-play. At the same time, because DAGS changes the initial-state distribution in an imperfect-information setting, it introduces the possibility of equilibrium bias, which we analyze empirically and work to address in ongoing work.

4 Benchmarking Exploration in Large Games

Accurately benchmarking exploration in large-scale imperfect-information games is often infeasible, since computing exploitability in realistic domains (e.g., RTS or FPS games) is intractable. To study methods like DAGS in a controlled but challenging setting, we construct a family of benchmark games that wrap simple analytically tractable two-player zero-sum games with structured gridworld control tasks. This construction greatly increases horizon and state space while preserving a tractable reduction back to the original base game, allowing us to measure exact exploitability.

We implement these benchmarks as two successive transformations around OpenSpiel [12] games like Kuhn and Leduc poker. For any given base game, we first introduce a forfeit action to every decision node to represent failure to perform required control tasks. Then, we introduce a multi-step gridworld navigation task at every base game decision node. The goal that the player reaches in this subtask determines the base game action to be taken. To necessitate effective exploration, failure to reach a goal results in a forfeit.

4.1 Forfeit Transformation

Given any base two-player zero-sum game with utilities in a bounded range $[u_{\min}, u_{\max}]$, the forfeit transformation augments every decision state of each player with an additional "surrender" action. If a player chooses this forfeit action, the episode terminates and payoffs are set to

$$u_{\text{forfeit}} \;=\; \max\big(|u_{\min}|, |u_{\max}|\big),$$

so that the forfeiting player receives $-u_{\text{forfeit}}$ and the opponent receives $+u_{\text{forfeit}}$. This makes forfeiting as bad as the worst outcome of the base game and serves as a simple model of failing to execute required mechanical control tasks. For simultaneous-move base games, we convert the base game to an equivalent sequential-move form (player 0 moves first, then player 1) so that at most one player can forfeit at any decision point.

4.2 Control Game Transformation

Given a game with added forfeit actions, the control game transformation turns each base-game decision into a gridworld navigation problem. At every decision point of the base game, the acting player is placed in a $G_x \times G_y$ grid that contains one designated "action square" for each available base-game action. The player observes the concatenation of their own grid coordinates and the base-game information state. The action space is transformed from that of the base game and forfeit actions to discrete navigation controls (up, down, left, right, stay). The player must reach one of the action squares before a fixed time limit, and their final location will select the action taken in the base game decision node, each action square corresponds to a base game action, and failure to land on an action square when the time limit reaches causes the player to taken the forfeit base game action.

A key property of this construction is that policies in the control game are analytically reducible to policies in the base game. For each base-game information state, we evaluate the control policy over grid states and compute the induced probabilities of terminating in each action square or in forfeiting before the timer expires. These probabilities define an equivalent mixed strategy over base-game actions plus an explicit forfeit probability, which we use to compute exact exploitability in the base Kuhn or Leduc game. The complexity of this reduction is linear in the number of base-game information states times the number of grid cells, and is negligible compared to self-play training. In the experiments, we instantiate these wrappers to obtain "Control Kuhn" and "Control Leduc" benchmarks at varying grid sizes, which produce long-horizon, sparse-reward games specifically designed to stress exploration and highlight the benefits of DAGS.

5 Experiments

In order to explore our method's empirical performance, we evaluate DAGS on "control" versions of Kuhn and Leduc poker using the OpenSpiel framework [12]. To serve as offline datasets for control Kuhn and Leduc, we generate 1000 games using a hard-coded policy to uniformly select base-game actions and correctly execute them in the control game subtasks by navigating to the corresponding action squares. By choosing base game actions at random, each game dataset's behavioural policy is suboptimal, and purely performing imitation learning does not result in a low-exploitability policy.

For our base 2p0s game solving algorithm, we use entropy-regularized PPO [13] in self-play. We performed widespread sweeps over hyperparameters for variations of the method for all environments, and report the best results for each method.

5.1 DAGS and Accelerated Exploration

In Figure 2, we compare the exploitability of PPO aginst our proposed variants that include DAGS and the optional imitation learning loss from eq. 2. We evaluate on 3 different gridworld variants of Kuhn and Leduc Poker, where the "size" parameter refers to the gridworld Manhattan distance between base-game action squares.

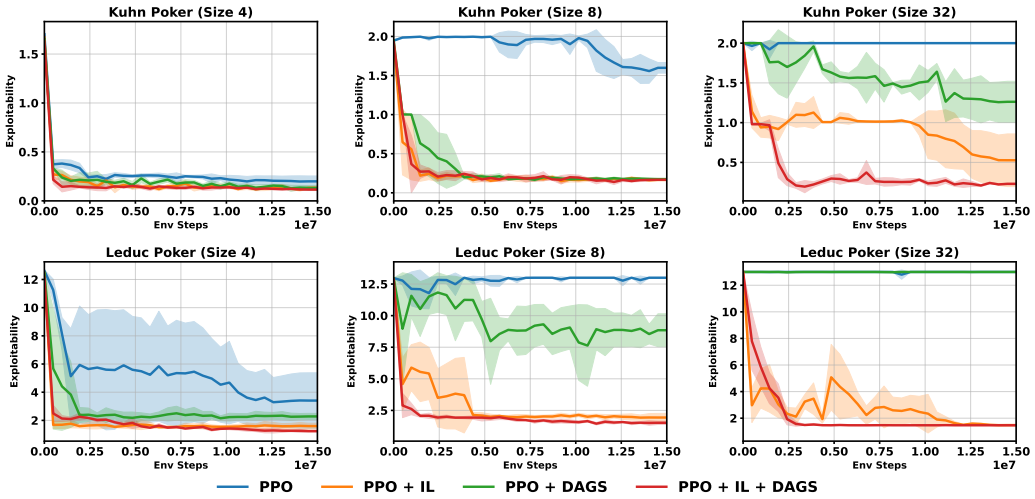

Figure 2: Comparison of exploitability curves across methods on Kuhn and Leduc poker. Using a combination of DAGS and IL can allow PPO to achieve low exploitability in significantly larger games.

After a wide sweep of hyperparameters across grid sizes 4, 8, and 32, we observe across four seeds that combining IL regularization with DAGS leads to the lowest exploitability, particularly as the grid size increases. While PPO is able to achieve low exploitability in the small size 4 game variants, it fails reach the same lever of performance in larger games. This suggests that DAGS assists with exploration, and this benefit becomes more prominent when exploration challenges become difficult in larger games. "PPO + IL" achieves low exploitability by the end of training in larger environments. As the size of the game increase, "PPO + IL + DAGS" decreases in exploitability faster due to the assisted exploration in larger in more sparsely-rewarded game environments.

5.2 Limitations and DAGS-Induced Bias

DAGS can improve RL exploration in 2p0s games, but it may also a create a biased game with a NE that is changed from the original game (exemplified in the "counterexample" game in Figure 3). We demonstrate and propose a fix to this bias by conditioning the policy with a value $f \in \{0, 1\}$, which corresponds to $\mathbf{1}\{s_0 \sim \tilde{d}_0^\beta\}$, the indicator representing if the starting state was sampled from the augmented distribution. When $\beta = 1$, we set $f = 1$ because the policy only sees the augmented distribution during training. For $\beta \in (0, 1)$, we condition the policy on f accordingly during training, and the f in Figure 4 represents the value of f at evaluation. Under appropriate β, equilibrium bias from DAGS may be mitigated.

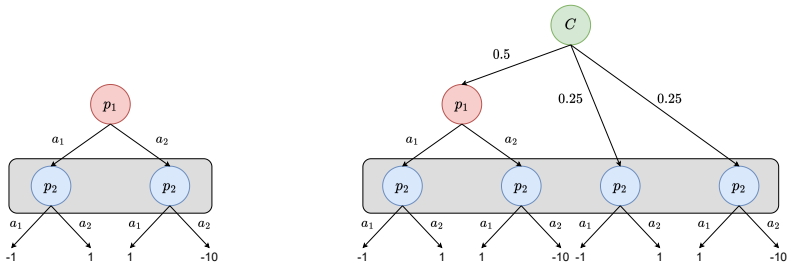

Figure 3: Counterexample Game: A two-player zero-sum game that has a different NE when its initial state distribution is augmented to uniform dataset over a uniform policy by using DAGS with $\beta = 1$. C indicates a chance node, p_i are the nodes corresponding to the turn of player i, and the leaf nodes represent the payoffs of player 0. The grey box indicates information states that appear the same to that tree-level's corresponding player.

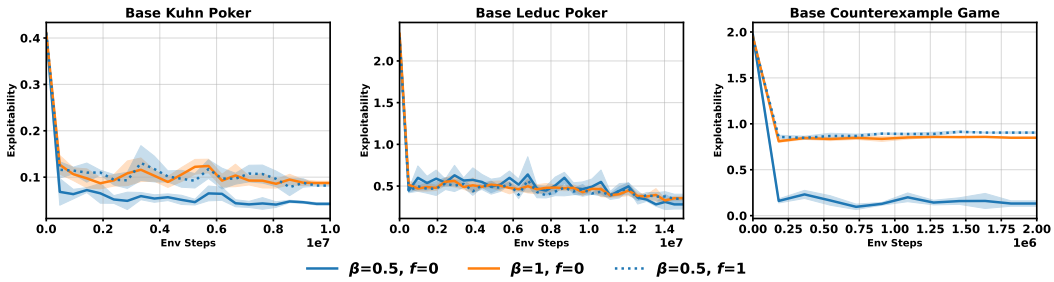

Figure 4: Comparison of exploitability curves across half and full augmented DAGS training. Experiments are performed on the base versions of each game rather than the modified gridworld control variants. A flag value $f = 0$ indicates to the agent that the current episode started from the original game root node, while $f = 1$ signals that the episode was resumed from a sampled dataset state. Training with $\beta = 0.5$ and evaluating with $f = 0$ allows the agent to better approximate the original game NE.

We observe that evaluating with $f = 1$ leads to equilibrium bias (i.e. higher exploitability). This bias is likely due to the augmented distribution inducing incorrect beliefs about opponent behavior. When $\beta = 0.5$ and $f = 0$, exploitability is noticeably lower. By training on both the modified and original starting state distributions with a in-observation flag indicating the source of the episode start, agents are able to benefit from improved DAGS exploration while also still an unbiased NE to the original game.

6 Conclusion and Future Work

In our experiments, we observed that having access to an offline dataset, even one that is very strategically suboptimal, can be used to improve learning in difficult two-player zero-sum games. Specifically, we propose a method for utilizing such an offline dataset to accelerate exploration and optionally combine it with regularization towards a behavior-cloned policy from the dataset. We then introduced a new suite of benchmarks involving high-level strategy alongside control, which is intended to be representative of much more challenging games while still maintaining tractable optimality. Empirically, we observe that our method improves performance over its baselines in the Kuhn and Leduc Poker instantiations of this benchmark.

While we continue this work, we intend to further explore the equilibrium bias produced by DAGS and how to counteract it. Finally, we would like to evaluate how representative our proposed benchmark is of more complex games by examining the relative performance improvements offered by DAGS on existing large games where exact exploitability analyses are not tractable.

References

[1] Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575:350–354, 2019. doi: 10.1038/s41586-019-1724-z. URL https://www.nature.com/articles/s41586-019-1724-z#citeas.

[2] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, Rafal Józefowicz, Scott Gray, Catherine Olsson, Jakub W. Pachocki, Michael Petrov, Henrique Pond'e de Oliveira Pinto, Jonathan Raiman, Tim Salimans, Jeremy Schlatter, Jonas Schneider, Szymon Sidor, Ilya Sutskever, Jie Tang, Filip Wolski, and Susan Zhang. Dota 2 with large scale deep reinforcement learning. *ArXiv*, 2019.

[3] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. 529(7587):484–489, 2016.

[4] Noam Brown, Anton Bakhtin, Adam Lerer, and Qucheng Gong. Combining deep reinforcement learning and search for imperfect-information games. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17057–17069. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c61f571dbd2fb949d3fe5ae1608dd48b-Paper.pdf.

[5] Seunghyun Lee, Younggyo Seo, Kimin Lee, Pieter Abbeel, and Jinwoo Shin. Offline-to-online reinforcement learning via balanced replay and pessimistic q-ensemble. In Aleksandra Faust, David Hsu, and Gerhard Neumann, editors, *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 1702–1712. PMLR, 08–11 Nov 2022. URL https://proceedings.mlr.press/v164/lee22d.html.

[6] Ikechukwu Uchendu, Ted Xiao, Yao Lu, Banghua Zhu, Mengyuan Yan, Joséphine Simon, Matthew Bennice, Chuyuan Fu, Cong Ma, Jiantao Jiao, Sergey Levine, and Karol Hausman. Jump-start reinforcement learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 34556–34583. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/uchendu23a.html.

[7] Stone Tao, Arth Shukla, Tse kai Chan, and Hao Su. Reverse forward curriculum learning for extreme sample and demonstration efficiency in reinforcement learning, 2024. URL https://arxiv.org/abs/2405.03379.

[8] HiFight. Footsies. https://hifight.github.io/footsies/, 2018. Video game.

[9] Karol Kurach, Anton Raichuk, Piotr Stańczyk, Michał Zając, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet, and Sylvain Gelly. Google research football: A novel reinforcement learning environment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):4501–4510, Apr. 2020. doi: 10.1609/aaai.v34i04.5878. URL https://ojs.aaai.org/index.php/AAAI/article/view/5878.

[10] Max Rudolph, Nathan Lichtle, Sobhan Mohammadpour, Alexandre Bayen, J. Zico Kolter, Amy Zhang, Gabriele Farina, Eugene Vinitsky, and Samuel Sokota. Reevaluating policy gradient methods for imperfect-information games, 2025. URL https://arxiv.org/abs/2502.08938.

[11] Samuel Sokota, Ryan D'Orazio, J Zico Kolter, Nicolas Loizou, Marc Lanctot, Ioannis Mitliagkas, Noam Brown, and Christian Kroer. A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=DpE5UYUQzZH.

[12] Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, Daniel Hennes, Dustin Morrill, Paul Muller, Timo Ewalds, Ryan Faulkner, János Kramár, Bart De Vylder, Brennan Saeta, James Bradbury, David Ding, Sebastian Borgeaud, Matthew Lai, Julian Schrittwieser, Thomas Anthony, Edward Hughes, Ivo Danihelka, and Jonah Ryan-Davis. Openspiel: A framework for reinforcement learning in games, 2020. URL https://arxiv.org/abs/1908.09453.

[13] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.