PREFERENCE-BASED MULTI-AGENT REINFORCEMENT LEARNING: DATA COVERAGE AND ALGORITHMIC TECHNIQUES

Anonymous authors
Paper under double-blind review

ABSTRACT

We initiate the study of Preference-Based Multi-Agent Reinforcement Learning (PbMARL), exploring both theoretical foundations and empirical validations. We define the task as identifying the Nash equilibrium from a preference-only offline dataset in general-sum games, a problem marked by the challenge of sparse feedback signals. Our theory establishes the upper complexity bounds for Nash Equilibrium in effective PbMARL, demonstrating that single-policy coverage is inadequate and highlighting the importance of unilateral dataset coverage. These theoretical insights are verified through comprehensive experiments. To enhance the practical performance, we further introduce two algorithmic techniques. (1) We propose a Mean Squared Error (MSE) regularization along the time axis to achieve a more uniform reward distribution and improve reward learning outcomes. (2) We propose an additional penalty based on the distribution of the data set to incorporate pessimism, improving stability and effectiveness during training. Our findings underscore the multifaceted approach required for PbMARL, paving the way for effective preference-based multi-agent systems.

1 INTRODUCTION

Large language models (LLMs) have achieved significant progress in natural language interaction, knowledge acquisition, instruction following, planning and reasoning, which has been recognized as the sparks for AGI (Bubeck et al., 2023). The evolution of LLMs fosters the field of agent systems, wherein LLMs act as the central intelligence (Xi et al., 2023). In these systems, multiple LLMs can interact with each other as well as with external tools. For instance, MetaGPT assigns LLM agents various roles, akin to those in a technology company, enabling them to cooperate on complex software engineering tasks (Hong et al., 2023).

Despite some empirical successes in agent systems utilizing closed-source LLMs, finetuning these systems and aligning them with human preferences remains a challenge. Reinforcement learning from human feedback (RLHF) has played an important role in aligning LLMs with human preferences (Christiano et al., 2017; Ziegler et al., 2019). However, unexpected behavior can arise when multiple LLMs interact with each other. In addition, reward design has been a hard problem in multi-agent reinforcement learning (Devlin et al., 2011). Thus, it is crucial to further align the multi-agent system from preference feedback.

We address this problem through both theoretical analysis and empirical experiments. Theoretically, we characterize the dataset coverage condition for PbMARL that enables learning the Nash equilibrium, which serves as a favorable policy for each player. Empirically, we validate our theoretical insights through comprehensive experiments utilizing the proposed algorithmic techniques.

1.1 CONTRIBUTIONS AND TECHNICAL NOVELTIES

1. Necessary and Sufficient Dataset Coverage Condition for PbMARL. In single-agent RLHF, (Zhu et al., 2023) demonstrated that single policy coverage is sufficient for learning the optimal policy. However, we prove that this condition no longer holds for PbMARL by providing a counterexample. Instead, we introduce an algorithm that operates under unilateral coverage, a condition derived from

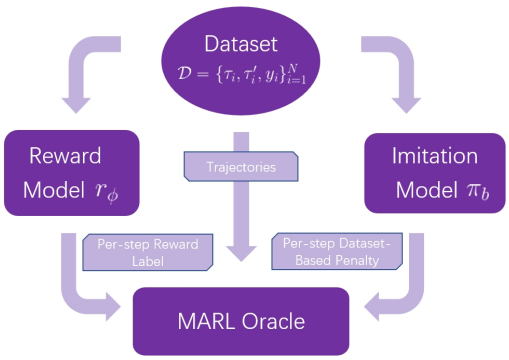


Figure 1: The overall pipeline of offline PbMARL. $\mathcal{D}$ is the preference dataset where τ_i, τ_i' are trajectories and $\mathbf{y}_i \in \{1, -1\}^m$ indicates which trajectory is preferred by each agent. r_ϕ is the learned reward. π_b is the learned reference policy using imitation learning.

offline MARL (Cui and Du, 2022a; Zhong et al., 2022). Specifically, this condition requires the dataset to cover all unilateral deviations from a Nash equilibrium policy. For further details, see Section 4.

2. Algorithmic Techniques for Practical Performance. As a foundational exploration into PbMARL research, we focus on employing the simplest learning framework, incorporating only the essential techniques necessary to ensure the approach's feasibility. The framework consists of three key components: 1) leveraging the preference dataset to learn a reward function, 2) mitigating extrapolation errors with pessimism, and 3) determining the final policy. Figure 1 provides an overview of the process.

However, additional algorithmic techniques are required to identify a robust policy, even when the dataset demonstrates good coverage according to our theoretical insights.

- **Reward regularization.** We observed that the reward learned through standard Maximum Likelihood Estimation (MLE) is sparse and spiky, making it difficult for standard RL algorithms to utilize effectively (cf. Figure 2 (b2)). To address this, we introduce an additional Mean Squared Error (MSE) loss between the predictions of adjacent time steps as a form of regularization. This regularization helps to prevent the model from accumulating reward signals solely at the final time step or relying on reward-irrelevant observation patterns, which could otherwise result in the complete failure in producing meaningful predictions.

- **Dataset Distribution-Based Pessimism.** To mitigate the extrapolation error in offline RL, we add an extra reward term based on the density of a certain state-action pair in the dataset to implement pessimism. In our approach, an imitation learning agent is trained to model the density function. The final policy is then trained using a DQN-based Value Decomposition Network (VDN) (Mnih et al., 2013; Sunehag et al., 2017). Our ablation study demonstrates the critical role of appropriately tuning the reward coefficient to ensure training stability and performance (see Table 4).

3. Experiment Results. Our experiments, following the pipeline described above, confirm the theoretical necessity of unilateral coverage. We conducted comprehensive ablation studies on three cooperative Multi-Agent Particle Environment (MPE) scenarios (Mordatch and Abbeel, 2017): Spread-v3, Tag-v3, and Reference-v3. These studies focused on the hyperparameter selection for the reward regularization coefficient α, pessimism coefficient β, and dataset diversity. The empirical results (Table 2) demonstrate that: 1) simply adding trivial trajectories to expert demonstrations can enhance performance, 2) unilateral datasets are advantageous, and 3) dataset diversity contributes to lower variance.

Our ablation experiments underscore the effectiveness of the proposed algorithmic techniques. Additionally, we introduced a principled standardization technique that can efficiently tune hyperparameters across all environments and datasets.

2 Related Works

Reinforcement Learning from Human Feedback (RLHF). RLHF, or preference-based RL (PbRL), plays a pivotal role in alignment with various tasks such as video games (Warnell et al., 2018; Brown et al., 2019), robotics (Jain et al., 2013; Kupcsik et al., 2016; Christiano et al., 2023; Shin et al., 2023), image augmentation (Metcalf et al., 2024), and large language models (Ziegler et al., 2020; Wu et al., 2021; Nakano et al., 2022; Menick et al., 2022; Stiennon et al., 2022; Bai et al., 2022; Glaese et al., 2022; Ganguli et al., 2022; Ouyang et al., 2022). Additionally, a body of work focuses on the reward models behind preference data (Sadigh et al., 2017; Bıyık and Sadigh, 2018; Gao et al., 2022; Hejna and Sadigh, 2023). Recent works like VIPO (Cen et al., 2024) incorporates uncertainty-aware regularization into the reward model, while (Liu et al., 2024) address over-optimization using adversarial regularization. Direct preference optimization (DPO, Rafailov et al. (2023)) and its variants (Azar et al., 2023; Rafailov et al., 2024) approach RLHF without directly handling the reward model. Theoretical studies have also explored guarantees, such as sample complexity and regret, and the limitations of certain RLHF algorithms (Novoseller et al., 2020; Xu et al., 2020; Pacchiano et al., 2023; Chen et al., 2022; Razin et al., 2023; Zhu et al., 2024a; Wang et al., 2023c; Xiong et al., 2024; Zhu et al., 2024b).

Offline Reinforcement Learning. Offline RL (Lange et al., 2012; Levine et al., 2020) has achieved success in a wide range of real-world applications, including robotics (Pinto and Gupta, 2015; Levine et al., 2016; Chebotar et al., 2021; Kumar et al., 2023), healthcare (Raghu et al., 2017; Wang et al., 2018), and autonomous driving (Shi et al., 2021; Lee et al., 2024). Key algorithms such as Behavior Cloning, BRAC (Wu et al., 2019), BEAR (Kumar et al., 2019), and CQL (Kumar et al., 2020; Lyu et al., 2024) have driven these successes. Theoretical research on offline RL has primarily focused on sample complexity under various dataset coverage assumptions Le et al. (2019); Chen and Jiang (2019); Yin et al. (2020); Rashidinejad et al. (2023); Yin et al. (2021; 2022); Shi et al. (2022); Nguyen-Tang et al. (2022); Xie et al. (2022); Xiong et al. (2023b); Li et al. (2024); Xie et al. (2023); Mete et al. (2021).

Multi-Agent Reinforcement Learning (MARL). Many real-world scenarios are naturally modeled as multi-agent environments, whether cooperative or competitive. As a result, MARL has gained popularity in video games (Tian et al., 2017; Vinyals et al., 2017; Silver et al., 2017; Vinyals et al., 2019), network design (Shamsoshoara et al., 2018; Kaur and Kumar, 2020), energy sharing (Prasad and Dusparic, 2018), and autonomous driving (Palanisamy, 2019; Yu et al., 2020; Zhou et al., 2022). Prominent algorithms in MARL include IQL (Tan, 2003), MADDPG (Lowe et al., 2020), COMA (Foerster et al., 2017), MAPPO (Yu et al., 2022), VDN (Sunehag et al., 2017), and QMIX (Rashid et al., 2018). Theoretical research has made great process in reducing the sample complexity(Wang et al., 2023b; Xiong et al., 2023a).

Offline MARL. Offline MARL is a practical solution for handling sophisticated multi-agent environments. Empirically, to address issues related to out-of-distribution actions and complex reward functions, previous works have developed algorithms such as MABCQ (Jiang and Lu, 2023), ICQ-MA (Yang et al., 2021), OMAR (Pan et al., 2022), and OMIGA (Wang et al., 2023a), which incorporate regularization or constraints on these actions and functions. MOMA-PPO (Barde et al., 2024) is a model-based approach to offline MARL that generates synthetic interaction data from offline datasets. Tseng et al. (2022) combines knowledge distillation with multi-agent decision transformers (Meng et al., 2022) for offline MARL. Theoretical understanding of offline MARL, particularly in the context of Markov games, has been advanced by works that provide sample complexity guarantees for learning equilibria Sidford et al. (2019); Cui and Yang (2020); Zhang et al. (2023a; 2020); Abe and Kaneko (2020); Cui and Du (2022a;b); Zhang et al. (2023b); Blanchet et al. (2023); Shi et al. (2023); Zhong et al. (2022).

3 PRELIMINARIES

General-sum Markov Games. We consider an episodic time-inhomogeneous general-sum Markov game $\mathcal{M}$, consisting of m players, a shared state space $\mathcal{S}$, an individual action space $\mathcal{A}_i$ for each player $i \in [m]$ and a joint action space $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \mathcal{A}_m$. The game has a time horizon H, an initial state s_1, state transition probabilities $\mathbb{P} = (\mathbb{P}_1, \mathbb{P}_2, \cdots, \mathbb{P}_H)$ with $\mathbb{P}_h : \mathcal{S}\mathcal{A} \to \Delta(\mathcal{S})$, and rewards $R = R_h(\cdot \mid s_h, a_h)_{h=1}^H$ where $R_{h,i} \in [0,1]$ represents the random reward for player i at step h. At each step $h \in [H]$, all players observe current state s_h and simultaneously choose their actions $\mathbf{a}_h = (a_{h,1}, a_{h,2}, \cdots, a_{h,m})$. The next state s_{h+1} is then sampled from $\mathbb{P}_h(\cdot \mid s_h, \mathbf{a}_h)$, and the reward $r_{h,i}$ for player i is sampled from $R_{h,i}(\cdot \mid s_h, \mathbf{a}_h)$. The game terminates at step $H+1$, with each player aiming to maximize the total collected rewards.

We use $\pi = (\pi_1, \pi_2, \cdots, \pi_m)$ to denote a joint policy, where the individual policy for player i is represented as $\pi_i = (\pi_{1,i}, \pi_{2,i}, \cdots, \pi_{H,i})$, with each $\pi_{h,i} : S \to \Delta(A_i)$ defined as the Markov policy for player i at step h. The state value function and state-action value function for each player $i \in [m]$ are defined as

$$V_{h,i}^{\pi}(s_h) := \mathbb{E}_{\pi}\left[\sum_{t=h}^{H} r_{t,i}(s_t, \mathbf{a}_t) \mid s_h\right], \; Q_{h,i}^{\pi}(s_h) := \mathbb{E}_{\pi}\left[\sum_{t=h}^{H} r_{t,i}(s_t, \mathbf{a}_t) \mid s_h, \mathbf{a}_h\right],$$

where $\mathbb{E}_{\pi} = \mathbb{E}_{s_1, \mathbf{a}_1, \mathbf{r}_1, \cdots, s_{H+1} \sim \pi, \mathcal{M}}$ denotes the expectation over the random trajectory generated by policy π. The best response value for player i is defined as

$$V_{h,i}^{\dagger, \pi_{-i}}(s_h) := \max_{\pi_i} V_{h,i}^{\pi_i, \pi_{-i}}(s_h),$$

which represents the maximal expected total return for player i given that the other players follow policy π_{-i}.

A Nash equilibrium is a policy configuration where no player has an incentive to change their policy unilaterally. Formally, we measure how closely a policy approximates a Nash equilibrium using the *Nash-Gap*:

$$\text{Nash-Gap}(\pi) := \sum_{i \in [m]} \left[V_{1,i}^{\dagger, \pi_{-i}}(s_1) - V_{1,i}^{\pi}(s_1)\right].$$

By definition, the Nash-Gap is always non-negative, and it quantifies the potential benefit each player could gain by unilaterally deviating from the current policy. A policy π is considered an ϵ-Nash equilibrium *iff* Nash-Gap$(\pi) \leq \epsilon$.

Offline Multi-agent Reinforcement Learning with Preference Feedback. In offline MARL with Preference Feedback, the algorithm has access to a pre-collected preference dataset generated by an unknown behavior policy interacting with an underlying Markov game. We consider two sampled trajectories, $\tau = (s_1, \mathbf{a}_1, s_2, \mathbf{a}_2, \cdots, s_{H+1})$ and $\tau' = (s_1', \mathbf{a}_1', s_2', \mathbf{a}_2', \cdots, s_{H+1}')$, drawn from distribution $\mathbb{P}(s_1, \mathbf{a}_1, s_2, \cdots, s_{H+1}) = \Pi_h \pi^b(\mathbf{a}_h \mid s_h)\mathbb{P}(s_{h+1} \mid s_h, \mathbf{a}_h)$ induced by the behavior policy π^b. In MARLHF, the reward signal is not revealed in the dataset. Instead, each player can observe a binary signal y_i from a Bernoulli distribution following the Bradley-Terry-Luce model (Bradley and Terry, 1952):

$$\mathbb{P}(y_i = 1 \mid \tau, \tau') = \frac{\exp(\sum_{h=1}^{H} r_i(s_h, \mathbf{a}_h))}{\exp(\sum_{h=1}^{H} r_i(s_h, \mathbf{a}_h)) + \exp(\sum_{h=1}^{H} r_i(s_h', \mathbf{a}_h'))}, \forall i \in [m].$$

We make the standard linear Markov game assumption (Zhong et al., 2022):

Assumption 1. *$\mathcal{M}$ is a linear Markov game with a feature map $\psi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ if we have*

$$\mathbb{P}_h(s_{h+1} \mid s_h, \mathbf{a}_h) = \langle \psi(s_h, \mathbf{a}_h), \mu_h(s_{h+1}) \rangle, \forall (s_h, \mathbf{a}_h, s_{h+1}, h) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times [H],$$

$$r_i(s_h, \mathbf{a}_h) = \langle \psi(s_h), \theta_{h,i} \rangle, \forall (s_h, \mathbf{a}_h, h, i) \in \mathcal{S} \times \mathcal{A} \times [H] \times [m],$$

where μ_h and $\theta_{h,i}$ are unknown parameters. Without loss of generality, we assume $\|\psi(s, \mathbf{a})\| \leq 1$ for all $(s, \mathbf{a}) \in \mathcal{S} \times \mathcal{A}$ and $\|\mu_h(s)\| \leq \sqrt{d}$, $\|\theta\|_h \leq \sqrt{d}$ for all $h \in [H]$.

The one-hot feature map is defined as $\overline{\psi}_h(s, \mathbf{a}) := [0, \cdots, 0, \psi(s, \mathbf{a}), 0, \cdots, 0] \in \mathbb{R}^{Hd}$, where $\psi(s, \mathbf{a})$ is at position $(h-1)d+1$ to hd.

Value-Decomposition Network (VDN). In our experiments, we utilize VDN as an offline MARL algorithm for its effectiveness and simplicity. VDN (Sunehag et al., 2017) is a Q-learning style MARL architecture for cooperative games. It takes the idea of decomposing the team value function into agent-wise value functions, expressed as: $Q_h(s, \mathbf{a}) = \sum_{i=1}^{n} Q_{h,i}(s, a_i)$. In our experiments, we applied Deep Q-Network (DQN) (Mnih et al., 2013) with VDN to learn the team Q function. We chose DQN to maintain the simplicity and controllability of the experimental pipeline, which facilitates a more accurate investigation of the impact of various techniques on the learning process.

4 Dataset Coverage Theory for MARLHF

In this section, we study the dataset coverage assumptions for offline MARLHF. For offline single-agent RLHF, Zhu et al. (2023); Zhan et al. (2023) show that single policy coverage is sufficient for learning the optimal policy. However, we prove that this assumption is insufficient in the multi-agent setting by constructing an counterexample. In addition, we prove that unilateral policy coverage is adequate for learning the Nash equilibrium.

4.1 Policy Coverages

We quantify the information contained in the dataset using covariance matrices, as the rewards and transition kernels are parameterized by a linear model. With a slight abuse of the notation, for trajectory $\tau = (s_1, \mathbf{a}_1, s_2, \mathbf{a}_2, \cdots, s_{H+1})$, we use $\psi(\tau) := [\psi(s_1, \mathbf{a}_1), \psi(s_2, \mathbf{a}_2), \cdots, \psi(s_H, \mathbf{a}_H)]$ to denote the concatenated trajectory feature. The reward coverage is measured by the preference covariance matrix:

$$\Sigma_{\mathcal{D}}^{r} = \lambda I + \sum_{(\tau, \tau') \in \mathcal{D}} (\psi(\tau) - \psi(\tau'))(\psi(\tau) - \psi(\tau'))^{\top},$$

where $\psi(\tau) - \psi(\tau')$ is derived from the preference model. Similarly, the transition coverage is measured by the covariance matrix:

$$\Sigma_{\mathcal{D},h}^{\mathbb{P}} = \lambda I + \sum_{(\tau, \tau') \in \mathcal{D}} \left[\psi(s_h, \mathbf{a}_h)\psi(s_h, \mathbf{a}_h)^{\top} + \psi(s_h', \mathbf{a}_h')\psi(s_h', \mathbf{a}_h')^{\top}\right].$$

For a given state and action pair $(s_h, \mathbf{a}_h)$, the term $\left\|\overline{\psi}_h(s_h, \mathbf{a}_h)\right\|_{[\Sigma_{\mathcal{D}}^{r}]^{-1}}$ measures the uncertainty in reward estimation and $\|\psi(s_h, \mathbf{a}_h)\|_{[\Sigma_{\mathcal{D},h}^{\mathbb{P}}]^{-1}}$ measures the uncertainty in transition estimation. As a result, the overall uncertainty of a given policy π with dataset $\mathcal{D}$ is measured by

$$U_{\mathcal{D}}(\pi) := \mathbb{E}_{\pi}\left[\sum_{h=1}^{H} \left\|\overline{\psi}_h(s_h, a_h)\right\|_{[\Sigma_{\mathcal{D}}^{r}]^{-1}} + \sum_{h=1}^{H} \|\psi(s_h, a_h)\|_{[\Sigma_{\mathcal{D},h}^{\mathbb{P}}]^{-1}}\right].$$

Definition 1. *For a Nash equilibrium π^*, different policy coverages are measured by the following quantities:*

- *Single policy coverage:* $U_{\mathcal{D}}(\pi^*)$.
- *Unilateral policy coverage:* $\max_{i, \pi_i} U_{\mathcal{D}}(\pi_i, \pi_{-i}^*)$.
- *Uniform policy coverage:* $\max_{\pi} U_{\mathcal{D}}(\pi)$.

Intuitively, small $U_{\mathcal{D}}(\pi^*)$ indicates that the dataset contains adequate information about π^*. A small $\max_{i, \pi_i} U_{\mathcal{D}}(\pi_i, \pi_{-i}^*)$ implies that the dataset covers all of the unilateral deviations of π^*, and small $\max_{\pi} U_{\mathcal{D}}(\pi^*)$ suggests that the dataset covers all possible policies.

4.2 Single Policy Coverage is Insufficient

Our objective is to learn a Nash equilibrium policy from the dataset, which necessitates that the dataset sufficiently covers the Nash equilibrium. In the single-agent scenario, if the dataset covers the

optimal policy, pessimism-based algorithms can be employed to recover the optimal policy. However, previous work (Cui and Du, 2022a; Zhong et al., 2022) has demonstrated that single policy coverage is insufficient for offline MARL. We extend this result to the context of offline MARL with preference feedback, as follows:

Theorem 1. *(Informal) If the dataset only has coverage on the Nash equilibrium policy (i.e. small $U_{\mathcal{D}}(\pi^*)$), it is not sufficient for learning an approximate Nash equilibrium policy.*

The proof is derived by a reduction from standard offline MARL to MARLHF. Suppose that MARLHF with single policy coverage suffices, we could construct an algorithm for standard offline MARL, which leads to a contradiction. The formal statement and the detailed proof are deferred to Appendix A.1.

4.3 Unilateral Policy Coverage is Sufficient

While single policy coverage is too weak to learn a Nash equilibrium, uniform policy coverage, though sufficient, is often too strong and impractical for many scenarios. Instead, we focus on unilateral policy coverage, which offers a middle ground between single policy coverage and uniform policy coverage.

Theorem 2. *(Informal) If the dataset has unilateral coverage on the Nash equilibrium policy, there exists an algorithm that can output an approximate Nash equilibrium policy.*

The detailed proof is deferred to Appendix A.2. We leverage a variant of Strategy-wise Bonus and Surrogate Minimization (SBSM) algorithm in (Cui and Du, 2022b) with modified policy evaluation and policy optimization subroutines. Intuitively, the algorithm identifies a policy that minimizes a pessimistic estimate of the Nash gap. As a result, if the dataset has unilateral coverage, the output policywill have a small Nash gap and serves as a good approximation of the Nash equilibrium.

5 Algorithmic Techniques for Practical Performance

In Section 4, we provided a theoretical characterization of the dataset requirements for MARLHF. However, the algorithm used in Theorem 2 is not computationally efficient. In this section, we propose a practical algorithm for MARLHF and validate our theoretical findings through experiments.

5.1 High-level Methodology

Our MARLHF pipeline consists of two phases: In the first step, we train a reward prediction model ϕ and approximate the behavior policy π_b using imitation learning; in the second step, we then apply an MARL algorithm to maximize a combination of the KL-divergence-based reward and standardized predicted reward r_ϕ, ultimately deriving the final policy $\pi_{\mathbf{w}}$.

Step 1: Reward Training and Dataset Modeling. Given the preference signals of trajectories, we use neural networks to predict step-wise rewards $r_\phi(s_h, a_h)$ for each agent, minimizing the loss defined in (1). The objective is to map (s, a_i)-pairs to reward values such that the team returns align with the preference signals. At the same time, in order to utilize distribution-based penalty term $\log \pi_b(s, a)$ to cope with the extrapolation error in offline learning, an imitation learner is trained over the entire dataset to model the behavior policy π_b.

Step 2: Offline MARL. Although in this work, VDN is chosen as the MARL oracle, it should be noted that other MARL architectures are also applicable. With the reward model r_ϕ and the approximated dataset distribution learned in Step 1, we are now able to construct a virtual step-wise reward for each agent. The agents are then trained to maximize the target defined in (3).

Given this framework, additional techniques are required to build a strong practical algorithm, which we provide more details below.

5.2 Reward Regularization

Compared to step-wise reward signals, preference signals are H times sparser, making them more challenging for a standard RL algorithm to utilize effectively. Concretely, this reward sparsity causes the naive optimization of the negative log-likelihood (NLL) loss to suffer from two key problems:

1. **Sparse and spiky reward output.** When calculating NLL losses, spreading the reward signal along the trajectories is equivalent to summing it at the last time step (Figure 2a). However, a sparse reward signal is harder for traditional RL methods to handdle due to the lack of continuous supervision. More uniformly distributed rewards across the entire trajectory generally leads to more efficient learning in standard RL algorithms.
2. **Over-reliance on irrelevant features.** The model may exploit redundant features as shortcuts to predict rewards. For instance, expert agents in cooperative games usually exhibit a fixed pattern of collaboration from the very beginning of the trajectory (such as specific actions or communication moves). The reward model might use these patterns to differentiate them from agents of other skill levels, thereby failing to capture the true reward-observation causal relationships.

To mitigate these problems, we introduce an extra Mean Squared Error (MSE) regularization along the time axis (Equation 1, 2). By limiting the sudden changes in reward predictions between adjacent time steps, this regularization discourages the reward model from concentrating its predictions on just a few time steps. While these issues can also be mitigated by using more diversified datasets and adding regularization to experts to eliminate reward-irrelevant action patterns, these approaches can be costly and sometimes impractical in real-world applications. In contrast, our MSE regularization is both easy to implement and has been empirically verified to be effective, creating more uniform reward distribution (Figure 2) and better performances.

$$L_{\text{RM}}(\phi) = -\mathbb{E}_{\mathcal{D}}\left[\sum_{i=1}^{m} \log \sigma(y_i(r_{\phi,i}(\tau_1) - r_{\phi,i}(\tau_2)))\right] + \frac{\alpha}{\text{Var}_{\mathcal{D}}(r_\phi)} L_{\text{MSE}}(\phi, \tau), \tag{1}$$

where the regularization term L_{MSE} is defined as:

$$L_{\text{MSE}}(\phi, \tau) = \mathbb{E}_{\mathcal{D}}\left[\sum_{h=1}^{H-1} \|r_\phi(s_h, \mathbf{a}_h) - r_\phi(s_{h+1}, \mathbf{a}_{h+1})\|_2^2\right]. \tag{2}$$

Here α is the regularization coefficient, which is set to be 1 in our experiments. The variance of r_ϕ is calculated over the training set to adaptively scale the regularization term. During training, $\text{Var}_{\mathcal{D}}(r_\phi)$ is detached to prevent gradients from flowing through it. The effectiveness of this method is validated in the ablation study (cf. Section 6.3).

5.3 Dataset Distribution-Based Pessimism

There are various methods to mitigate the over-extrapolation errors in offline RL (Peng et al., 2019; Nair et al., 2021), including conservative loss over the Q-function (Kumar et al., 2020) and directly restricting the learned policy actions to those within within the dataset (Fujimoto et al., 2019). We add a per-step dataset-based penalty term, $\log \pi_b(s, \mathbf{a})$, as pessimism towards less explored states. Imitation learning is utilized to estimate the behavior policy π_b from the dataset distribution. To stabilize training, we standardize predicted reward r_ϕ over $\mathcal{D}$ before combining it with the penalty term to make them comparable:

$$\text{objective}(\mathbf{w}) = \mathbb{E}_{\tau\sim\pi_{\mathbf{w}}}\left[\sum_{h=1}^{H} r_{\text{std}}(s_h, \mathbf{a}_h, \phi) + \text{clip}(\beta \log \pi_b(s_h, \mathbf{a}_h), -10, 1)\right], \tag{3}$$

where β is the pessimism coefficient, set to be $(1, 1, 10, 10)$ in Spread-v3, Reference-3, Tag-v3 and Overcooked respectively in the main experiments. The clip operator is defined by $\text{clip}(x, a, b) = \min(b, \max(a, x))$. The standardized reward r_{std} is defined as:

$$r_{\text{std}}(s_h, \mathbf{a}_h, \phi) = \sum_{i=1}^{m} \frac{r_\phi(s_h, a_{h,i}) - \mathbb{E}_{\mathcal{D}}(r_\phi)}{\sqrt{\text{Var}_{\mathcal{D}}(r_\phi)}}. \tag{4}$$

Intuitively, the penalty term $\log \pi_b(s_h, \mathbf{a}_h)$ discourages the agents from deviating from the most preferred actions in the dataset. The effectiveness of this method is validated in the ablation study (cf. Section 6.4).

6 EXPERIMENTS

We design a series of experiments to validate our theories and methods in common general-sum games. Specifically, we first use online RL algorithms to train expert agents, and take intermediate checkpoints as rookie agents. Then, we use these agents to collect datasets and use the Bradley-Terry model over standardized returns to simulate human preference. Experiments are carried out to verify the efficiency of our approach with unilateral policy dataset coverage (in Theorem 2) while single policy coverage is insufficient (stated in Theorem 1). We also design ablation studies to showcase the importance of our methods, particularly focusing on reward regularization and dataset distribution-based pessimism.

6.1 ENVIRONMENTS

Our experiments involved 3 Multi-Agent Particle Environments (MPE), including Spread-v3, Tag-v3 and Reference-v3, and Overcooked environment implemented with JaxMARL codebase (Rutherford et al., 2023). **Spread-v3** contains a group of agents and target landmarks, where the objective is to cover as many landmarks as possible while avoiding collisions. **Tag-v3** contains two opposing groups, where quicker "preys" need to escape from "predators". To ensure a fair comparison of different predator cooperation policies, we fixed a pretrained prey agent. **Reference-v3** involves two agents and three potential landmarks, where the agents need to find each one's target landmark to receive a high reward. The target landmark of each agent is only known by the other agent at first. **Overcooked** involves two agents moving and operating objects in a gridworld. A more detailed description of the tasks and their associated challenges is provided in Appendix B.2.

6.2 THE IMPORTANCE OF DATASET DIVERSITY

To study the influence of diversity of dataset, we manually designed 4 kinds of mixed joint behavior polices, and change their ratios to form different datasets.

- Expert policy: n expert agents. Trained with online RL algorithms till convergence.
- Rookie policy: n rookie agents. Trained with online RL algorithms with early stop.
- Trivial policy: n random agents. All actions are uniformly sampled from the action space.
- Unilateral policy: $n-1$ expert agents and 1 rookie agent of different proficiency level.

Table 1 presents the ratio of trajectories collected by the four different policies. The experiments are designed to hierarchically examine the roles of diversity (Diversified vs. Mix-Unilateral), unilateral coverage (Mix-Unilateral vs. Mix-Expert), and trivial comparison (Mix-Expert vs. Pure-Expert).

The ranking of diversity follows the order:

$$\text{Pure-Expert} < \text{Mix-Expert} < \text{Mix-Unilateral} < \text{Diversified}$$

Due to the inherent limitations of offline reinforcement learning (RL) in action selection dictated by the dataset, the effectiveness of learning is often strongly correlated with dataset quality, i.e. the level of expertise demonstrated in the dataset. However, the results in preference-based MARL experiments partially diverge from this conventional conclusion. While the quality of the dataset remains critical, experiments on Reference-v3 and Overcooked (Table 2) indicate that diversity and unilateral data can significantly enhance the performance of the reward model, thereby facilitating learning.

The main experimental results are presented in Table 2 and Table 3.Among all the experiments, apart from the experiments on Tag-v3, where the high operational precision requirements make data quality more critical than diversity, the other three environments validate our conclusions across all algorithms.

	Expert	Unilateral	Rookie	Trivial
Diversified	1	1	1	1
Mix-Unilateral	2	1	0	1
Mix-Expert	3	0	0	1
Pure-Expert	4	0	0	0

Table 1: Final datasets mixed with various ratios. The overall dataset size is kept to 38400 trajectories for MPE, and 960 trajectories for Overcooked. (cf. B.1)

Algorithm	Dataset	Spread-v3	Tag-v3	Reference-v3	Overcooked
VDN with Pessimism Penalty	Diversified	-21.16 $\pm$ 0.54	29.28 $\pm$ 1.08	-18.89 $\pm$ 0.60	**238.89** $\pm$ **3.50**
	Mix-Unilateral	-21.03 $\pm$ 0.44	36.65 $\pm$ 0.70	-18.80 $\pm$ 0.63	221.80 $\pm$ 26.66
	Mix-Expert	-20.98 $\pm$ 0.54	35.96 $\pm$ 0.86	-18.80 $\pm$ 0.44	35.26 $\pm$ 55.19
	Pure-Expert	-21.01 $\pm$ 0.57	**39.55** $\pm$ **0.77**	-28.97 $\pm$ 2.89	3.36 $\pm$ 7.19

Table 2: In the simplest environment, Spread-v3, different dataset gives similar performance. In Tag-v3 environment, where precise actions are required, the quality of the dataset (proportion of expert demonstration) is more important than diversity. In contrast, in Overcooked environment, which focuses on strategy learning and demands less on precision, dataset diversity contributes to improved stability, with Unilateral playing a particularly critical role. In the Reference-v3 environment, which balances the need for precision and strategic, the importance of both factors is more balanced, but non-expert data is still necessary.

6.3 Experiments for Reward Regularization

In Figure 2, we examined the effectiveness of our proposed reward regularization technique. Figure 2a demonstrates that without regularization, the learned rewards tend to be sparse and spiky compared to the ground truth rewards.

We also observe that the rewards often exhibit temporal continuity, which can create greater discrepancies with the sparse, pulse-like ground truth. Notably, we found that adding stronger regularization does not necessarily lead to underfitting of the reward model; in some cases, it even helps the model converge to a lower training loss. Detailed parameters and experimental results are provided in the appendix (cf. Table 8). We attribute this to the role of regularization in preventing the model from overly relying on shortcuts.

6.4 Other Ablation Studies

Pessimism coefficient Due to the clipping in 3, excessively large β values will not dominate the entire reward function. As a result, larger β values almost never degrade the agent's performance in our experiments (Table 4). This allows us to increase β with relative confidence. Therefore, we generally recommend setting β to a value between 10 and 100 for optimal performances.

Scalability We also tested the scalability on Spread-v3. While our current approach manages the scaling of agents without introducing new problems, it does not specifically address the inherent issues of instability and complexity that are well-documented in traditional MARL (cf. Appendix B.4).

7 Discussion

In this paper, we proposed dedicated algorithmic techniques for offline PbMARL and provided theoretical justification for the unilateral dataset coverage condition. We believe our work is a significant step towards systematically studying PbMARL and offers a foundational framework for future research in this area. The flexibility of our framework allows for application across a wide

Algorithm	Dataset	Spread-v3	Reference-v3	Overcooked
MAIQL	Diversified	-25.33 ± 1.40	-22.15 ± 0.55	**16.59** ± 11.22
	Mix-Unilateral	-23.25 ± 1.06	-23.22 ± 1.37	0.00 ± 0.00
	Mix-Expert	-23.26 ± 0.90	-24.21 ± 1.60	0.00 ± 0.00
	Pure-Expert	-26.01 ± 1.53	-29.47 ± 1.65	0.00 ± 0.00
MABCQ	Diversified	-20.02 ± 0.64	-17.64 ± 0.43	**239.34** ± 1.67
	Mix-Unilateral	-19.47 ± 0.33	-17.64 ± 1.11	215.01 ± 65.43
	Mix-Expert	-19.42 ± 0.17	-17.88 ± 0.78	50.32 ± 82.82
	Pure-Expert	-20.56 ± 0.38	-25.90 ± 1.11	1.14 ± 3.46

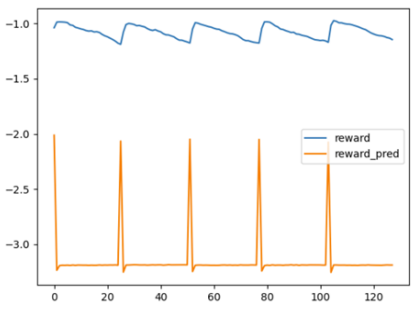
Table 3: Test returns of MAIQL and MABCQ. In the experimental results, we can observe a clear preference toward more diversified datasets. Compared to our method and BCQ, which directly calculate $\max_a Q$ for Bellman updates, IQL employs expectile regression to estimate it. So MAIQL demands higher accuracy of the reward model. Consequently, the performance improvements brought by dataset diversity are also more pronounced in MAIQL experiments.

	$\beta = 0$	$\beta = 0.1$	$\beta = 1$	$\beta = 10$	$\beta = 100$	$\alpha = 0$
Spread-v3	-22.56 ± 1.61	-22.03 ± 0.67	-20.82 ± 0.53	-20.46 ± 0.51	-20.35 ± 0.43	-22.21 ± 0.72
Tag-v3	4.11 ± 1.66	4.25 ± 0.53	10.96 ± 1.20	28.88 ± 1.02	29.53 ± 1.35	30.77 ± 0.57
Reference-v3	-19.69 ± 0.36	-19.37 ± 0.53	-18.89 ± 0.78	-18.33 ± 0.42	-18.54 ± 0.46	-21.86 ± 0.73
Overcooked	0.00 ± 0.00	0.00 ± 0.00	149.53 ± 86.74	238.89 ± 3.50	**240** ± 0.00	**240** ± 0.00

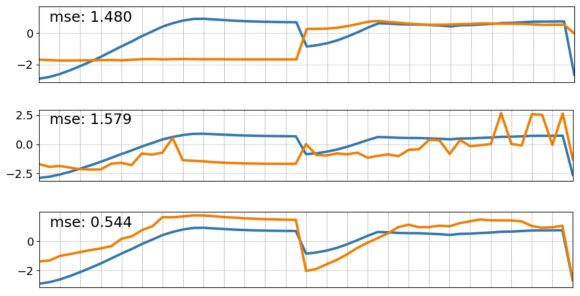
Table 4: Comparison of test return with different hyperparameters. Standard pipeline take pessimism coefficient $\beta = 1$ for Spread-v3, Reference-v3 and $\beta = 10$ for Tag-v3, Overcooked, and the MSE reward regularization coefficient α is set to the optimal value for fixed β. All the agents are trained on Diversified Dataset across 10 random seeds. Results show that larger β always gives better performance and a proper positive α can improve performance.

range of general games, and our empirical results validate the effectiveness of our proposed methods in various scenarios.

Looking ahead, there is significant potential to extend this work to more complex, real-world scenarios, particularly by integrating Large Language Models (LLMs) into multi-agent systems. Future research will focus on fine-tuning and aligning LLMs within PbMARL, addressing challenges such as increased complexity and the design of effective reward structures.



(a) Mix-Expert $\alpha = 0$ (spread-v3)



(b) reference-v3

(b1) Pure-Expert (b2) Diversified $\alpha = 0$ (b3) Diversified

Figure 2: (a) Averaged reward predictions and ground truth of a trajectory sample on spread-v3. (b) Standardized reward predictions and ground truth of a trajectory sample in reference-v3. When trained with expert data only (b1), ϕ experiences a mode collapse, failing to give informative signals. Reward function trained without regularization (b2) shows spiky patterns and tends to accumulate predictions at certain time steps when trained with less diversified datasets as (a). Our method with diversified dataset (b3) gives predictions that approximate the ground truth well.

8 Reproducibility Statement

All code used for our experiments is included in the supplementary material (`codebase.zip`). Appendix A provides detailed proofs of the theoretical bounds, along with necessary assumptions. Key experimental details and hyperparameters are also outlined in Appendix B. We believe these resources provide a comprehensive foundation for reproducing both the theoretical and empirical results presented in this work.

References

Kenshi Abe and Yusuke Kaneko. Off-policy exploitability-evaluation in two-player zero-sum markov games, 2020.

Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. A general theoretical paradigm to understand learning from human preferences, 2023.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.

Paul Barde, Jakob Foerster, Derek Nowrouzezahrai, and Amy Zhang. A model-based solution to the offline multi-agent reinforcement learning coordination problem, 2024.

José H. Blanchet, Miao Lu, Tong Zhang, and Han Zhong. Double pessimism is provably efficient for distributionally robust offline reinforcement learning: Generic algorithm and robust partial coverage. *ArXiv*, abs/2305.09659, 2023. URL https://api.semanticscholar.org/CorpusID:258714763.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Daniel S. Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations, 2019.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Erdem Bıyık and Dorsa Sadigh. Batch active preference-based learning of reward functions, 2018.

Shicong Cen, Jincheng Mei, Katayoon Goshvadi, Hanjun Dai, Tong Yang, Sherry Yang, Dale Schuurmans, Yuejie Chi, and Bo Dai. Value-incentivized preference optimization: A unified approach to online and offline rlhf, 2024. URL https://arxiv.org/abs/2405.19320.

Yevgen Chebotar, Karol Hausman, Yao Lu, Ted Xiao, Dmitry Kalashnikov, Jake Varley, Alex Irpan, Benjamin Eysenbach, Ryan Julian, Chelsea Finn, and Sergey Levine. Actionable models: Unsupervised offline reinforcement learning of robotic skills, 2021.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning, 2019.

Xiaoyu Chen, Han Zhong, Zhuoran Yang, Zhaoran Wang, and Liwei Wang. Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation, 2022.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Qiwen Cui and Simon S Du. When are offline two-player zero-sum markov games solvable? *Advances in Neural Information Processing Systems*, 35:25779–25791, 2022a.

Qiwen Cui and Simon S Du. Provably efficient offline multi-agent reinforcement learning via strategy-wise bonus. *Advances in Neural Information Processing Systems*, 35:11739–11751, 2022b.

Qiwen Cui and Lin F. Yang. Minimax sample complexity for turn-based stochastic game, 2020.

Sam Devlin, Daniel Kudenko, and Marek Grześ. An empirical study of potential-based reward shaping and advice in complex, multi-agent systems. *Advances in Complex Systems*, 14(02): 251–278, 2011.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients, 2017.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration, 2019.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned, 2022.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization, 2022.

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Uesato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Iason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. Improving alignment of dialogue agents via targeted human judgements, 2022.

Joey Hejna and Dorsa Sadigh. Inverse preference learning: Preference-based rl without a reward function, 2023.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 2023.

Ashesh Jain, Brian Wojcik, Thorsten Joachims, and Ashutosh Saxena. Learning trajectory preferences for manipulators via iterative improvement, 2013.

Jiechuan Jiang and Zongqing Lu. Offline decentralized multi-agent reinforcement learning, 2023.

Amandeep Kaur and Krishan Kumar. Energy-efficient resource allocation in cognitive radio networks under cooperative multi-agent model-free reinforcement learning schemes. *IEEE Transactions on Network and Service Management*, 17(3):1337–1348, 2020. doi: 10.1109/TNSM.2020.3000274.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning, 2021. URL https://arxiv.org/abs/2110.06169.

Aviral Kumar, Justin Fu, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction, 2019.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning, 2020.

Aviral Kumar, Anikait Singh, Frederik Ebert, Mitsuhiko Nakamoto, Yanlai Yang, Chelsea Finn, and Sergey Levine. Pre-training for robots: Offline rl enables learning new tasks from a handful of trials, 2023.

Andras Kupcsik, David Hsu, and Wee Sun Lee. Learning dynamic robot-to-human object handover from human feedback, 2016.

Sascha Lange, Thomas Gabel, and Martin Riedmiller. *Batch Reinforcement Learning*, pages 45–73. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-27645-3. doi: 10.1007/978-3-642-27645-3_2. URL https://doi.org/10.1007/978-3-642-27645-3_2.

Hoang M. Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints, 2019.

Dongsu Lee, Chanin Eom, and Minhae Kwon. Ad4rl: Autonomous driving benchmarks for offline reinforcement learning with value-based dataset, 2024.

Sergey Levine, Peter Pastor, Alex Krizhevsky, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection, 2016.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems, 2020.

Gen Li, Laixi Shi, Yuxin Chen, Yuejie Chi, and Yuting Wei. Settling the sample complexity of model-based offline reinforcement learning, 2024.

Zhihan Liu, Miao Lu, Shenao Zhang, Boyi Liu, Hongyi Guo, Yingxiang Yang, Jose Blanchet, and Zhaoran Wang. Provably mitigating overoptimization in rlhf: Your sft loss is implicitly an adversarial regularizer, 2024. URL https://arxiv.org/abs/2405.16436.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments, 2020.

Jiafei Lyu, Xiaoteng Ma, Xiu Li, and Zongqing Lu. Mildly conservative q-learning for offline reinforcement learning, 2024.

Linghui Meng, Muning Wen, Yaodong Yang, Chenyang Le, Xiyun Li, Weinan Zhang, Ying Wen, Haifeng Zhang, Jun Wang, and Bo Xu. Offline pre-trained multi-agent decision transformer: One big sequence model tackles all smac tasks, 2022.

Jacob Menick, Maja Trebacz, Vladimir Mikulik, John Aslanides, Francis Song, Martin Chadwick, Mia Glaese, Susannah Young, Lucy Campbell-Gillingham, Geoffrey Irving, and Nat McAleese. Teaching language models to support answers with verified quotes, 2022.

Katherine Metcalf, Miguel Sarabia, Natalie Mackraz, and Barry-John Theobald. Sample-efficient preference-based reinforcement learning with dynamics aware rewards, 2024.

Akshay Mete, Rahul Singh, Xi Liu, and P. R. Kumar. Reward biased maximum likelihood estimation for reinforcement learning, 2021. URL https://arxiv.org/abs/2011.07738.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning, 2013.

Igor Mordatch and Pieter Abbeel. Emergence of grounded compositional language in multi-agent populations. *arXiv preprint arXiv:1703.04908*, 2017.

Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. Awac: Accelerating online reinforcement learning with offline datasets, 2021.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. Webgpt: Browser-assisted question-answering with human feedback, 2022.

Thanh Nguyen-Tang, Sunil Gupta, Hung Tran-The, and Svetha Venkatesh. Sample complexity of offline reinforcement learning with deep relu networks, 2022.

Ellen R. Novoseller, Yibing Wei, Yanan Sui, Yisong Yue, and Joel W. Burdick. Dueling posterior sampling for preference-based reinforcement learning, 2020.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

Aldo Pacchiano, Aadirupa Saha, and Jonathan Lee. Dueling rl: Reinforcement learning with trajectory preferences, 2023.

Praveen Palanisamy. Multi-agent connected autonomous driving using deep reinforcement learning, 2019.

Ling Pan, Longbo Huang, Tengyu Ma, and Huazhe Xu. Plan better amid conservatism: Offline multi-agent reinforcement learning with actor rectification, 2022.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning, 2019.

Lerrel Pinto and Abhinav Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours, 2015.

Amit Prasad and Ivana Dusparic. Multi-agent deep reinforcement learning for zero energy communities. *2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe)*, pages 1–5, 2018. URL https://api.semanticscholar.org/CorpusID:52948132.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2023.

Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. From r to q^*: Your language model is secretly a q-function, 2024.

Aniruddh Raghu, Matthieu Komorowski, Imran Ahmed, Leo Celi, Peter Szolovits, and Marzyeh Ghassemi. Deep reinforcement learning for sepsis treatment, 2017.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning, 2018.

Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism, 2023.

Noam Razin, Hattie Zhou, Omid Saremi, Vimal Thilak, Arwen Bradley, Preetum Nakkiran, Joshua Susskind, and Etai Littwin. Vanishing gradients in reinforcement finetuning of language models, 2023.

Alexander Rutherford, Benjamin Ellis, Matteo Gallici, Jonathan Cook, Andrei Lupu, Gardar Ingvarsson, Timon Willi, Akbir Khan, Christian Schroeder de Witt, Alexandra Souly, Saptarashmi Bandyopadhyay, Mikayel Samvelyan, Minqi Jiang, Robert Tjarko Lange, Shimon Whiteson, Bruno Lacerda, Nick Hawes, Tim Rocktaschel, Chris Lu, and Jakob Nicolaus Foerster. Jaxmarl: Multi-agent rl environments in jax. 2023.

Dorsa Sadigh, Anca D. Dragan, S. Shankar Sastry, and Sanjit A. Seshia. Active preference-based learning of reward functions. In *Robotics: Science and Systems*, 2017. URL https://api.semanticscholar.org/CorpusID:12226563.

Alireza Shamsoshoara, Mehrdad Khaledi, Fatemeh Afghah, Abolfazl Razi, and Jonathan Ashdown. Distributed cooperative spectrum sharing in uav networks using multi-agent reinforcement learning, 2018.

Chengshuai Shi, Wei Xiong, Cong Shen, and Jing Yang. Provably efficient offline reinforcement learning with perturbed data sources. *ArXiv*, abs/2306.08364, 2023. URL https://api.semanticscholar.org/CorpusID:259165155.

Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity, 2022.

Tianyu Shi, Dong Chen, Kaian Chen, and Zhaojian Li. Offline reinforcement learning for autonomous driving with safety and exploration enhancement, 2021.

Daniel Shin, Anca D. Dragan, and Daniel S. Brown. Benchmarks and algorithms for offline preference-based reward learning, 2023.

Aaron Sidford, Mengdi Wang, Lin F. Yang, and Yinyu Ye. Solving discounted stochastic two-player games with near-optimal time and sample complexity, 2019.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy P. Lillicrap, Fan Hui, L. Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nature*, 550:354–359, 2017. URL https://api.semanticscholar.org/CorpusID:205261034.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, 2022.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. Value-decomposition networks for cooperative multi-agent learning, 2017.

Ming Tan. Multi agent reinforcement learning independent vs cooperative agents. 2003. URL https://api.semanticscholar.org/CorpusID:260435822.

Yuandong Tian, Qucheng Gong, Wenling Shang, Yuxin Wu, and C. Lawrence Zitnick. Elf: An extensive, lightweight and flexible research platform for real-time strategy games, 2017.

Wei-Cheng Tseng, Tsun-Hsuan Johnson Wang, Yen-Chen Lin, and Phillip Isola. Offline multi-agent reinforcement learning with knowledge distillation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 226–237. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/01d78b294d80491fecddea897cf03642-Paper-Conference.pdf.

Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John Agapiou, Julian Schrittwieser, John Quan, Stephen Gaffney, Stig Petersen, Karen Simonyan, Tom Schaul, Hado van Hasselt, David Silver, Timothy Lillicrap, Kevin Calderone, Paul Keet, Anthony Brunasso, David Lawrence, Anders Ekermo, Jacob Repp, and Rodney Tsing. Starcraft ii: A new challenge for reinforcement learning, 2017.

Oriol Vinyals, Igor Babuschkin, Wojciech M. Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, L. Sifre, Trevor Cai, John P. Agapiou, Max Jaderberg, Alexander Sasha Vezhnevets, Rémi Leblond, Tobias Pohlen, Valentin Dalibard, David Budden, Yury Sulsky, James Molloy, Tom Le Paine, Caglar Gulcehre, Ziyun Wang, Tobias Pfaff, Yuhuai Wu, Roman Ring, Dani Yogatama, Dario Wünsch, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy P. Lillicrap, Koray Kavukcuoglu, Demis Hassabis, Chris Apps, and David Silver. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575:350 – 354, 2019. URL https://api.semanticscholar.org/CorpusID:204972004.

Lu Wang, Wei Zhang, Xiaofeng He, and Hongyuan Zha. Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation, 2018.

Xiangsen Wang, Haoran Xu, Yinan Zheng, and Xianyuan Zhan. Offline multi-agent reinforcement learning with implicit global-to-local value regularization, 2023a.

Yuanhao Wang, Qinghua Liu, Yu Bai, and Chi Jin. Breaking the curse of multiagency: Provably efficient decentralized multi-agent rl with function approximation. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 2793–2848. PMLR, 12–15 Jul 2023b. URL https://proceedings.mlr.press/v195/wang23b.html.

Yuanhao Wang, Qinghua Liu, and Chi Jin. Is rlhf more difficult than standard rl?, 2023c.

Garrett Warnell, Nicholas Waytowich, Vernon Lawhern, and Peter Stone. Deep tamer: Interactive agent shaping in high-dimensional state spaces, 2018.

Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. Recursively summarizing books with human feedback, 2021.

Yifan Wu, George Tucker, and Ofir Nachum. Behavior regularized offline reinforcement learning, 2019.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023.

Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning, 2022.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning, 2023. URL https://arxiv.org/abs/2106.06926.

Nuoya Xiong, Zhihan Liu, Zhaoran Wang, and Zhuoran Yang. Sample-efficient multi-agent rl: An optimization perspective, 2023a. URL https://arxiv.org/abs/2310.06243.

Wei Xiong, Han Zhong, Chengshuai Shi, Cong Shen, Liwei Wang, and Tong Zhang. Nearly minimax optimal offline reinforcement learning with linear function approximation: Single-agent mdp and markov game, 2023b.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint, 2024.

Yichong Xu, Ruosong Wang, Lin F. Yang, Aarti Singh, and Artur Dubrawski. Preference-based reinforcement learning with finite-time guarantees, 2020.

Yiqin Yang, Xiaoteng Ma, Chenghao Li, Zewu Zheng, Qiyuan Zhang, Gao Huang, Jun Yang, and Qianchuan Zhao. Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning, 2021.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning, 2020.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal offline reinforcement learning via double variance reduction, 2021.

Ming Yin, Yaqi Duan, Mengdi Wang, and Yu-Xiang Wang. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism, 2022.

Chao Yu, Xin Wang, Xin Xu, Minjie Zhang, Hongwei Ge, Jiankang Ren, Liang Sun, Bingcai Chen, and Guozhen Tan. Distributed multiagent coordinated learning for autonomous driving in highways based on dynamic coordination graphs. *IEEE Transactions on Intelligent Transportation Systems*, 21(2):735–748, 2020. doi: 10.1109/TITS.2019.2893683.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative, multi-agent games, 2022.

Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason D Lee, and Wen Sun. Provable offline preference-based reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2023.

Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Basar. Finite-sample analysis for decentralized batch multi-agent reinforcement learning with networked agents, 2020.

Kaiqing Zhang, Sham M. Kakade, Tamer Basar, and Lin F. Yang. Model-based multi-agent rl in zero-sum markov games with near-optimal sample complexity, 2023a.

Yuheng Zhang, Yunru Bai, and Nan Jiang. Offline learning in markov games with general function approximation. In *International Conference on Machine Learning*, 2023b. URL https://api.semanticscholar.org/CorpusID:256615864.

Han Zhong, Wei Xiong, Jiyuan Tan, Liwei Wang, Tong Zhang, Zhaoran Wang, and Zhuoran Yang. Pessimistic minimax value iteration: Provably efficient equilibrium learning from offline datasets. In *International Conference on Machine Learning*, pages 27117–27142. PMLR, 2022.

Wei Zhou, Dong Chen, Jun Yan, Zhaojian Li, Huilin Yin, and Wanchen Ge. Multi-agent reinforcement learning for cooperative lane changing of connected and autonomous vehicles in mixed traffic. *Autonomous Intelligent Systems*, 2(1), March 2022. ISSN 2730-616X. doi: 10.1007/s43684-022-00023-5. URL http://dx.doi.org/10.1007/s43684-022-00023-5.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, pages 43037–43067. PMLR, 2023.

Banghua Zhu, Jiantao Jiao, and Michael I. Jordan. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons, 2024a.

Banghua Zhu, Michael I. Jordan, and Jiantao Jiao. Iterative data smoothing: Mitigating reward overfitting and overoptimization in rlhf, 2024b.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020.