Mean-Field Multi-Agent Learning: A Trust Region Approach

Antonio Ocello
Centre de Mathématiques Appliquées
École Polytechnique
Palaiseau, 91120, France
antonio.ocello@polytechnique.edu

Lorenzo Mancini
Centre de Mathématiques Appliquées
École Polytechnique
Palaiseau, 91120, France
lorenzo.mancini@polytechnique.edu

Safwan Labbi
Centre de Mathématiques Appliquées
École Polytechnique
Palaiseau, 91120, France
safwan.labbi@polytechnique.edu

Daniil Tiapkin
Centre de Mathématiques Appliquées
École Polytechnique
Palaiseau, 91120, France
daniil.tiapkin@polytechnique.edu

Adel Belouchrani
École Nationale Polytechnique
LDCCP
Algiers, Algeria
adel.belouchrani@g.enp.edu.dz

Éric Moulines
Centre de Mathématiques Appliquées
École Polytechnique
Palaiseau, 91120, France
eric.moulines@polytechnique.edu

***Abstract*—Multi-Agent Reinforcement Learning (MARL) has achieved remarkable success in various applications, yet scalability and non-stationarity remain fundamental challenges, particularly in large-scale multi-agent environments. To address these issues, we propose Mean-Field Trust Region Policy Optimization (MF-TRPO), an algorithm that extends the Trust Region Policy Optimization (TRPO) framework to the mean-field setting. Our approach leverages mean-field approximations to mitigate the complexity of multi-agent interactions while preserving decentralized decision-making. By incorporating entropic regularization, MF-TRPO ensures stable and robust policy updates, enhancing convergence properties and enabling structured optimization in non-linear MARL settings. We validate our algorithm through numerical simulations of different scenarios, demonstrating that MF-TRPO achieves competitive performance compared to standard mean-field algorithms such as Fictitious Play and Online Mirror Descent. Our results highlight the effectiveness of MF-TRPO in handling large-scale interactions while maintaining stability and adaptability.**

***Index Terms*—multi-agent reinforcement learning, mean-field games, trust region policy optimization, Nash equilibrium, scalable learning.**

I. Introduction

Multi-Agent Reinforcement Learning (MARL) has achieved significant success across various domains, including telecommunications (see, *e.g.*, [4]). Many real-world problems naturally involve multiple agents whose interactions can be cooperative, competitive, or a combination of both. MARL's capacity to model and optimize these dynamic multi-agent interactions makes it a compelling approach for addressing complex decision-making challenges.

One of the core difficulties in MARL is *scalability*, as the joint state-action space grows exponentially with the number of agents. Additionally, during training, each agent's policy evolves while others simultaneously update theirs, leading to a severe non-stationarity problem that intensifies with larger populations of players [12]. A widely adopted solution to mitigate these issues is Centralized Training with Decentralized Execution (CTDE) [5], [8]. This framework leverages centralized information during training to counteract non-stationarity while still enabling agents to make independent decisions during execution.

However, most existing approaches primarily focus on scenarios with tens of agents. As the scale increases to multi-agent systems (MAS) with hundreds of agents, the issues of non-stationarity and independent learning become significantly more pronounced. In this context, Mean-Field Reinforcement Learning (MFRL) [10] provides a feasible framework to address scalability issues. MFRL models players' interactions by approximating the influence of many agents through an averaged effect. This approximation reduces computational complexity and enables more scalable learning while preserving decentralized decision-making capabilities and mitigating instability in training.

The fully decentralized nature of self-governed systems enables *independent learning*, as agents adapt to collective dynamics using only local information. This aspect is well captured by the Mean-Field Nash Equilibrium (MFNE) framework, which provides a structured approach to analyzing equilibrium behavior.

In this paper, we introduce Mean-Field Trust Region Policy Optimization (MF-TRPO), an algorithm that extends proximal methods to the mean-field setting. By leveraging trust-region updates, MF-TRPO enhances stability in policy updates while ensuring smooth convergence—key challenges in nonlinear MFRL models.

The work of A.O. and E.M. was funded by the European Union (ERC-2022-SYG-OCEAN-101071601). Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. The work of L.M., S.L., and E.M., has been supported by Technology Innovation Institute (TII), project Fed2Learn.

II. Related works

a) Mean-Field Approaches in MARL: Many approaches have been proposed in the MFRL setting. One notable method is Fictitious Play (FP) [7], which employs a regularized version of the softmax best response combined with a regularized mean-field update, providing a structured iterative approach for agents to adapt their strategies based on empirical population distributions. Additionally, Q-learning-based methods have been explored [1], [3], but these approaches often suffer from instability due to their reliance on value function estimation. To mitigate these challenges, strong regularization in the softmax policy updates is commonly applied, which can lead to overly constrained exploration and slow adaptation in dynamic environments.

b) Proximal methods: Beyond FP, some works have explored mirror descent approaches [6], [11], leveraging convex optimization techniques to refine policy updates while ensuring stable learning dynamics. These methods have demonstrated strong empirical performance but often rely on stringent assumptions regarding policy smoothness and regularization.

c) Regularization: Many of these methods [1], [3], [11] impose strong constraints on policy updates by enforcing a minimum temperature for softmax policies. This restriction ensures a degree of exploration but significantly limits the adaptability of policies, effectively forcing them to remain close to the uniform distribution. Such a constraint can hinder the learning process, preventing the policy from converging efficiently to optimal strategies in complex environments.

A more natural approach is to introduce regularization directly into the optimization problem rather than artificially constraining the policy structure. By incorporating a controlled bias through entropic regularization, policies retain flexibility while ensuring stable updates. This perspective allows for smoother convergence and enhances stability without imposing rigid constraints that could undermine learning efficiency.

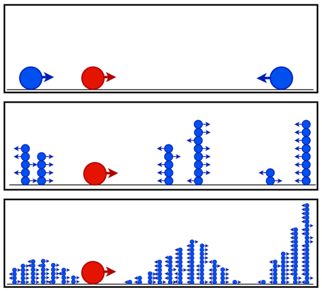

Fig. 1. This image illustrates the intuition on how *anonymity* and *homogeneity* in a MARL problem lead to a MFRL formulation. As the number of players increases, the influence of each individual agent becomes negligible, leading to a MFRL formulation. In this setting, a representative agent (in red) optimizes its policy while interacting with a general population of similar players (in blue), represented as a probability distribution over states. This abstraction allows the analysis to shift from an explicit multi-agent system to a single-agent optimization problem within a mean-field framework.

III. MF-TRPO

a) Framework: A Mean-Field Markov Decision Process (MF-MDP) is defined as a tuple $M = (\mathcal{S}, \mathcal{A}, \mathsf{P}, \mathsf{r}, \gamma)$, where $\mathcal{S}$ represents the finite state space, $\mathcal{A}$ the finite action space, P the transition kernel, r the reward function, and γ the discount factor. In this framework, an agent interacts with the mean-field distribution μ, which captures the aggregate behavior of the population. The system evolves according to the transition operator P^{π}_{μ}, which is induced by the policy π mapping states to distributions over actions.

In this setting, an agent aims to maximize its expected discounted return, given by:

$$J(\pi, \mu, \xi) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \left(\mathsf{r}\left(s_t, a_t, \mu\right) + \eta \log\left(\pi(a_t|s_t)\right)\right)\right],$$

where the initial state s_0 is drawn from the distribution ξ, and each action a_t is selected according to the policy π. The regularization parameter η serves as a tradeoff balancing learning stability, exploration, and exploitation. The value function of the Mean-Field Game (MFG), $v(\mu, \xi) = \max_{\pi} J(\pi, \mu, \xi)$, represents the optimal achievable return over all policies, ensuring that the agent maximizes its expected reward. Setting η on the order of $\widetilde{O}(1/\sqrt{N})$ ensures consistency with the approximation accuracy of the MARL problem, aligning with established trade-offs in MFRL approximations for large-scale MARL settings.

b) Nash equilibrium: A Nash equilibrium in game theory represents a stable state where no agent benefits from unilaterally changing their strategy. In MFGs, this concept extends to large populations, where each agent optimally responds to the overall population dynamics. This results in a self-consistent equilibrium in which the mean-field distribution influences individual decisions, which, in turn, shape the population dynamics.

A Mean-Field Nash Equilibrium (MFNE) is defined by a policy π^* and a population distribution μ^* that satisfy two conditions: *(rationality)* π^* is the best response given μ^*, ensuring no agent has an incentive to deviate; and *(consistency)* μ^* remains stable under π^*, forming a fixed point of the mean-field dynamics.

A. MF-TRPO

a) μ-parametric TRPO: Trust Region Policy Optimization (TRPO) [9] is an algorithm designed to iteratively refine policies while ensuring stability through constrained updates. By enforcing a trust region constraint, TRPO prevents abrupt policy shifts that could lead to instability or performance degradation. Entropic regularization further enhances this stability, enabling smooth optimization and reliable convergence.

In this framework, the policy update admits a closed-form solution based on the Q-function, which quantifies the expected return of taking action a in state s under policy π, *i.e.*,

$$\begin{aligned}&\texttt{PolicyUpdate}(\pi, Q; \ell)(a, s)\\ &:= \frac{\pi(a|s)\exp\left(\frac{1}{\eta(\ell+2)}\left(Q(s,a) - \eta\log\pi(a|s)\right)\right)}{\sum_{a'\in\mathcal{A}}\pi(a'|s)\exp\left(\frac{1}{\eta(\ell+2)}\left(Q(s,a') - \eta\log\pi(a'|s)\right)\right)} \ .\end{aligned}$$

Formally, the Q-function in the mean-field setting is defined as:

$$Q^{\pi}_{\mu}(s,a) := \mathsf{r}\left(s,a,\mu\right) + \gamma\sum_{s'\in\mathcal{S}}\mathsf{P}\left(s'|s,a,\mu\right)\cdot J(\pi,\mu,s') \ .$$

This formulation captures both immediate rewards and future expected returns, guiding stable and efficient policy improvements while ensuring that the updated policy remains within the probabilistic simplex. This algorithm is considered in a parametric setting dependent on the population distribution μ, where it optimizes policies under a fixed μ. This formulation allows the algorithm to be embedded within a broader population update framework, ensuring a structured integration of policy learning and population dynamics.

Algorithm 1 $\texttt{TRPO}(\mu)$

```
Initialize: $\pi_0$ is the uniform policy.
for $\ell \in [L]$ do
    $J(\pi_\ell,\mu,\mu) \leftarrow \mu(\mathrm{I} - \gamma\mathsf{P}^{\pi_\ell}_{\mu})^{-1}\mathsf{r}^{\pi_k}_{\mu}$
    $\mathcal{S}_{\mathsf{d}^{\pi_\ell}_{\nu,\mu}} := s \in \mathcal{S} : \mathsf{d}_{\nu,\mu}\pi_\ell > 0$
    for $s \in \mathcal{S}_{\mathsf{d}^{\pi_\ell}_{\nu,\mu}}$ do
        for $a \in \mathcal{A}$ do
            $Q_{\pi_\ell,\mu}(s,a) \leftarrow \mathsf{r}\left(s,a,\mu\right)$
                    $+\gamma\sum_{s'}\mathsf{P}\left(s'|s,a,\mu\right)J(\pi_\ell,\mu,s')$
        end for
        $\pi_{\ell+1}(a|s) \leftarrow \texttt{PolicyUpdate}(\pi_\ell, Q_{\pi_\ell,\mu}; \ell)(a,s)$
    end for
end for
Output: $\pi_L$.
```

b) Mean-Field TRPO: MF-TRPO extends TRPO to solve the MFG problem. The algorithm alternates between optimizing a policy under a fixed mean-field distribution μ using $\texttt{TRPO}(\mu)$ and later updating it accordingly. This procedure ensures smooth convergence to the equilibrium distribution, preserving stability while adapting to the evolving dynamics of agent interactions.

This mechanism mitigates non-stationarity and ensures convergence toward a MFNE. By iteratively refining both policy and population distribution, MF-TRPO provides a scalable approach to equilibrium learning in MFGs. The inclusion of entropic regularization further enhances its robustness, leading to smoother policy updates and improved convergence stability.

Algorithm 2 Tabular $\texttt{MF-TRPO}$

```
Input: Initial distribution $\mu_0$, number of iterations $K$.
Initialize: Initial policy $\pi_0$ is the uniform policy.
for $k \in [K]$ do
    $\pi_k \leftarrow \texttt{TRPO}(\mu_{k-1})$.
    $\mu_k := \mu_{k-1} + \beta_k\left(\mu_{k-1}\left(\mathsf{P}^{\pi_k}_{\mu_{k-1}}\right)^M - \mu_{k-1}\right)$
            # Update population distribution
end for
Output: $\mu_K$.
```

IV. Simulations

We evaluate MF-TRPO in three settings related to *Crowd Modeling games*. In these frameworks, agents navigate a structure that can be either a grid or a connected graph to avoid congestion. We benchmark our approach against FP and Online Mirror Descent (OMD), two standard algorithms for computing MFGs, demonstrating that MF-TRPO achieves performance on par with state-of-the-art methods.

These examples have been extensively studied to assess scalability, convergence, and adaptability of different approaches [2], [6], [7]. Their structured yet challenging dynamics serve as standard benchmarks, providing a rigorous testing ground for validating algorithmic advancements in this field.

a) Four Rooms Crowd Modeling Game: In this environment, inspired by the Four Rooms example from [2], a population of agents navigates and distributes themselves across a grid, which may include obstacles. Our model consists of a two-dimensional grid with discrete positions, structured into four interconnected rooms separated by walls with narrow passageways. Each agent's state is defined by their position, and they can choose between five possible actions: moving left, right, up, down, or staying in place.

The reward function is designed to naturally discourage overcrowding by penalizing agents based on the population density at their location. Specifically, agents receive a negative reward proportional to the logarithm of the density at their destination, incentivizing them to distribute more evenly across the state space. Additionally, a small bonus is awarded for staying in place, while moving in any direction incurs a penalty. Formally, the reward function is defined as:

$$\mathsf{r}\left(s,a,\mu\right) = -K\log(\mu(s)) + \Gamma(a) \ , \qquad (1)$$

where

$$\Gamma(a) = \begin{cases} 0.2 \ , & \text{if } a = 0 \ , \qquad \text{(Stay)} \\ -0.2 \ , & \text{if } a \in \{\text{Left}, \text{Right}, \text{Up}, \text{Down}\} \ . \quad \text{(Move)}\end{cases}$$

This formulation highlights the role of the parameter K in balancing crowd aversion with the incentive to remain static. Specifically, K regulates the balance between agents spreading out to mitigate congestion and the resistance encountered during movement. This means that, for high values of K, the population distribution converges toward a stationary state resembling a uniform distribution over the state space, as the crowd aversion dominates. Conversely, for low values of K, a

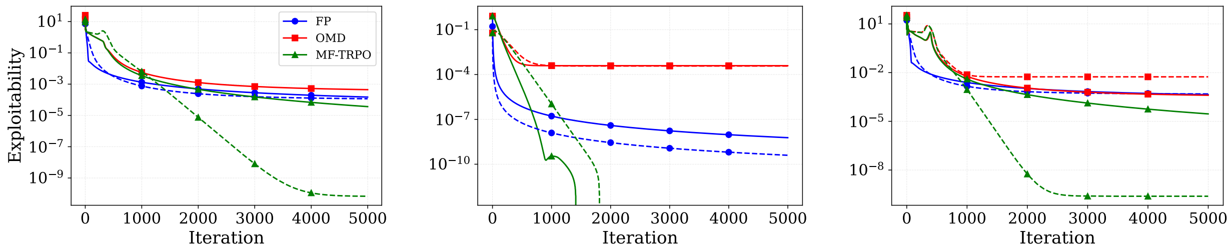

Fig. 2. Reading order: **(a)** Grid-based crowd modeling, **(b)** Islands crowd modeling, and **(c)** Grid-based crowd modeling with point of interest. Each environment is evaluated using the Exploitability metric, providing a comprehensive assessment of equilibrium approximation and learning stability across different settings. The solid lines represent values for $\eta = 0.05$, while the dashed ones represent values for $\eta = 0.3$.

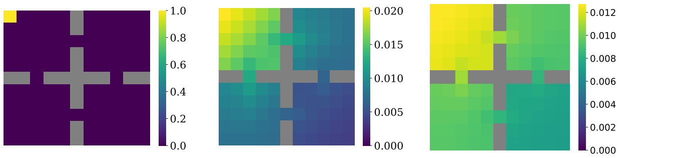

Fig. 3. The image illustrates the evolution of the mean field distribution for the standard Four Rooms Game with a regularization parameter of $\eta = 0.05$, starting from a highly concentrated distribution where agents are clustered in a single cell. The first image corresponds to the initial distribution, the second to the distribution after 10^3 iterations, and the third after $5 \cdot 10^3$ iterations. As learning progresses, exploration increases, leading to a gradual spread of agents across the state space. Eventually, the system converges to equilibrium, with the final distribution shaped by the crowd aversion parameter $K = 0.2$.

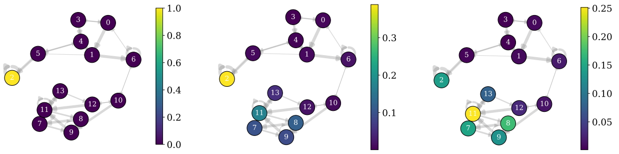

Fig. 4. The image illustrates the evolution of the mean field distribution for the Two-Islands Crowd Game, starting from a highly concentrated distribution, where agents are clustered in node 2, with $\eta = 0.05$. The first image corresponds to the initial distribution, the second to the distribution after 10^2 iterations, and the third after $5 \cdot 10^3$ iterations. The crowd penalty is $K = 0.2$.

non-trivial equilibrium emerges, where the initial distribution of the agents plays a significant role in shaping the long-term dynamics of the system.

The choice of the regularization parameter η significantly influences the final population distribution. This effect is closely tied to the fact that a higher regularization parameter enforces a stronger proximity to the uniform policy, thereby enforcing exploration. Conversely, lower values of η allow for more deterministic policies, enabling agents to better exploit local rewards and settle into non-trivial equilibrium distributions.

In this model, state transitions do not depend on the mean-field parameter but instead incorporate a degree of randomness through slipperiness. Specifically, when an agent chooses to move, she has a high probability of moving in the intended direction while also having a smaller probability of deviating toward other possible directions.

b) Two Islands Graph Crowd Modeling: The Two Islands variation of the Crowd Modeling Game replaces the grid with two interconnected graphs, referred to as *islands*, connected by a single narrow bridge.

We consider $|\mathcal{S}| = 14$ and $|\mathcal{A}| = 2$, with a branching factor of 2. In this setup, the primary challenge arises from the limited connectivity between the two sub-populations. The transition matrix is built randomly, *i.e.*, each node is pushed to visit one its neighbor with a certain probability. The equilibrium now depends not only on local congestion but also on the strategic decision of agents regarding whether to remain on their starting island or transition to the other one. Here, the reward function solely penalizes the logarithm of the mean field distribution. This heightened non-linearity in the environment makes cautious policy updates crucial, as abrupt changes can lead to instability in the learning procedure.

This framework allows for a wider range of experimental variations. Imbalances in rewards between the two islands or an increase in the cost of transitioning between them allow for an investigation of the impact of environmental asymmetries on agent behavior. These modifications offer valuable insights into how equilibrium distributions shift under different constraints and emphasize the importance of stability in large-scale multi-agent reinforcement learning.

A lack of communication between the islands can be observed in Figure 4, as point 6 remains sparsely populated, while the agent mass redistributes into two well-defined clusters. This clustering effect is evident when analyzing the transition matrices, which reveal strong internal connectivity within each island. Notably, convergence occurs rapidly after 10^3 iterations, as illustrated in Figure 2, where exploitability sharply decreases during, indicating a swift stabilization of the learning dynamics.

c) Four Rooms Crowd Modeling Game with point of interest: This game builds upon the previously introduced Four Rooms Game, with the additional feature of guiding players toward a specific point of interest s_{target}. In our case, the latter is set to the bottom-right cell of the grid, encouraging agents to navigate toward this target while trying to avoid crowded situations. The new reward function reads as:

$$\tilde{\mathsf{r}}(s, a, \mu) = \mathsf{r}(s, a, \mu) + \max(0.3 - 0.1 \cdot d(s, s_{\text{target}}), 0) \,,$$

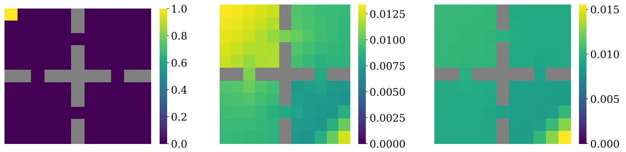

Fig. 5. The image shows the evolution of the mean field distribution for the standard Four Rooms Game with the addition of a point of interest. The first image corresponds to the initial distribution, the second to the state after 10^3 iterations, and the third after $5 \cdot 10^3$ iterations. In this scenario,x the crowd penalty is set to $K = 0.4$.

where $\mathsf{r}(s, a, \mu)$ is a reward function defined in (1), and $d(s, s_{\text{target}})$ represents the distance between between the state s and the state s_{target}, measured as a ℓ_1 distance between the coordinates.

d) Hyperparameters: To ensure the reproducibility of our results, we experiment with a range of hyperparameter values, tuning them to achieve stable learning dynamics.

Hyperparameter	**Value**
Discount factor	$\gamma = 0.9$
Regularization coefficient	$\lambda = \{0.05, 0.3\}$
Number of training iterations	$T = 5 \cdot 10^3$
Population update rate	$\beta = 0.01$
Crowd penalization	$K = \{0.2, 0.4\}$

e) Performance Metrics: To assess the effectiveness of our approach, we evaluate performance using the *Exploitability* metrics. This quantity measures the deviation from equilibrium by quantifying the best response improvement possible for any agent:

$$\phi(\pi, \mu) = \max_{\pi'} J(\pi', \mu^{\pi}, \mu^{\pi}) - J(\pi, \mu^{\pi}, \mu^{\pi}) \ ,$$

with $\mu^{\pi} = \mu(\mathsf{P}_{\pi}^{\mu})^{\infty}$. A lower value indicates that the learned policy is close to a Nash equilibrium.

f) Visualization and Results: Benchmarking ourselves against FP and OMD, as illustrated in the first image of Figure 2, our model demonstrates competitive performance compared to existing approaches in the three examples, achieving superior results in the long run. We observe that our algorithm exhibits less aggressive performances in the early phases of training with respect to the other methods. This behavior is expected, as our approach prioritizes stability by preventing overly greedy updates, ensuring a more cautious adaptation process. As training progresses, our model effectively refines its policy and ultimately surpasses the performance of the competing algorithms, confirming the robustness and effectiveness of MF-TRPO in long-term learning.

These results highlight the robustness and applicability of our approach, demonstrating its effectiveness in capturing equilibrium strategies in large-scale multi-agent settings.

V. Conclusion

In this work, we introduced MF-TRPO, a novel Mean-Field Trust Region Policy Optimization algorithm designed to address scalability challenges in MARL by leveraging mean-field approximations. By extending TRPO to the MFG setting, we formulated a structured approach that ensures stable policy updates while maintaining computational efficiency. Through different experiments, we show that our algorithm effectively balances agent interactions, outperforming traditional benchmarks like FP and OMD. Overall, this work highlights the potential of proximal-based methods in stabilizing multi-agent learning and opens promising directions for further exploration in scalable MARL frameworks.

References

[1] Andrea Angiuli, Jean-Pierre Fouque, Mathieu Laurière, and Mengrui Zhang. Convergence of multi-scale reinforcement Qlearning algorithms for mean field game and control problems. *arXiv preprint arXiv:2312.06659*, 2023.

[2] Matthieu Geist, Julien Pérolat, Mathieu Laurière, Romuald Elie, Sarah Perrin, Olivier Bachem, Rémi Munos, and Olivier Pietquin. Concave utility reinforcement learning: The mean-field game viewpoint. *arXiv preprint arXiv:2106.03787*, 2021.

[3] Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. Learning mean-field games. *Advances in neural information processing systems*, 32, 2019.

[4] Tianxu Li, Kun Zhu, Nguyen Cong Luong, Dusit Niyato, Qihui Wu, Yang Zhang, and Bing Chen. Applications of multi-agent reinforcement learning in future internet: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 24(2):1240–1279, 2022.

[5] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

[6] Julien Pérolat, Sarah Perrin, Romuald Elie, Mathieu Laurière, Georgios Piliouras, Matthieu Geist, Karl Tuyls, and Olivier Pietquin. Scaling mean field games by online mirror descent. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pages 1028–1037, 2022.

[7] Sarah Perrin, Julien Pérolat, Mathieu Laurière, Matthieu Geist, Romuald Elie, and Olivier Pietquin. Fictitious play for mean field games: Continuous time analysis and applications. *Advances in neural information processing systems*, 33:13199–13213, 2020.

[8] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(178):1–51, 2020.

[9] John Schulman, Sergey Levine, Philipp Moritz, Michael Jordan, and Pieter Abbeel. Trust region policy optimization. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*, pages 1889–1897, 2015.

[10] Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean field multi-agent reinforcement learning. In *International conference on machine learning*, pages 5571–5580. PMLR, 2018.

[11] Batuhan Yardim, Semih Cayci, Matthieu Geist, and Niao He. Policy mirror ascent for efficient and independent learning in mean field games. In *International Conference on Machine Learning*, pages 39722–39754. PMLR, 2023.

[12] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pages 321–384, 2021.