Journal of Machine Learning **ISSN**: 2790-2048(e), 2790-203X(p)

Deep Reinforcement Learning for Infinite Horizon Mean Field Problems in Continuous Spaces

Andrea Angiuli [*] [1], Jean-Pierre Fouque [†] [2], Ruimeng Hu [‡] [2,3], and Alan Raydan [§] [3]

[1]Prime Machine Learning Team, Amazon, SEA83, Seattle, WA, 98109, USA.
[2]Department of Statistics and Applied Probability, University of California, Santa Barbara, CA 93106-3110, USA.
[3]Department of Mathematics, University of California, Santa Barbara, CA 93106-3080, USA.

Abstract. We present the development and analysis of a reinforcement learning algorithm designed to solve continuous-space mean field game (MFG) and mean field control (MFC) problems in a unified manner. The proposed approach pairs the actor-critic (AC) paradigm with a representation of the mean field distribution via a parameterized score function, which can be efficiently updated in an online fashion, and uses Langevin dynamics to obtain samples from the resulting distribution. The AC agent and the score function are updated iteratively to converge, either to the MFG equilibrium or the MFC optimum for a given mean field problem, depending on the choice of learning rates. A straightforward modification of the algorithm allows us to solve mixed mean field control games. The performance of our algorithm is evaluated using linear-quadratic benchmarks in the asymptotic infinite horizon framework.

Keywords:
Actor-critic, Linear-quadratic control, Mean field game, Mean field control, Mixed mean field control game, Score matching, Reinforcement learning, Timescales.

Article Info.:
Volume: 4
Number: 1
Pages: 11 - 47
Date: March/2025
doi.org/10.4208/jml.230919

Article History:
Received: 19/09/2023
Accepted: 08/11/2024

Communicated by:
Jiequn Han

1 Introduction

Mean field games and mean field control – collectively dubbed mean field problems – are mathematical frameworks used to model and analyze the behavior and optimization of large-scale, interacting agents in settings with varying degrees of cooperation. Since the early 2000s, with the seminal works [21, 28], MFGs have been used to study the equilibrium strategies of competitive agents in a large population, accounting for the aggregate behavior of the other agents. Alternately, MFC, which is equivalent to optimal control of McKean-Vlasov SDEs [31, 32], focuses on optimizing the behavior of a central decision-maker controlling the population in a cooperative fashion. Cast in the language

[*]aangiuli@amazon.com
[†]fouque@pstat.ucsb.edu
[‡]Corresponding author. rhu@ucsb.edu
[§]alanraydan@ucsb.edu

of stochastic optimal control, both frameworks center on finding an optimal control α_t, which minimizes a cost functional objective $J(\alpha)$ subject to given state dynamics in the form of a stochastic differential equation. What distinguishes mean field problems from classical optimal control is the presence of the mean field distribution μ_t, which may influence both the cost functional and the state dynamics. The mean field is characterized by a flow of probability measures that emulates the effect of a large number of participants whose individual states are negligible but whose influence appears in the aggregate. In this setting, the state process X_t models a representative player from the crowd in the sense that the mean field should ultimately be the law of the state process: $\mu_t = \mathcal{L}(X_t)$. The distinction between MFG and MFC, a competitive game versus a cooperative governance, is made rigorous by precisely how we enforce the relationship between μ_t and X_t. We will address the details of the MFG/MFC dichotomy in greater depth in Section 2.

MFG and MFC theories have been instrumental in understanding and solving problems in a wide range of disciplines, such as economics, social sciences, biology, and engineering. In finance, mean field problems have been applied to model and analyze the behavior of investors and markets. For instance, MFG can be used to model the trading strategies of individual investors in a financial market, taking into account the impact of the overall market dynamics. Similarly, MFC can help optimize the management of large portfolios, where the central decision-maker seeks to maximize returns while considering the average behavior of other investors. For in-depth examples of mean field problems in finance, we refer the reader to [10–12].

Although traditional numerical methods for solving MFG and MFC problems have proceeded along two avenues, solving a pair of coupled partial differential equations (PDE) [13] or a forward-backward system of stochastic differential equations (FBSDE) [6], there has been growing interest in solving mean field problems in a model-free way [3, 4, 14, 19, 29, 34]. With this in mind, we turn to reinforcement learning (RL), an area of machine learning that trains an agent to make optimal decisions through interactions with a "black box" environment. RL can be employed to solve complex problems, such as those found in finance, traffic control, and energy management, in a model-free manner. A key feature of RL is its ability to learn from trial-and-error experiences, refining decision-making policies to maximize cumulative rewards. Temporal difference (TD) methods [37] are a class of RL algorithms that are particularly well-suited for this purpose. They estimate value functions by updating estimates based on differences between successive time steps, combining the benefits of both dynamic programming and Monte Carlo approaches for efficient learning without requiring a complete model of the environment. For a comprehensive overview of the foundations and numerous families of RL strategies, consult [38]. Actor-critic (AC) algorithms – the modern incarnations of which were introduced in [17] – are a popular subclass of TD methods where separate components, the actor and the critic, are used to update estimates of both a policy and a value function. The actor is responsible for selecting actions based on the current policy, while the critic evaluates the chosen actions and provides feedback to update the policy. By combining the strengths of both policy- and value-based approaches, AC algorithms achieve more stable and efficient learning.