Scalable Offline Reinforcement Learning for Mean Field Games

Axel Brunnbauer
TU Wien
Vienna, Austria
DatenVorsprung GmbH
Vienna, Austria
axel.brunnbauer@tuwien.ac.at

Julian Lemmel
TU Wien
Vienna, Austria
DatenVorsprung GmbH
Vienna, Austria
julian.lemmel@tuwien.ac.at

Zahra Babaiee
TU Wien
Vienna, Austria
DatenVorsprung GmbH
Vienna, Austria
zahra.babaiee@tuwien.ac.at

Sophie A. Neubauer
DatenVorsprung GmbH
Vienna, Austria
sophie@datenvorsprung.at

Radu Grosu
TU Wien
Vienna, Austria
radu.grosu@tuwien.ac.at

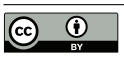
ABSTRACT

Reinforcement learning (RL) algorithms for mean-field games offer a scalable framework for optimizing policies in large populations of interacting agents. Existing methods often depend on online interactions or assume access to system dynamics, limiting their practicality in real-world scenarios where such interactions are infeasible or difficult to model. In this paper, we present Offline Munchausen Mirror Descent (Off-MMD), a novel mean-field RL algorithm that approximates equilibrium policies in mean-field games using purely offline data. By leveraging iterative mirror descent and importance sampling, Off-MMD estimates the mean-field distribution from static datasets without relying on simulation or environment dynamics. Additionally, we incorporate techniques from offline RL to address common issues like Q-value overestimation, ensuring robust policy learning even with limited data coverage. Our algorithm scales to complex environments and demonstrates strong performance on benchmark tasks like crowd exploration or navigation, highlighting its applicability to real-world multi-agent systems where online experimentation is infeasible. We empirically demonstrate the robustness of Off-MMD to low-quality datasets and conduct experiments to investigate its sensitivity to hyperparameter choices.

KEYWORDS

Mean-Field Games; Deep Reinforcement Learning; Offline Reinforcement Learning

ACM Reference Format:
Axel Brunnbauer, Julian Lemmel, Zahra Babaiee, Sophie A. Neubauer, and Radu Grosu. 2025. Scalable Offline Reinforcement Learning for Mean Field Games. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 10 pages.



Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

1 INTRODUCTION

Reinforcement Learning (RL) has emerged as a foundational tool for solving sequential decision-making problems across a diverse range of domains, including robotics, healthcare, autonomous systems, and game theory. However, while RL techniques have seen significant success in single-agent settings, the transition to multi-agent reinforcement learning (MARL) presents unique challenges, such as the exponential growth in the state and action spaces as the number of interacting agents increases, making the problem significantly more complex in terms of computation.

Traditional MARL methods often do not scale to many-agent settings and rapidly become infeasible in environments with large populations of agents. As the number of agents grows, learning effective strategies can become computationally prohibitive. To address this challenge, Mean-Field Games (MFGs), introduced by Lasry and Lions [23] and Huang et al. [17], provide a scalable approximation for large N-player games. The core idea is to represent the collective state of the population as a distribution over individual agent states, known as the *mean-field*. This reduces the decision-making problem to interactions between a single representative agent and the mean-field, enabling more efficient analysis and computation. This approach has shown great promise in reducing the complexity of multi-agent interactions, allowing for the development of more tractable solutions in environments with many agents. Early works on MFGs tackled problems of relatively small scale, often under restrictive assumptions such as linear dynamics and quadratic cost functions. These simplified models, while mathematically elegant, limit the applicability of MFGs to real-world systems that exhibit complex, nonlinear behavior. However, recent advances in the field have focused on scaling MFG solutions by leveraging deep reinforcement learning (DRL) techniques. These methods use neural network function approximators to compute equilibrium policies in MFGs, as demonstrated in [24, 31, 32]. Such approaches have enabled significant progress in applying MFG theory to more practical and large-scale environments.

Despite these advances, a critical issue remains: most existing MFG-based RL algorithms rely on online interaction with the environment. In many real-world applications, particularly those involving large populations of agents or human interactions (e.g., traffic routing, crowd dynamics, or recommendation systems), online interaction is either impractical or ethically unjustifiable. For

example, in systems with human agents, it is often costly or intrusive to collect real-time data, and continuous exploration could lead to unintended consequences such as user dissatisfaction or safety risks. Furthermore, environments with many agents can be difficult to model accurately, and real-time experimentation in such systems may not be possible. In single-agent RL, offline learning is a well researched area and allows to solve this problem by learning policies from pre-collected, fixed datasets, eliminating the need for online interactions. These offline methods have proven highly effective in settings where real-time interaction is limited or expensive. However, the application of offline RL techniques to MFGs remains underexplored. Current MFG methods have largely overlooked the offline setting, where no online interaction with the environment is available during learning.

To bridge this gap, we propose **Offline Munchausen Mirror Descent (Off-MMD)**, an offline mean-field RL algorithm. Off-MMD extends the recently introduced Deep Munchausen Online Mirror Descent (D-MOMD) method by Lauriere et al. [24] and combines the scalability of MFG approximations with the data efficiency of offline RL. To the best of our knowledge, Off-MMD is the first deep offline RL algorithm specifically designed for MFGs, capable of handling arbitrarily sized datasets. This innovation opens the door for applying MFG theory in real-world systems where online data collection is prohibitive. The primary challenge in adapting MFGs to the offline setting stems from the fact that these systems typically require estimating the distribution of agents (the mean-field), which complicates the direct adaptation of online algorithms. To address this, we repurpose ideas from off-policy policy evaluation (OPE) to approximate the mean-field distribution using offline data. Additionally, we apply a robust regularization mechanism to mitigate the effects of distributional shift, a well-known issue in offline RL. This regularization stabilizes the policy learning process, ensuring more reliable performance even in underrepresented areas of the state space. We empirically validate Off-MMD on a suite of benchmark tasks to demonstrate its efficacy. Specifically, we assess its performance on tasks that involve both fully cooperative and non-cooperative behaviors across two common benchmarks in the field of MFGs. Additionally, we explore the sensitivity of Off-MMD to the quality of the datasets used for training, assessing how variations in state-action coverage and trajectory quality impact performance. We further investigate the importance of the proposed regularization term, designed to prevent the overestimation of Q-values. In summary, our contributions are:

(1) Offline Munchausen Mirror Descent, a novel deep RL algorithm for offline learning in MFGs.
(2) Extensive evaluation of the performance and ablation studies with respect to regularization and dataset quality.

2 BACKGROUND

Sequential decision making problems commonly make use of finite horizon Markov decision processes (MDP). A finite horizon MDP is a tuple $\langle \mathcal{S}, \mathcal{A}, r, p, \gamma, H, \mu_0 \rangle$ consisting of a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, a reward function $r : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$, stochastic dynamics $p : \mathcal{S} \times \mathcal{A} \mapsto \Delta_{\mathcal{S}}$, a discount factor $\gamma \in (0, 1)$, a horizon $H \in \mathbb{N}$ and an initial state distribution $\mu_0 \in \Delta_{\mathcal{S}}$. Problems involving a large number of interacting agents become intractable as the size of the state and action space grows exponentially with the number of agents. In many-agent games, where agents are anonymous and identical, MFGs offer an effective framework to approximate the population dynamics. First introduced by Lasry and Lions [23] and Huang et al. [17], MFGs address this complexity by modeling interactions through the distribution of agent states, rather than tracking individual agents. This reduces the dimensionality of the problem, making it more tractable and allows to approximate finite, N-player games. In MFGs, a representative agent interacts with the mean-field rather than directly interacting with each individual agent. Consequently, the problem becomes optimizing a single policy with respect to this population distribution.

In this work, we consider *stochastic, finite-horizon* MFGs with a finite set of states $\mathcal{S}$ and actions $\mathcal{A}$. The *mean-field*, which is the distribution of agent states at time t, is denoted by $\mu_t \in \Delta_{\mathcal{S}}$. In the most general form, MFGs allow for mean-field-dependent dynamics, rewards, and even policies. However, in this work, we focus on the case where only the rewards depend on the mean-field:

$$r_t : \mathcal{S} \times \mathcal{A} \times \Delta_{\mathcal{S}} \mapsto \mathbb{R}. \tag{1}$$

This allows an individual agent to incorporate the behavior of a large number of other agents into its decision-making process. Given a population policy $\pi : \mathcal{S} \mapsto \Delta_{\mathcal{A}}$, the mean-field flow μ^π is defined by the recursive relation

$$\mu^\pi_{t+1}(s') = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} p(s'|s, a)\pi(a|s)\mu^\pi_t(s), \tag{2}$$

$$\mu^\pi_0 = \mu_0, \tag{3}$$

where $p(s'|s, a)$ represents the transition dynamics of the environment, and μ_0 is the initial distribution over states. Given a mean-field flow μ, the goal for an agent is to find a policy π which maximizes the expected sum of rewards:

$$\begin{aligned} \max_\pi \quad & J(\pi, \mu) = \mathbb{E}_\pi \Big[\sum_{t=0}^{H} \gamma^t \, r(s_t, a_t, \mu_t) \Big] \\ \text{subject to:} \quad & s_0 \sim \mu_0 \\ & a_t \sim \pi(\cdot|s_t) \\ & s_{t+1} \sim p(\cdot|s_t, a_t). \end{aligned}$$

By making the reward depend only on other agents via the mean-field instead of the individual states and actions of all other agents, we obtain a much smaller optimization problem to solve.

Learning in MFGs. In contrast to single-agent RL, where we optimize a stationary reward signal, algorithms for MFGs typically aim to find policies that are close to some equilibrium, as changes in the policy influence the mean-field and vice-versa, making the optimization problem non-stationary. The concept of Nash-Equilibria (NE), a common solution concept in game theory, has been extended to MFGs [23] and is the main optimization target for algorithms solving non-cooperative MFGs.

Definition 2.1. The best response (BR) to a mean-field flow μ is the solution of the optimization problem

$$\pi^* = \operatorname*{argmax}_\pi J(\pi, \mu) = BR(\mu).$$

Definition 2.2 (ϵ-MFNE). An ϵ-Mean-Field Nash Equilibrium with $\epsilon \geq 0$ is defined as a tuple (π, μ^π) for which the following holds:

$$\sup_{\pi' \in \Pi} J(\pi', \mu^\pi) \leq J(\pi, \mu^\pi) + \epsilon.$$

Intuitively, a MFNE entails that no agent has the incentive to deviate unilaterally from the population policy. Early methods to solve MFGs primarily involved solving coupled partial differential equations, typically a forward-backward system of Hamilton-Jacobi-Bellman and Fokker-Planck-Kolmogorov equations, to compute the value function and distribution flow of agents [1, 2]. However, these methods struggle with scalability in high-dimensional state-action spaces and complex environments.

Algorithms for solving MFGs using RL commonly rely on some form of fixed point iteration, alternating between policy updates and the mean-field distribution computation, denoted by the mean-field evaluation operator $\phi(\pi) = \mu^\pi$. Thereby, they use a best response computation step before evaluating the mean-field. On convergence, the fixed point iteration

$$\phi(\pi^*) = \phi(BR(\mu^*)) = \mu^*$$

yields the MFNE (π^*, μ^*). Generally, convergence is not guaranteed [12] and methods from algorithmic game theory, such as Fictitious Play (FP) [4], are used to stabilize training. In recent years, machine learning approaches, particularly RL, have been explored as a promising alternative for solving MFGs [8, 14, 40]. One of the central challenges in using RL to solve MFGs lies in the non-stationarity introduced by multi-agent interactions, which complicates the learning process. Deep Learning variants of FP [15] were adapted to the MFG settings to scale to larger state and action spaces [6, 31, 32, 37].

In this work, we focus on a class of algorithms that evaluate a policy instead of computing a BR at each iteration [5]. Specifically, we adapt the Online Mirror Descent algorithm (OMD) [24, 29] to the offline setting. OMD alternates between policy evaluation and mean-field updates, as outlined in Algorithm 1. The main difference to FP-style algorithms is that OMD tracks the sum of previous Q-functions instead of average policies. The policy update in OMD is a softmax over the sum of previous Q-functions. However, it is not straightforward to sum up Q-functions in the case of nonlinear function approximators, such as neural networks. This can be avoided by leveraging the following identity, as proposed in [24]:

$$\pi^i = \text{softmax}\Big(\frac{1}{\tau}\sum_{j=0}^{i} Q^j\Big) \tag{4}$$

$$= \text{argmax}_{\pi \in \Delta_{\mathcal{A}}}\Big(\langle \pi, Q^i \rangle - \tau\, \text{KL}\big(\pi || \pi^{i-1}\big)\Big) \tag{5}$$

$$= \text{argmax}_{\pi \in \Delta_{\mathcal{A}}}\Big(\langle \pi, \underbrace{Q^i + \tau\, \ln \pi^{i-1}}_{\tilde{Q}^i} \rangle - \tau\, \underbrace{\langle \pi, \ln \pi \rangle}_{\mathcal{H}(\pi)}\Big) \tag{6}$$

$$= \text{softmax}\Big(\frac{1}{\tau}\,\tilde{Q}^i\Big). \tag{7}$$

This insight allows us to apply policy evaluation directly on $\tilde{Q}$ to compute the sum of Q-functions. In Equations (5) and (6), the inner product $\langle \pi, Q^i \rangle$ is the shorthand notation for $\sum_a \pi(a|s)\, Q^i(s,a)$.

The modified Bellman Operator is then defined as:

$$(\mathcal{B}_\mu^\pi\, \tilde{Q})(s,a) = \tilde{r} + \gamma\, \mathbb{E}_{s',a'}\left[\tilde{Q}(s',a') - \tau\, \ln \pi^{i-1}(a'|s')\right] \tag{8}$$

$$\tilde{r}(s,a,\mu) = r(s,a,\mu) + \tau\, \alpha \ln \pi^{i-1}(a|s), \tag{9}$$

with modified reward $\tilde{r}$, which penalizes deviations from the previous policy π^{i-1}. The hyperparameter τ acts as a temperature and scales the sum of Q-values to avoid premature convergence whereas α is a regularization parameter to control how far a new policy can be from the previous policy. For a more detailed derivation, we refer to [24].

Algorithm 1 Munchausen Online Mirror Descent for MFGs [24]

1: **for** $i = 1 \ldots L$ **do**
2: **Mean-Field Update:** $\mu^i \leftarrow \phi(\pi^i)$
3: **Regularized Policy Evaluation:** $\tilde{Q}^{i+1} \leftarrow \mathcal{B}_{\mu^i}^{\pi^i}\, \tilde{Q}^i$
4: **Policy Update:** $\pi^{i+1}(\cdot|s) \leftarrow \text{softmax}(\frac{1}{\tau}\tilde{Q}^{i+1}(s,\cdot))$
5: **end for**

3 RELATED WORK

Recent years brought up many works that address various limitations of MFGs. Lauriere et al. [24] introduce the basis of our work, Deep Munchausen Mirror Descent, a deep neural network based variant of OMD [29] with strong empirical performance. Also Cui and Koeppl [12] and Perrin [30] introduce deep RL based approaches for MFGs. Other works address the assumption of identical agents and extend MFGs to multi-population games [7, 13]. Subramanian et al. [35] propose decentralized MFGs, allowing to lift the assumption of indistinguishable agents. Inverse-RL (IRL) methods are applied to the MFG setting to infer unknown reward signals [10, 11]. Yang et al. [39] develop an IRL approach that learns the dynamics and the reward model from data. However, their work focuses on behavior prediction rather than finding MFNE. Recent work on model-based algorithms in the mean-field control (MFC) setting, a subclass of MFGs in which agents fully cooperate, can learn a model of the environment and use that to optimize a MFC policy with better sample efficiency [16, 34]. Jusup et al. [18] extend this to safety-constrained problems. However, although those approaches learn a model of the environment, they still assume access to the environment for exploration.

Despite many advances in the field of MFGs, the direction of offline learning remains an underexplored topic. SAFARI [9] is, to our knowledge, the only approach specifically designed for offline mean-field RL, which is not directly comparable to our approach as it does not approximate MFNE. Its main innovation lies in using Reproducing Kernel Hilbert Space (RKHS) embeddings to model the mean-field distribution, combined with an uncertainty-regularized value iteration based on fixed trajectories. While theoretically robust, with dataset-dependent performance bounds, SAFARI faces significant scalability issues. The RKHS embeddings require inverting a Gram matrix that grows quadratically with the dataset size, leading to substantial memory demands and computational bottlenecks.

4 METHOD

In this section we discuss the foundational components of Off-MMD. In particular, we elaborate how we can leverage methods from offline policy evaluation to estimate the mean-field μ from static datasets and provide a modified version of the D-MOMD algorithm to adapt it to the offline learning regime.

4.1 Offline Mean-Field Estimation

Fixed point algorithms for solving mean-field games, as discussed in Section 2, iterate between (1) evaluating the mean-field distribution of agents and (2) best-response computation or policy evaluation. The first step, mean-field estimation, can be done via direct computation if one has access to the transition model or via Monte-Carlo samples if a simulator is provided. In scenarios where only a static dataset of previously collected environment interactions is available, online algorithms like D-MOMD [24] are not applicable.

Therefore, we seek to approximate the mean-field flow

$$\mu^{\pi}_{t+1}(s') = \sum_{s\in S}\sum_{a\in A} p(s'|s,a)\,\pi(a|s)\,\mu^{\pi}_{t}(s) \tag{10}$$

without having access to the transition model $p(s'|s,a)$. One approach to mitigate this is to leverage the data to learn a model of the dynamics. However, the models may be inaccurate in large action spaces, where not all actions are frequently visited. Moreover, approximating the environment dynamics with neural networks might cause additional biases from covariate shifts due to the change of policies [38]. In this work, we leverage the fact that this problem can be equivalently treated as an off-policy state density estimation problem. OPE methods are designed to estimate quantities such as rewards or value functions from off-policy samples. We repurpose this idea to estimate the state distribution under a new policy. In particular, we are interested to estimate μ^{π} from samples that are not collected from π but some other, possibly unknown, behavior policy π_β and without having access to environment dynamics $p(s'|s,a)$.

Let $d^{\pi}_{\mu}(s,a) = \pi(a|s)\,\mu(s)$ be the joint state-action distribution given a mean-field μ and policy π. We restate the mean-field flow as an expectation over d^{π}_{μ}:

$$\mu^{\pi}_{t+1}(s') = \mathbb{E}_{(s,a)\sim d^{\pi}_{\mu_t}}\Big[p(s'|s,a)\Big]. \tag{11}$$

In general, we could apply any OPE method capable of estimating density (ratios) such as model-based estimators [19, 41] or DICE-style approaches for estimating stationary distributions [26, 36], because Off-MMD is agnostic to the estimation method. In this work we choose importance sampling to estimate μ^{π}_{t+1}. In particular, we make use of the marginalized importance sampling (MIS) estimator of Xie et al. [38] because of its theoretical properties and its simplicity compared to other approaches, which typically require additional steps, such as solving an inner optimization problem [28, 42] or fitting another model [19, 41].

Let $d^{\beta}(s,a) = \pi_\beta(a|s)\,d(s)$ be the joint state-action distribution of the dataset collected under behavior policy π_β. In practice, π_β is often unknown and can be approximated using the state-conditional empirical distribution over actions in $\mathcal{D}$ [21]. We can apply importance sampling to reformulate Equation (11) as

$$\mu^{\pi}_{t+1}(s') = \mathbb{E}_{(s,a)\sim d^{\beta}}\left[\frac{d^{\pi}_{\mu_t}(s,a)}{d^{\beta}(s,a)}\,p(s'|s,a)\right] \tag{12}$$

$$= \mathbb{E}_{(s,a)\sim d^{\beta}}\left[\frac{\pi(a|s)\,\mu^{\pi}_{t}(s)}{\pi_\beta(a|s)\,d(s)}\,p(s'|s,a)\right]. \tag{13}$$

This factorization of the state-action distribution allows us to apply MIS to approximate Equation (13) using samples $(s^i_t, a^i_t, s^i_{t+1})$ from finite dataset $\mathcal{D}$. Let $\hat{d}(s_t) = \frac{1}{|\mathcal{D}|}\sum_i^{|\mathcal{D}|}\mathbb{1}[s^{(i)}_t = s_t]$ denote the empirical state distribution at time t, then the marginalized state distribution can be estimated recursively by

$$\begin{aligned}\mu^{\pi}_{t+1}(s) &\approx \frac{1}{|\mathcal{D}|}\sum_{i=0}^{|\mathcal{D}|}\frac{\pi(a^{(i)}_t|s^{(i)}_t)}{\pi^{\beta}(a^{(i)}_t|s^{(i)}_t)}\,\frac{\mu^{\pi}_t(s^{(i)}_t)}{\hat{d}(s^{(i)}_t)}\,\mathbb{1}[\,s^{(i)}_{t+1} = s]\\ \mu^{\pi}_0(s) &= \hat{d}(s_0).\end{aligned} \tag{14}$$

This yields an unbiased estimator of μ^{π} with polynomial error bound with respect to time horizon H, which reduces to $O(H)$ in some cases, such as bounded maximum expected returns [38]. Note that, unlike in typical single-agent offline RL scenarios, we can not directly estimate reward r_t, as it depends nonlinearly on μ_t in general. Thus, for the general case, we require access to the reward function. For special cases, such as reward functions monotonic in μ_t, we could in principle approximate r_t directly.

4.2 Offline Munchausen Mirror Descent

In Section 4.1, we introduced an offline method for estimating the mean-field distribution of a policy. This method can, in principle, be directly applied to D-MOMD to adapt it to the offline learning setting. The update rule in our algorithm follows the classic TD-error minimization approach, as used in DQN [27], where the target Q-values are parameterized by $\bar{\theta}$. Specifically, the objective is to minimize the temporal-difference error where the policy evaluation operator is defined as in Equation (8):

$$\min_{\theta}\mathbb{E}_{(s,a,s')\sim\mathcal{D}}\Big[\big(Q_\theta(s,a) - (\mathcal{B}^{\pi}_{\mu}Q_{\bar{\theta}})(s,a)\big)^2\Big]. \tag{15}$$

However, naively applying off-policy algorithms to offline RL tasks typically leads to the overestimation of Q-values for actions not well represented in the dataset. To address this issue, we incorporate a regularization term following the idea of *Conservative Q-Learning* (CQL) [21], which is designed to learn a conservative lower bound of the true Q-function. CQL introduces a regularized version of the Bellman equation, where the objective balances the maximization of the Q-values over the dataset and the minimization of the temporal-difference error. The general CQL optimization problem is given as:

$$\begin{aligned}\min_{Q}\max_{\tilde{\pi}}\ &\mathbb{E}_{(s,a)\sim\mathcal{D},\tilde{\pi}}\big[Q(s,a)\big] - \mathbb{E}_{(s,a)\sim\mathcal{D},\pi_\beta}\big[Q(s,a)\big]\\ &+\big|Q - \mathcal{B}^{*}Q\big|^2 + \mathcal{R}(\tilde{\pi}),\end{aligned} \tag{16}$$

where $\tilde{\pi}$ represents a policy used to define the joint-state-action distribution over which we minimize the state-action values. The second term encourages tighter bounds by maximizing Q-values under the dataset distribution. The last term is the classic Bellman equation minimizing the TD error with a regularization term $\mathcal{R}$ applied to $\tilde{\pi}$. For specific choices of $\mathcal{R}$, the inner maximization

problem can be solved in closed form. A common choice for $\mathcal{R}$ is to use the KL divergence to some prior action distribution ρ. If we chose ρ to be the uniform distribution over actions, we obtain an entropy regularized, closed-form loss function for Off-MMD:

$$\mathcal{L}(\theta) = \eta\, \mathbb{E}_{(s,a)\sim\mathcal{D}}\Big[\log \sum_{a'} \exp Q_\theta(s,a') - Q_\theta(s,a)\Big] + \mathbb{E}_{(s,a,s')\sim\mathcal{D}}\Big[\big(Q_\theta(s,a) - (\mathcal{B}^\pi_\mu Q_{\bar{\theta}})(s,a)\big)^2\Big], \tag{17}$$

where η is a hyperparameter to control the importance of the CQL-regularization term.

The pseudo-code for Off-MMD is shown in Algorithm 2. We use Equation (14) to compute the offline mean-field and Equation (17) to compute the loss function and update the parameters θ via stochastic gradient descent. Off-MMD thus follows the same iterative schema as Algorithm 1.

Algorithm 2 Offline Munchausen Mirror Descent (Off-MMD)

Input: Dataset $\mathcal{D}$, initial parameters θ^1
for $i = 1 \dots L$ **do**
  Estimate mean-field μ^i for current policy using eq. (14)
  for $j = 1 \dots B$ **do**
    Sample batch $\mathcal{B}$: $\{(s_t^k, a_t^k, s_{t+1}^k)\}_{k=1}^N \sim \mathcal{D}$
    Relabel reward using μ_t^i: $r_t^k = r(s_t^k, a_t^k, \mu_t^i)$
    Update: $\theta_i \leftarrow \theta_i - \nabla_\theta \mathcal{L}(\theta_i)$ using eq. (17)
  end for
  $\theta^{i+1} \leftarrow \theta^i$
  Update policy: $\pi(a|s) = \text{softmax}\big(\frac{1}{\tau}\bar{Q}_{\theta_{i+1}}(s,a)\big)$
end for

5 EXPERIMENTAL EVALUATION

We empirically evaluate Off-MMD on two grid-world problems introduced by Lauriere et al. [24] and compare its performance against the online variant. We also conduct experiments to investigate the sensitivity to the quality of the dataset and the importance of the regularization term in the loss function. The algorithms and the environments are implemented in JAX [3] and build on code by Kostrikov [20] and Lanctot et al. [22]. The code for the experiments and hyperparameters are available on GitHub.[1]

5.1 Experiment Setup

To make runs comparable with each other, we employ *Exploitability* as an evaluation criteria (also often referred to as *Regret*):

$$\mathcal{E}(\pi,\mu) = \max_{\pi'} J(\pi',\mu) - J(\pi,\mu).$$

It directly measures how far a learned policy is from a MFNE by quantifying the potential utility an agent can gain by deviating from its policy, with lower exploitability indicating better equilibrium approximation. We use the same evaluation protocol as in [24] and compute the ground truth mean-field and exploitability.

Both, Off-MMD and D-MOMD, optimize a Q function represented as a fully connected neural network with 3 layers of 128 nodes and ReLU activations. The hyperparamer settings are the same for all instances of Off-MMD over all tasks, except the ablation studies.

[1] https://github.com/axelbr/offline-mmd

5.2 Performance Evaluation

We evaluate the performance of our algorithm on two distinct tasks within a gridworld environment consisting of four separated rooms connected by narrow corridors, as described in [24, 25]. Agents can choose from five actions: move up, down, left, right, or stay in place. If an action results in a collision with a wall, the agent remains in its current position. The time horizon for each episode is set to 40 timesteps. The two tasks we evaluate are exploration and crowd navigation. For each task, we train Off-MMD on three datasets of varying quality and compare its exploitability against the baseline. The following variants of Off-MMD are included in the evaluation:

- **D-MOMD**: The online baseline algorithm.
- **Off-MMD (Expert)**: Trained on data collected by a fully trained D-MOMD policy.
- **Off-MMD (Int)**: Trained on data collected from an intermediate checkpoint.
- **Off-MMD (Rand)**: Trained on data collected from a uniform random policy.

All datasets contain 100K episodes with 40 timesteps each. The subsequent sections present the evaluation results for the exploration and crowd navigation tasks.

Exploration Task. In this task, agents start in the left upper corner and must spread evenly across all four rooms. The reward function is defined as

$$r(s_t, a_t, \mu_t) = -\log \mu_t(s_t),$$

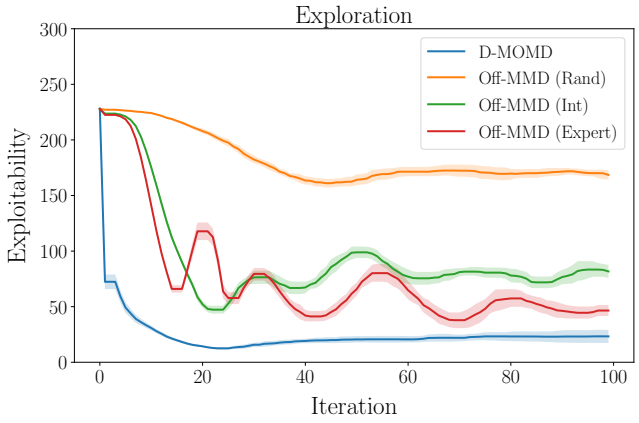
which incentivizes agents to occupy less crowded states, leading to higher rewards when low-density states are reached. The optimal policy for this task spreads evenly across the whole state-space.

In Figure 1a, we present the exploitability over 100 iterations of our algorithm. The online variant rapidly converges to the expected outcome. We evaluate the performance of three instances of Off-MMD, each trained on a dataset of different quality. When trained on sufficiently high-quality datasets, Off-MMD consistently learns policies that perform well. Figure 1b illustrates the evolution of the mean-field over time, supporting this observation. As expected, the policy trained on data generated by a uniform random policy fails to spread evenly across all rooms, particularly in the lower-right room, due to insufficient coverage of this region in the dataset.

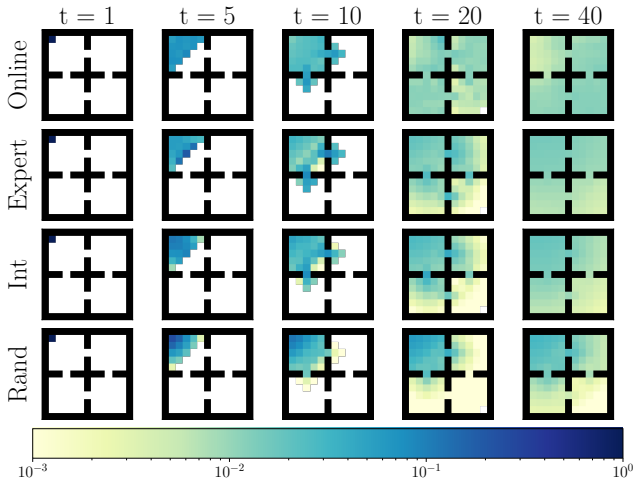
Crowd Modelling with Congestion. In this task, agents also start in the upper left corner. Differently to the exploration task, agents must navigate to the target position in the lower right corner while avoiding high-density areas. Furthermore, we simulate congestion effects by penalizing movements when agents are in crowded areas. The reward function is defined as

$$r(s_t, a_t, \mu_t) = -||s_t - s_{\text{target}}|| - ||a_t||\, \mu_t(s_t) - \log \mu_t(s_t),$$

where $||s_t - s_{\text{target}}||$ denotes the distance to the target and $||a_t||$ is 1 if the agent moves in any direction, 0 otherwise. This is a more complex reward function, as it poses a trade-off of conflicting goals for the agents. The results, shown in Figure 2a, demonstrate convergence of Off-MMD (Exp) and Off-MMD (Int) towards the baseline. Notably, the policy trained on the random dataset also

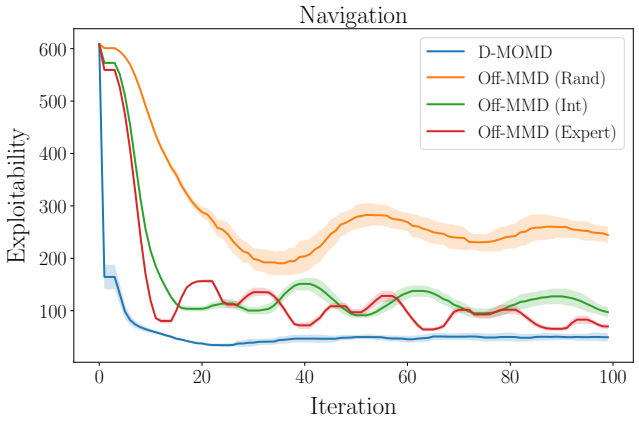


(a) Exploitability

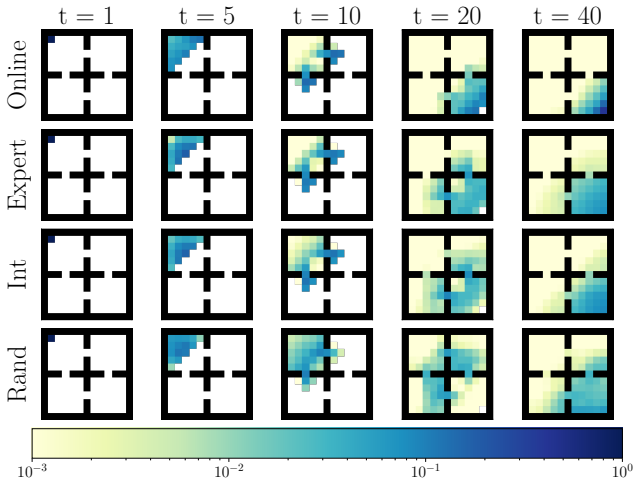


(b) Mean-Field Evolution

Figure 1: (a) Off-MMD can approximate the performance of D-MOMD on the Exploration task when being trained on reasonably good datasets. Training runs were conducted over 10 seeds for 100 iterations of Off-MMD and D-MOMD. We report the mean exploitability and the 95% confidence interval. (b) Evolution of the mean-field over timesteps t. Darker areas indicate higher state density.



(a) Exploitability



(b) Mean-Field Evolution

Figure 2: Off-MMD performs best with intermediate and expert quality datasets. Compared to the exploration task, the policy trained on the random behavior dataset performs better. Experiment settings are the same as in Figure 1.

performs reasonably well. We hypothesize that the distance penalty provides effective guidance, enabling the policy to solve the task even with limited state-action coverage in certain parts of the state space.

5.3 Impact of Dataset Quality

In offline RL, we are interested in the robustness of policy performance to dataset quality. In this experiment, we aim to investigate the sensitivity of Off-MMD to changes in the quality of the trajectories in the dataset and the coverage of the state space.

Building on the methodology of Schweighofer et al. [33], we characterize datasets based on two quality criteria: *state-action coverage* and *trajectory quality*.

Definition 5.1 (State-Action Coverage). Let $u_{s,a}(\mathcal{D})$ denote the number of unique state-action pairs in a dataset $\mathcal{D}$, then the state-action coverage of this dataset is defined as

$$\text{Coverage}(\mathcal{D}) = \frac{u_{s,a}(\mathcal{D})}{|S||A|}. \tag{18}$$

Definition 5.2 (Trajectory Quality). Let $g(\mathcal{D})$ denote the average episode return of a dataset. Furthermore, let $\mathcal{D}_{\text{min}}$ and $\mathcal{D}_{\text{expert}}$ be reference datasets, collected by a suboptimal and an expert policy, respectively. The trajectory quality of dataset $\mathcal{D}$ is defined as

$$\text{Quality}(\mathcal{D}) = \frac{g(\mathcal{D}) - g(\mathcal{D}_{\text{min}})}{g(\mathcal{D}_{\text{expert}}) - g(\mathcal{D}_{\text{min}})}. \tag{19}$$

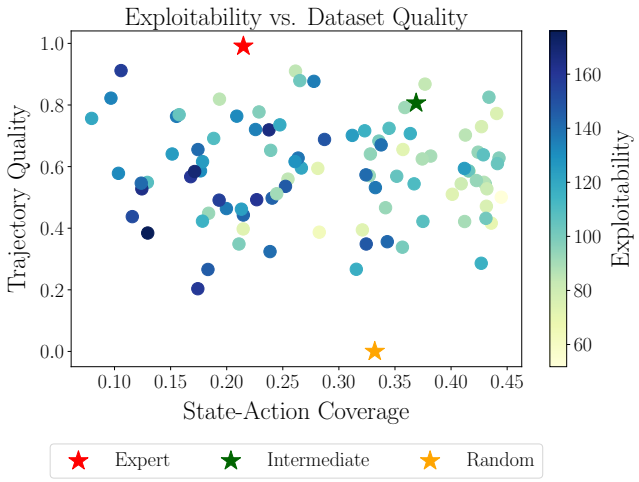


Figure 3: Exploitability vs. Data-Quality: Each point represents a training run of Off-MMD on a dataset with specific state-action coverage and trajectory quality. The color indicates the exploitability of the policy after 100 iterations with darker colors indicating higher exploitability. For reference, we also mark the datasets used in previous experiments.

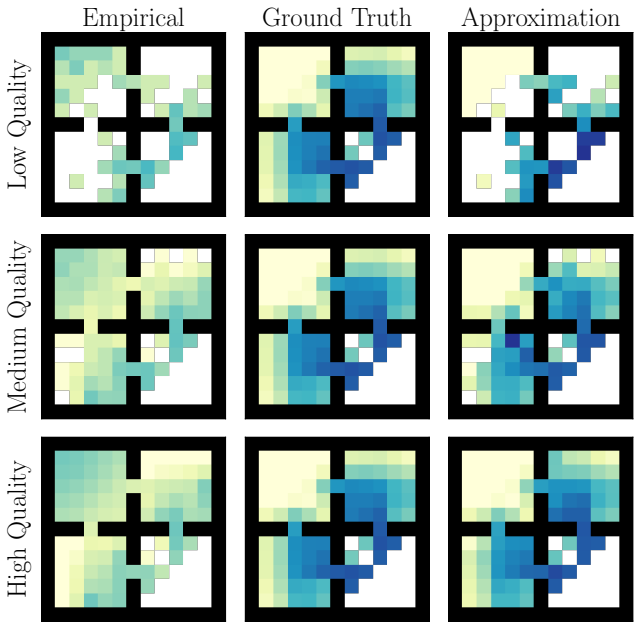


Figure 4: The left column shows the empirical state distribution of datasets collected by behavior policy π_β (from the navigation task) with different state coverages. The center column shows the ground-truth mean-field that is generated by the new policy to evaluate and the right column shows the approximated mean-field of that policy using just the dataset. The mean-fields are picked at $t = 15$. White spots indicate no state-action coverage in this area.

In our experiments, we use datasets collected by the expert policy and the random policy for computing the normalization bounds in Equation (19). Using the three datasets introduced previously for the navigation task, we generate 100 synthetic datasets by randomly subsampling episodes. The dataset sizes range from 1,000 to 100,000 episodes. We then train policies using Off-MMD on these datasets to evaluate the effect of dataset quality on performance.

Figure 3 presents the exploitability of Off-MMD when trained on datasets with varying state-action coverage and trajectory quality. The results indicate a strong correlation between state-action coverage and performance, whereas trajectory quality appears to be a weaker predictor, except in extreme cases such as expert demonstrations or fully random datasets (highlighted in Figure 3). This behavior can be explained by the challenges inherent in multi-agent systems: the policy return is strongly influenced by the behavior of other agents. Therefore, performance achieved under one mean-field setting may not be comparable to another.

In Figure 4, we visualize the approximation of the mean-field under datasets of varying quality for a fixed policy. The left-most column shows the empirical state distribution of the datasets, the center column shows the true mean-field for the policy and serves as a reference, and the right-most column shows the offline approximation. Figure 4 shows how the approximation of the mean-field changes with the state coverage in the dataset. This is particularly prevalent in the first row, where we chose the dataset with the lowest state coverage from the set of synthetic datasets. The approximation can not provide estimates for unvisited parts of the state space. However, for the states that are in the dataset, it produces correct estimates of the mean-field distribution. In scenarios with higher quality datasets, we observe accurate approximations of the ground-truth distribution.

5.4 Effect of Regularization

In the following, we investigate the importance of the CQL regularization term with respect to the robustness to varying levels of state-coverage in the dataset and the training stability of Off-MMD.

The first experiment, shown in Figure 5, aims to examine the effect of the regularization hyperparameter η on the quality of the policy. We conduct training runs with varying regularization strengths, ranging from 0 (e.g. no regularization) to 5 across datasets with different levels of state-action coverage. Specifically, we select five datasets closest to each state-action coverage bin, with coverage values ranging from 0.15 to 0.45. Off-MMD is trained for 100 iterations on each of these datasets, allowing us to analyze the influence of regularization under diverse coverage conditions. Figure 5 shows that moderate regularization allows to reach comparable exploitability as D-MOMD on higher state-action coverages while being significantly more robust than Off-MMD without regularization. The best final performance is achieved with $\eta = 2.0$, which coincides with the recommended setting for CQL [21]. Furthermore, we can see that larger values for the regularization hyperparameter η dampen the difference in performance over different coverage levels, but hurt the maximum achievable performance.

In another ablation experiment, we investigate the training stability of Off-MMD, specifically focusing on monotonic improvement, under different values of the regularization parameter η. Figure 6

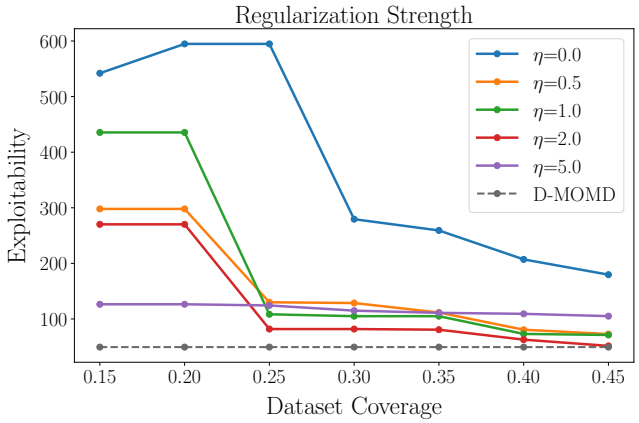


Figure 5: We report the exploitability of policies with different values of η in eq. (17). We train each configuration on 5 datasets that have state-action coverages close to a specific value, ranging from 0.15 to 0.45. We report the mean exploitability of the policies after 100 iterations. For reference, we also plot the performance of the online baseline after 100 iterations.

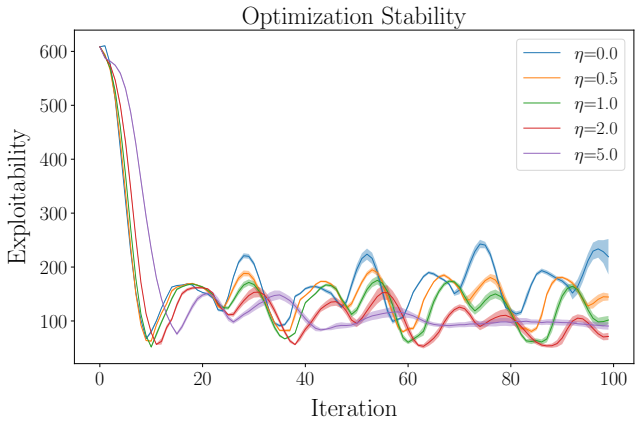


Figure 6: We optimize policies with different values of η on the same expert dataset of the navigation tasks and plot the exploitability over training iterations. We show the mean and 95% confidence interval over 10 seeds.

presents the results of five training runs on the expert dataset of the navigation task, each with a different value of η. The results show that the regularization term plays a crucial role in stabilizing the training dynamics. Lower values of η lead to oscillations in policy performance and, in some cases, even result in divergence, as seen in the unregularized case. In contrast, higher values of η reduce performance fluctuations between iterations, contributing to more stable and consistent learning progress.

6 CONCLUSION

We present Offline Munchausen Mirror Descent (Off-MMD), a novel algorithm designed for learning equilibrium policies in mean-field games using only offline data. This approach addresses the limitations of existing methods that rely on costly and often impractical online interactions. By leveraging importance sampling and Q-value regularization techniques, Off-MMD provides an efficient way to approximate the mean-field distribution from static datasets, ensuring scalability and robustness in complex environments. Our empirical evaluations demonstrated the algorithm's strong performance across two common benchmarks for MFGs, even in scenarios with limited data coverage or sub-optimal datasets. With its ability to scale and adapt to real-world multi-agent systems, Off-MMD opens new avenues for applying RL based algorithms for MFGs to settings where online experimentation is infeasible, irresponsible or difficult to model. We believe that this work lays the foundation for future research into offline learning methods for complex, large-scale multi-agent interactions, bridging the gap between offline RL and MFGs. Future research directions include applications to real-world use-cases such as location recommendation systems or traffic routing, two problem domains suffering from overcrowding effects due to selfish agents.

6.1 Limitations

While Off-MMD marks a first step towards scalable offline RL algorithms for MFGs, it is currently limited to environments where the dynamics are independent of the mean-field. This assumption restricts its applicability to scenarios where the environment dynamics are not arbitrarily influenced by the collective behavior of agents. Addressing this limitation offers a promising avenue for future research. One potential solution could involve adapting model-based algorithms specifically designed for offline RL settings, to handle mean-field dependencies in dynamics. Such advancements would broaden the scope of Off-MMD, making it applicable to a wider range of multi-agent systems where interactions between agents and the environment are more intertwined.

ACKNOWLEDGMENTS

This work is supported by the Austrian Research Promotion Agency (FFG) Project grant No. FO999887513.

REFERENCES

[1] Yves Achdou, Fabio Camilli, and Italo Capuzzo-Dolcetta. 2012. Mean Field Games: Numerical Methods for the Planning Problem. *SIAM Journal on Control and Optimization* 50, 1 (2012), 77–109. https://doi.org/10.1137/100790069 arXiv:https://doi.org/10.1137/100790069

[2] Yves Achdou and Italo Capuzzo-Dolcetta. 2010. Mean Field Games: Numerical Methods. *SIAM J. Numer. Anal.* 48, 3 (2010), 1136–1162. https://doi.org/10.1137/090758477 arXiv:https://doi.org/10.1137/090758477

[3] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. 2018. *JAX: composable transformations of Python+NumPy programs.* Google. http://github.com/jax-ml/jax

[4] George W. Brown. 1951. Iterative Solution of Games by Fictitious Play. In *Activity Analysis of Production and Allocation*, T. C. Koopmans (Ed.). Wiley, New York.

[5] Cacace, Simone, Camilli, Fabio, and Goffi, Alessandro. 2021. A policy iteration method for mean field games. *ESAIM: COCV* 27 (2021), 85. https://doi.org/10.1051/cocv/2021081

[6] Pierre Cardaliaguet and Saeed Hadikhanloo. 2017. Learning in mean field games: The fictitious play. *ESAIM: Control, Optimisation and Calculus of Variations* 23, 2 (2017), 569–591. https://doi.org/10.1051/cocv/2016004

[7] Rene Carmona, Daniel Cooney, Christy Graves, and Mathieu Lauriere. 2019. Stochastic Graphon Games: I. The Static Case. arXiv:1911.10664 [math.OC] https://arxiv.org/abs/1911.10664

[8] René Carmona, Mathieu Laurière, and Zongjun Tan. 2023. Model-free mean-field reinforcement learning: Mean-field MDP and mean-field Q-learning. *The Annals of Applied Probability* 33, 6B (2023), 5334 – 5381. https://doi.org/10.1214/23-AAP1949

[9] Minshuo Chen, Yan Li, Ethan Wang, Zhuoran Yang, Zhaoran Wang, and Tuo Zhao. 2021. Pessimism Meets Invariance: Provably Efficient Offline Mean-Field Multi-Agent RL. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 17913–17926. https://proceedings.neurips.cc/paper_files/paper/2021/file/9559fc73b13fa721a816958488a5b449-Paper.pdf

[10] Yang Chen, Libo Zhang, Jiamou Liu, and Shuyue Hu. 2022. Individual-Level Inverse Reinforcement Learning for Mean Field Games. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems* (Virtual Event, New Zealand) *(AAMAS '22)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 253–262.

[11] Yang Chen, Libo Zhang, Jiamou Liu, and Michael Witbrock. 2023. Adversarial Inverse Reinforcement Learning for Mean Field Games. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems* (London, United Kingdom) *(AAMAS '23)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1088–1096.

[12] Kai Cui and Heinz Koeppl. 2021. Approximately Solving Mean Field Games via Entropy-Regularized Deep Reinforcement Learning. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 130)*, Arindam Banerjee and Kenji Fukumizu (Eds.). PMLR, 1909–1917. https://proceedings.mlr.press/v130/cui21a.html

[13] Christian Fabian, Kai Cui, and Heinz Koeppl. 2023. Learning Sparse Graphon Mean Field Games. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 206)*, Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (Eds.). PMLR, 4486–4514. https://proceedings.mlr.press/v206/fabian23a.html

[14] Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. 2019. Learning Mean-Field Games. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/030e65da2b1c944090548d36b244b28d-Paper.pdf

[15] Johannes Heinrich, Marc Lanctot, and David Silver. 2015. Fictitious Self-Play in Extensive-Form Games. In *Proceedings of the 32nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 37)*, Francis Bach and David Blei (Eds.). PMLR, Lille, France, 805–813. https://proceedings.mlr.press/v37/heinrich15.html

[16] Jiawei Huang, Batuhan Yardim, and Niao He. 2024. On the Statistical Efficiency of Mean-Field Reinforcement Learning with General Function Approximation. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 238)*, Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (Eds.). PMLR, 289–297. https://proceedings.mlr.press/v238/huang24a.html

[17] Minyi Huang, Roland Malhame, and Peter Caines. 2006. Large population stochastic dynamic games: Closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. *Commun. Inf. Syst.* 6 (01 2006). https://doi.org/10.4310/CIS.2006.v6.n3.a5

[18] Matej Jusup, Barna Pásztor, Tadeusz Janik, Kenan Zhang, Francesco Corman, Andreas Krause, and Ilija Bogunovic. 2024. Safe Model-Based Multi-Agent Mean-Field Reinforcement Learning. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems* (Auckland, New Zealand) *(AAMAS '24)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 973–982.

[19] Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. 2020. MOReL: Model-Based Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 21810–21823. https://proceedings.neurips.cc/paper_files/paper/2020/file/f7efa4f864ae9b88d43527f4b14f750f-Paper.pdf

[20] Ilya Kostrikov. 2021. JAXRL: Implementations of Reinforcement Learning algorithms in JAX. https://doi.org/10.5281/zenodo.5535154

[21] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. 2020. Conservative Q-Learning for Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1179–1191. https://proceedings.neurips.cc/paper_files/paper/2020/file/0d2b2061826a5df3221116a5085a6052-Paper.pdf

[22] 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. 2019. OpenSpiel: A Framework for Reinforcement Learning in Games. *CoRR* abs/1908.09453 (2019). arXiv:1908.09453 [cs.LG] http://arxiv.org/abs/1908.09453

[23] J. M. Lasry and Pierre-Louis Lions. 2007. Mean field games. *Japanese Journal of Mathematics* 2 (2007), 229–260. https://api.semanticscholar.org/CorpusID:1963678

[24] Mathieu Lauriere, Sarah Perrin, Sertan Girgin, Paul Muller, Ayush Jain, Theophile Cabannes, Georgios Piliouras, Julien Perolat, Romuald Elie, Olivier Pietquin, and Matthieu Geist. 2022. Scalable Deep Reinforcement Learning Algorithms for Mean Field Games. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 12078–12095. https://proceedings.mlr.press/v162/lauriere22a.html

[25] Mathieu Laurière, Sarah Perrin, Julien Pérolat, Sertan Girgin, Paul Muller, Romuald Élie, Matthieu Geist, and Olivier Pietquin. 2024. Learning in Mean Field Games: A Survey. arXiv:2205.12944 [cs.LG] https://arxiv.org/abs/2205.12944

[26] Jongmin Lee, Wonseok Jeon, Byungjun Lee, Joelle Pineau, and Kee-Eung Kim. 2021. OptiDICE: Offline Policy Optimization via Stationary Distribution Correction Estimation. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 6120–6130. https://proceedings.mlr.press/v139/lee21f.html

[27] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Kirkeby Fidjeland, Georg Ostrovski, Stig Petersen, Charlie Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature* 518 (2015), 529–533. https://api.semanticscholar.org/CorpusID:205242740

[28] Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. 2019. DualDICE: Behavior-Agnostic Estimation of Discounted Stationary Distribution Corrections. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/cf9a242b70f45317ffd281241fa66502-Paper.pdf

[29] Julien Pérolat, Sarah Perrin, Romuald Elie, Mathieu Laurière, Georgios Piliouras, Matthieu Geist, Karl Tuyls, and Olivier Pietquin. 2022. Scaling Mean Field Games by Online Mirror Descent. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems* (Virtual Event, New Zealand) *(AAMAS '22)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1028–1037.

[30] Sarah Perrin. 2022. *Scaling up Multi-agent Reinforcement Learning with Mean Field Games and Vice-versa*. Theses. Université de Lille. https://theses.hal.science/tel-04284876

[31] Sarah Perrin, Mathieu Laurière, Julien Pérolat, Matthieu Geist, Romuald Élie, and Olivier Pietquin. 2021. Mean Field Games Flock! The Reinforcement Learning Way. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Zhi-Hua Zhou (Ed.). International Joint Conferences on Artificial Intelligence Organization, 356–362. https://doi.org/10.24963/ijcai.2021/50 Main Track.

[32] Sarah Perrin, Julien Perolat, Mathieu Lauriere, Matthieu Geist, Romuald Elie, and Olivier Pietquin. 2020. Fictitious Play for Mean Field Games: Continuous Time Analysis and Applications. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 13199–13213. https://proceedings.neurips.cc/paper_files/paper/2020/file/995ca733e3657ff9f5f3c823d73371e1-Paper.pdf

[33] Kajetan Schweighofer, Marius-constantin Dinu, Andreas Radler, Markus Hofmarcher, Vihang Prakash Patil, Angela Bitto-nemling, Hamid Eghbal-zadeh, and Sepp Hochreiter. 2022. A Dataset Perspective on Offline Reinforcement Learning. In *Proceedings of The 1st Conference on Lifelong Learning Agents (Proceedings of Machine Learning Research, Vol. 199)*, Sarath Chandar, Razvan Pascanu, and Doina Precup (Eds.). PMLR, 470–517. https://proceedings.mlr.press/v199/schweighofer22a.html

[34] Pier Giuseppe Sessa, Maryam Kamgarpour, and Andreas Krause. 2022. Efficient Model-based Multi-agent Reinforcement Learning via Optimistic Equilibrium Computation. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 19580–19597. https://proceedings.mlr.press/v162/sessa22a.html

[35] Sriram Ganapathi Subramanian, Matthew E. Taylor, Mark Crowley, and Pascal Poupart. 2022. Decentralized Mean Field Games. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 9 (Jun. 2022), 9439–9447. https://doi.org/10.1609/aaai.v36i9.21176

[36] Junfeng Wen, Bo Dai, Lihong Li, and Dale Schuurmans. 2020. Batch stationary distribution estimation. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*. JMLR.org, Article 945, 11 pages.

[37] Qiaomin Xie, Zhuoran Yang, Zhaoran Wang, and Andreea Minca. 2021. Learning While Playing in Mean-Field Games: Convergence and Optimality. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 11436–11447. https://proceedings.mlr.press/v139/xie21g.html

[38] Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. 2019. Towards Optimal Off-Policy Evaluation for Reinforcement Learning with Marginalized Importance Sampling. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/4ffb0d2ba92f664c2281970110a2e071-Paper.pdf

[39] Jiachen Yang, Xiaojing Ye, Rakshit Trivedi, Huan Xu, and Hongyuan Zha. 2018. Deep Mean Field Games for Learning Optimal Behavior Policy of Large Populations. In *International Conference on Learning Representations*. https://openreview.net/forum?id=HktK4BeCZ

[40] Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. 2018. Mean Field Multi-Agent Reinforcement Learning. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 5571–5580. https://proceedings.mlr.press/v80/yang18d.html

[41] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. 2020. MOPO: Model-based Offline Policy Optimization. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 14129–14142. https://proceedings.neurips.cc/paper_files/paper/2020/file/a322852ce0df73e204b7e67cbbef0d0a-Paper.pdf

[42] Ruiyi Zhang*, Bo Dai*, Lihong Li, and Dale Schuurmans. 2020. GenDICE: Generalized Offline Estimation of Stationary Values. In *International Conference on Learning Representations*. https://openreview.net/forum?id=HkxlcnVFwB