A Single Online Agent Can Efficiently Learn Mean Field Games

Chenyu Zhang: Data Science Institute, Columbia University
Xu Chen: Department of Civil Engineering & Engineering Mechanics, Columbia University
Xuan (Sharon) Di: Department of Civil Engineering & Engineering Mechanics, Columbia University (sharon.di@columbia.edu)

Introduction

This work explores single-agent model-free online learning for mean field games (MFGs), where the impact of other agents is encapsulated in the *mean field*, i.e., the population distribution. Solving an MFG aims to find an equilibrium policy and its induced population distribution such that no individual agent can improve its performance by unilaterally deviating from the equilibrium.

Limitations of existing methods:

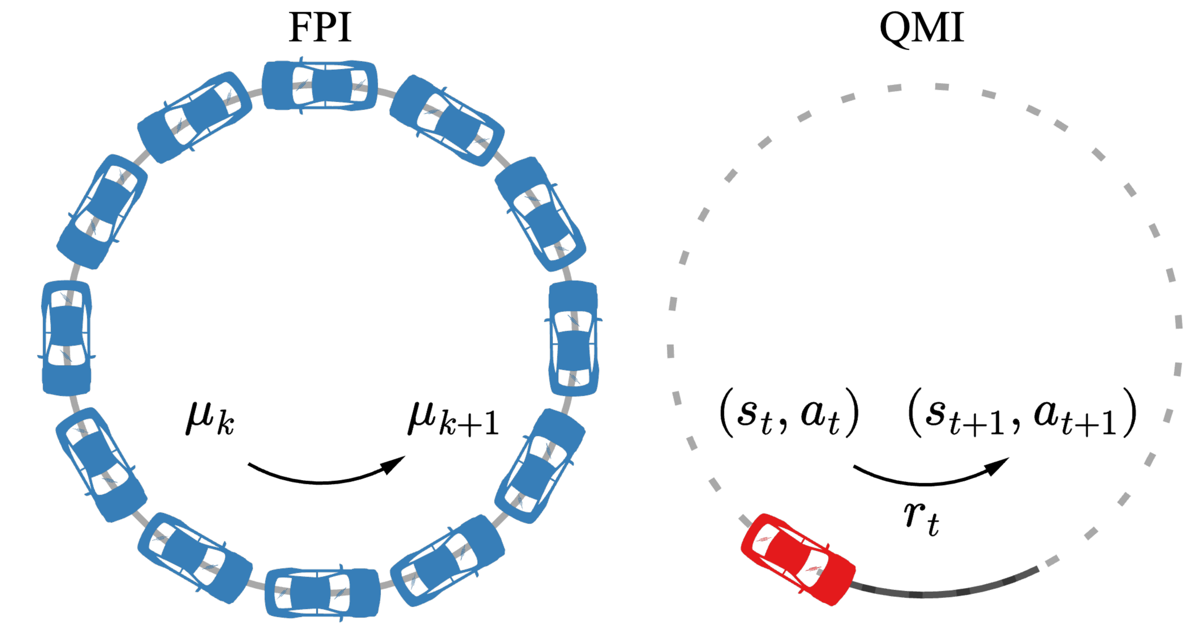
- *Fixed-point iteration* (FPI) and its variants calculate the best responses (BR) and the induced population (IP) distribution *sequentially*, impeding parallel computing and increasing the *computational complexity*.
- Calculating IPs typically requires the knowledge of the transition dynamics, limiting the use of *model-free* methods.
- Without prior knowledge, direct observability of population dynamics is required, restricting the feasibility of learning with a *single online agent* on a single sample trajectory.

> *Can a single online agent efficiently learn the equilibria of mean field games without any prior knowledge?*

Contributions

1. Develop **QM iteration (QMI)**, a novel single-agent model-free scheme for learning MFGs using online samples without prior knowledge of the environment or population.
2. QMI updates the BR and IP estimates *simultaneously* using the same batch of online observations, rendering it *sample-efficient* and *parallelizable*.
3. Two variants, *off-policy* and *on-policy* QMI, are proposed, each with distinct features.
4. Finite time sample complexity guarantees are provided.

Illustration:

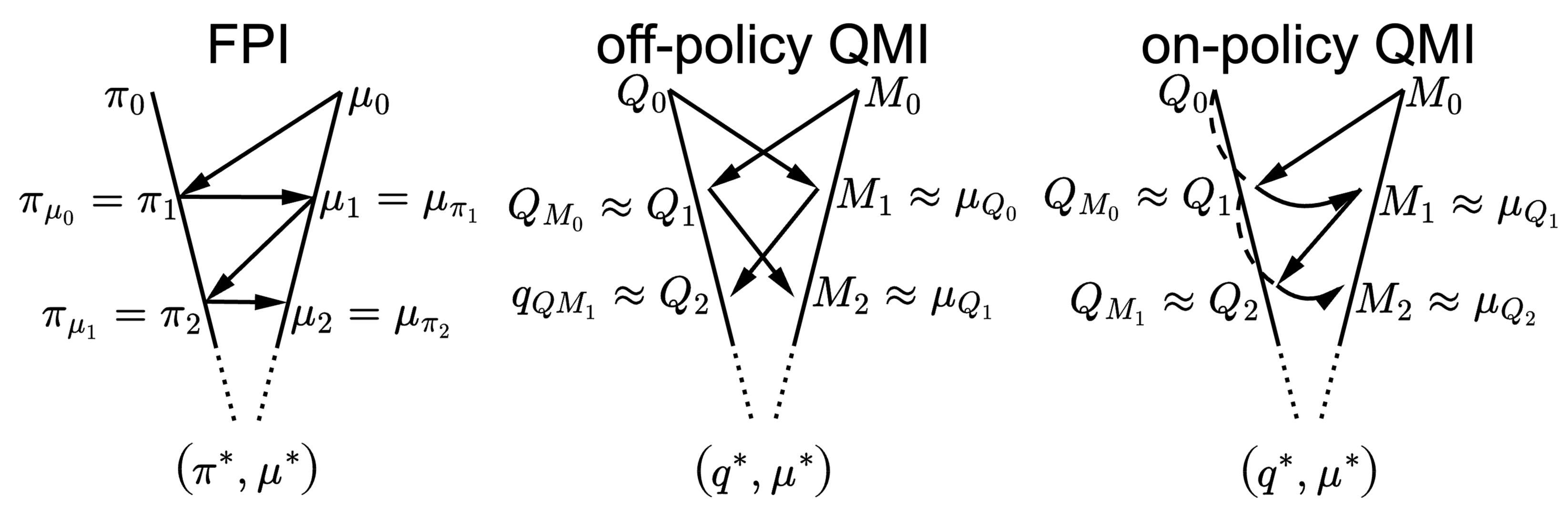


Off-Policy and On-Policy QMI

Pseudocode:

Input: initial value functions $Q_{-1,T} = Q_0$ and $M_{-1,T} = M_0$; initial state s_0; option `off-policy` or `on-policy`
for $k = 0, 1, ..., K$ **do**
  $Q_{k,0} = Q_{k-1,T}$, $M_{k,0} = M_{k-1,T}$
  $\pi_{k,0} = \Gamma_\pi(Q_{k,0})$
  for $t = 0, 1, \ldots, T$ **do**
    sample one Markovian observation tuple $(s_t, a_t, s_{t+1}, a_{t+1})$ following policy $\pi_{k,t}$
    observe the reward $r_{k,t} = r(s_t, a_t, M_{k,0})$
    $Q_{k,t+1}(s_t, a_t) = Q_{k,t}(s_t, a_t) - \alpha_t(Q_{k,t}(s_t, a_t) - r_{k,t} - \gamma Q_{k,t}(s_{t+1}, a_{t+1}))$
    $M_{k,t+1} = M_{k,t} - \beta_t(M_{k,t}(s_t) - \delta_{s_{t+1}})$
    if `off-policy` **then**
      $\pi_{k,t+1} = \pi_{k,0}$
    else if `on-policy` **then**
      $\pi_{k,t+1} = \Gamma_\pi\left(\text{mix}\left(\{Q_{k,l}\}_{l=0}^{t+1}\right)\right)$
    end if
  end for
end for
return $Q_{K,T}, M_{K,T}$

Learning process:



Comparison of two variants:

	Off-Policy	On-Policy
Behavior policy within an outer iteration	fixed	adaptive
Policy type	greedy	soft
MFNE	original	regularized
Sample efficiency boost mechanism	parallel	concurrent
Population-dependent transition kernels	✗	✔

Theorem (Sample complexity of QMI)

Suppose the underlying MDP is ergodic and MFG is $(1-\kappa)$-contractive, as well as the transition kernel and policy operator are L-Lipschitz continuous for off- and on-policy QMI, respectively. Let μ^* be the MFNE population distribution. Then the algorithm returns an ϵ-approximate MFNEwith the number of iterations being at most

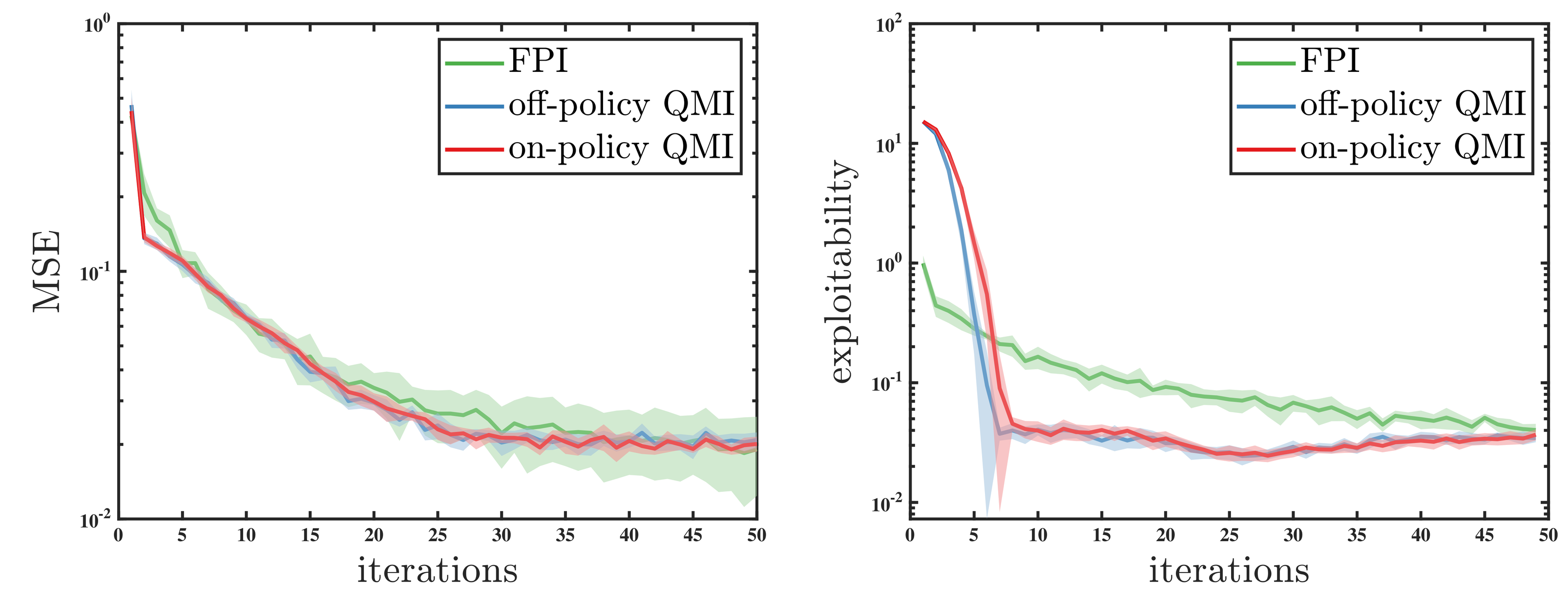
$$K = O\left(\kappa^{-1} \log \epsilon^{-1}\right), \quad T = C \cdot O\left(\kappa^{-2} \epsilon^{-2} \log \epsilon^{-1}\right),$$

where

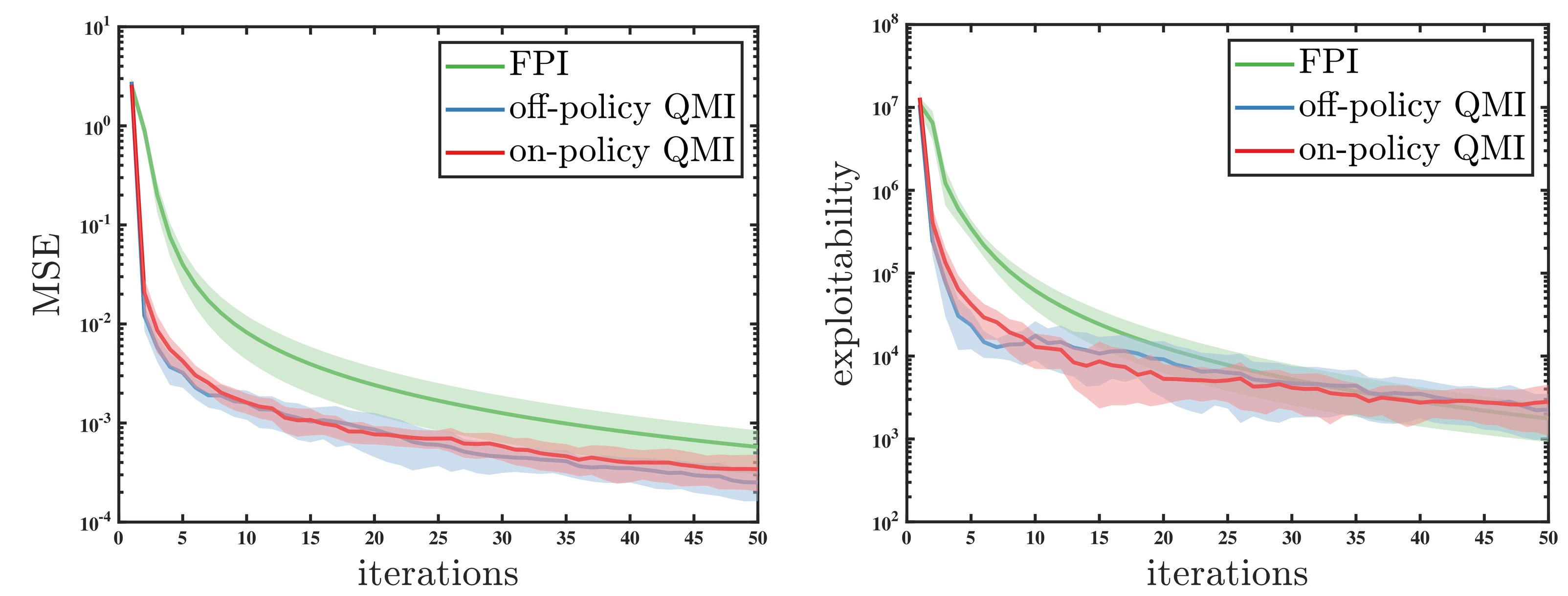
$$C \leq \frac{SAR^2 L^2 \sigma^2}{\lambda_{\min}^2 (1-\gamma)^5}.$$

Numerical Experiments

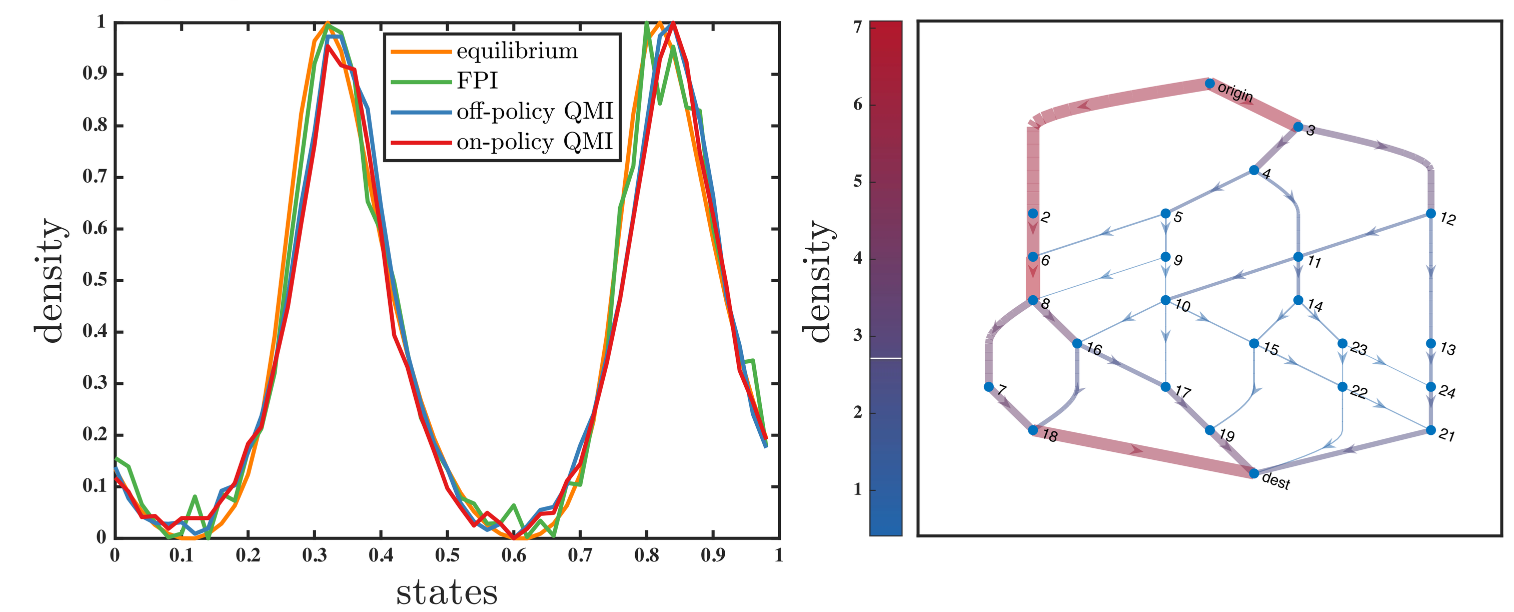
Ring road speed control:



Routing game on the Sioux Falls network:



Learned population distributions:



(a) Ring road (b) Sioux Falls network