Journal of Machine Learning Research 26 (2025) 1-53 Submitted 8/24; Revised 3/25; Published 6/25

Learning with Linear Function Approximations in Mean-Field Control

Erhan Bayraktar ERHAN@UMICH.EDU
Department of Mathematics, University of Michigan, Ann Arbor, MI, USA

Ali Devran Kara AKARA@FSU.EDU
Department of Mathematics, Florida State University, FL, USA

Editor: Quanquan Gu

Abstract

The paper focuses on mean-field type multi-agent control problems with finite state and action spaces where the dynamics and cost structures are symmetric and homogeneous, and are affected by the distribution of the agents. A standard solution method for these problems is to consider the infinite population limit as an approximation and use symmetric solutions of the limit problem to achieve near optimality. The control policies, and in particular the dynamics, depend on the population distribution in the finite population setting, or the marginal distribution of the state variable of a representative agent for the infinite population setting. Hence, learning and planning for these control problems generally require estimating the reaction of the system to all possible state distributions of the agents. To overcome this issue, we consider linear function approximation for the control problem and provide coordinated and independent learning methods. We rigorously establish error upper bounds for the performance of learned solutions. The performance gap stems from (i) the mismatch due to estimating the true model with a linear one, and (ii) using the infinite population solution in the finite population problem as an approximate control. The provided upper bounds quantify the impact of these error sources on the overall performance.

1. Introduction

The goal of the paper is to present various learning methods for mean-field control problems under linear function approximations and to provide provable error bounds for the learned solutions.

1.1 Literature Review

Learning for multi agent control problems is a practically relevant and a challenging problem where there has been as a growing interest in recent years. A general solution methodology for multi-agent control problems is difficult to obtain and the solution, in general, is intractable except for special information structures between the agents. We refer the reader to the survey paper by Zhang et al. (2021) for a substantive summary of learning methods in the context of multi-agent decision making problems.

In this paper, we study a particular case of multi-agent problems in which both the agents and their interactions are symmetric and homogeneous. For these mean-field type decision making problems, the agents are coupled only through the so-called mean-field term. These problems can be broadly divided into two categories; mean-field game problems where the agents are competitive and interested in optimizing their self objective functions, and mean-field control problems, where

the agents are interested in a common objective function optimization. We cite some papers by Gomes and Saúde (2014); Carmona and Delarue (2013); Bensoussan et al. (2013); Tembine et al. (2013); Huang et al. (2007); Anahtarci et al. (2022); Elie et al. (2020); Fu et al. (2020); Guo et al. (2019); Perrin et al. (2020); Subramanian and Mahajan (2019); Saldi et al. (2018, 2019) and references therein, for papers in mean-field game setting. We do not discuss these in detail as our focus will be on mean-field control problems which are significantly different in both analysis and the nature of the problems of interest.

For mean-field control problems, where the agents are cooperative and work together to minimize (or maximize) a common cost (or reward) function, see Bayraktar et al. (2018); Djete et al. (2022); Laurière and Pironneau (2014); Carmona and Laurière (2022); Pham and Wei (2017); Carmona and Laurière (2021); Germain et al. (2022); Bayraktar et al. (2023); Bayraktar and Zhang (2023) and references therein for the study of dynamic programming principle and learning methods in continuous time. In particular, we point out the papers Lacker (2017); Djete et al. (2022) which provide the justification for studying the centralized limit problem by rigorously connecting the large population decentralized setting and the infinite population limit problem.

For papers studying mean-field control in discrete time, we refer the reader to Motte and Pham (2023); Bäuerle (2023); Gu et al. (2021, 2023); Motte and Pham (2022); Carmona et al. (2023). Motte and Pham (2023, 2022) study existence of solutions to the control problem in both infinite and finite population settings, and they rigorously establish the connection between the finite and infinite population problems. Bäuerle (2023) studies the finite population mean-field control problems and their infinite population limit, and provide solutions of the ergodic control problems for some special cases.

In the context of learning, Gu et al. (2021, 2023) study dynamic programming principle and Q learning methods directly for the infinite population control problem. The value functions and the Q functions are defined for the lifted problem, where the state process is the measure-valued mean-field flow. They consider dynamics without common noise, and thus the learning problem from the perspective of a coordinator becomes a deterministic one.

Carmona et al. (2023) also considers the limit (infinite population) problem and studies different classes of policies that achieve optimal performance for the infinite population (limit problem) and focuses on Q learning methods for the problem after establishing the optimality of randomized feedback policies for the agents. The learning problem considers the state as the measure valued mean-field term and defines a learning problem over the set of probability measures where various approximations are considered to deal with the high dimension issues.

Angiuli et al. (2023, 2022) have studied learning methods for the mean-field game and control problems from a joint lens. However, for the control setup, they consider a different control objective compared to the previously cited papers. In particular, they aim to optimize the asymptotic phase of the control problem where the agents are assumed to reach to their stationary distributions under joint symmetric policies. Furthermore, the agents only use their local state variables, and thus the objective is to find a stationary measure for the agents where the cost is minimized under this stationary regime. Since the agents only use their local state variables (and not the mean-field term) for their control, the authors can define a Q function over the finite state and action spaces of the agents.

Pásztor et al. (2023) consider a closely related problem to our setting, where they propose model-based learning methods for the mean-field control. Similar to Gu et al. (2021, 2023), they directly work with the infinite population dynamics without analyzing the approximation consistency

between the finite-population dynamics and their infinite-population counterpart. Furthermore, they restrict the dynamics to the models with additive noise, and the optimality search is within deterministic and Lipschitz continuous controls.

We also note that there are various studies that focus on the application of the mean-field modeling using numerical methods based on machine learning techniques, see e.g. the works by Ruthotto et al. (2020); Achdou et al. (2020); Lauriere et al. (2022).

In this paper, we will consider the learning problem using an alternative formulation where the state is represented as the measure valued mean-field term. To approximate this uncountable space, and the cost and transition functions, different from the previous works in the mean-field control setting, we will consider linear function approximation methods. These methods have been studied well for single agent discrete time stochastic control problems. We cite papers by Melo et al. (2008); Carvalho et al. (2020); Szepesvári and Smart (2004); Jin et al. (2020); Meyn (2024) in which reinforcement learning techniques are used to study Markov decision problems with continuous state spaces using linear function approximations.

Contributions.

- In Section 2, we present the learning methods using linear function approximation. We focus on various scenarios.
  - We first consider the ideal case where we assume that the team has infinitely many agents. For this case, we study; (i) learning by a coordinator who has access to information about every agent in the team, and estimates a model from a data set by fitting a linear model that minimizes the L_2 distance between the training data and the estimate linear model, (ii) each agent estimates their own linear model using their local information via an iterative algorithm from a single sequence of data.
  - In Section 2.3, we consider the practical case, where the team has finitely many agents, and they aim to estimate a linear model from a single sequence of data, using their local information variables.
- The methods we study in Section 2 minimize the L_2 distance between the learned linear model and the actual model under a probability measure that depends on the training data. However, to find upper bounds for the performance loss of the policies designed for the learned linear estimates in any scenario, we need uniform estimation errors rather than L_2 estimation errors. In Section 3, we generalize L_2 error bounds to uniform error bounds.
- The proposed learning methods do not match the true model perfectly in general, due to linear approximation mismatch. Therefore, finally, in Section 4, we provide upper bounds on the performance of the policies that are designed for the learned models when they are applied on the true control problem. We note that the flow of the mean-field term is deterministic for infinitely many agents, and thus can be estimated using the dynamics without observing the mean-field term. Therefore, for the execution of the policies we focus on two methods, (i) *open loop control*, where the agents only observe their local states and estimate the mean-field term with the learned dynamics, (ii) *closed loop control* where the agents observe both their local information and the mean-field term. For each of these execution procedures, we provide upper bounds for the performance loss. As in Section 2, we first consider the ideal case where it is assumed that the system has infinitely many agents. In this case, the error bound depends

on the uniform model mismatch between the learned model and the true model. We then consider the case with finitely many agents. We assume that each agent follows the policy that they calculate considering the limit (infinite population) model. In this case, the error upper bounds depend on both the uniform model mismatch, and an empirical concentration bound since we estimate the finitely many agent model with the infinite population limit problem.

1.2 Problem formulation.

The dynamics for the model are presented as follows: suppose N agents (decision-makers or controllers) act in a cooperative way to minimize a cost function, and the agents share a common state and an action space denoted by $\mathbb{X}$ and $\mathbb{U}$. We assume that $\mathbb{X}$ and $\mathbb{U}$ are finite. We refer the reader to the paper by Bayraktar et al. (2025), for finite approximations of mean-field control problems where the state and actions spaces of the agents are continuous. For any time step t, and agent $i \in \{1, \dots, N\}$ we have

$$x_{t+1}^i = f(x_t^i, u_t^i, \mu_{\mathbf{x_t}}, w_t^i) \tag{1}$$

for a measurable function f, where $\{w_t^i\}$ denotes the i.i.d. idiosyncratic noise process.

Furthermore, $\mu_{\mathbf{x}} \in \mathcal{P}_N(\mathbb{X})$ denotes the empirical distribution of the agents on the state space $\mathbb{X}$ such that for a given joint state of the team of agents $\mathbf{x} := (x^1, \dots, x^N) \in \mathbb{X}^N$

$$\mu_{\mathbf{x}}(\cdot) := \frac{1}{N} \sum_{i=1}^{N} \delta_{x^i}(\cdot)$$

where δ_{x^i} represents the Dirac measure centered at x^i. Throughout this paper, we use the notation

$$\mathbb{X}^N := \underbrace{\mathbb{X} \times \cdots \times \mathbb{X}}_{N \text{ times}}$$

to denote the space of all joint state variables of the team equipped with the product topology on $\mathbb{X}^N$. We further define $\mathcal{P}_N(\mathbb{X})$, the set of all empirical measures on $\mathbb{X}$ constructed using sequences of N states in $\mathbb{X}$, such that

$$\mathcal{P}_N(\mathbb{X}) := \{\mu_{\mathbf{x}} : \mathbf{x} = (x^1, \dots, x^N) \in \mathbb{X}^N\}.$$

Note that $\mathcal{P}_N(\mathbb{X}) \subset \mathcal{P}(\mathbb{X})$ where $\mathcal{P}(\mathbb{X})$ denotes the set of all probability measures on $\mathbb{X}$ equipped with the weak convergence topology.

Equivalently, the next state of the agent i is determined by some stochastic kernel, that is, a regular conditional probability distribution:

$$\mathcal{T}(\cdot | x_t^i, u_t^i, \mu_{\mathbf{x_t}}). \tag{2}$$

At each time stage t, each agent receives a cost determined by a measurable stage-wise cost function $c : \mathbb{X} \times \mathbb{U} \times \mathcal{P}_N(\mathbb{X}) \to \mathbb{R}$. If the state, action, and empirical distribution of the agents are given by $x_t^i, u_t^i, \mu_{\mathbf{x_t}}$, then the agent receives the cost.

$$c(x_t^i, u_t^i, \mu_{\mathbf{x_t}}).$$

For the remainder of the paper, by an abuse of notation, we will sometimes denote the dynamics in terms of the vector state and action variables, $\mathbf{x} = (x^1, \ldots, x^N)$, and $\mathbf{u} = (u^1, \ldots, u^N)$, and vector noise variables $\mathbf{w} = (w^1, \ldots, w^N)$ such that

$$\mathbf{x_{t+1}} = f(\mathbf{x_t}, \mathbf{u_t}, \mathbf{w_t}).$$

For the initial formulation, every agent is assumed to know the state and action variables of every other agent. We define an admissible policy for an agent i, as a sequence of functions $\gamma^i := \{\gamma_t^i\}_t$, where γ_t^i is a $\mathbb{U}$-valued (possibly randomized) function which is measurable with respect to the σ-algebra generated by

$$I_t = \{\mathbf{x}_0, \ldots, \mathbf{x}_t, \mathbf{u}_0, \ldots, \mathbf{u}_{t-1}\}. \tag{3}$$

Accordingly, an admissible *team* policy, is defined as $\gamma := \{\gamma^1, \ldots, \gamma^N\}$, where γ^i is an admissible policy for the agent i. In other words, agents share the complete information.

The objective of the agents is to minimize the following cost function

$$J_\beta^N(\mathbf{x}_0, \gamma) = \sum_{t=0}^{\infty} \beta^t E_\gamma \left[\mathbf{c}(\mathbf{x}_t, \mathbf{u}_t)\right]$$

where E_γ denotes the expectation with respect to the probability measure induces by the team policy γ, and where

$$\mathbf{c}(\mathbf{x}_t, \mathbf{u}_t) := \frac{1}{N} \sum_{i=1}^{N} c(x_t^i, u_t^i, \mu_{\mathbf{x}_t}).$$

The optimal cost is defined by

$$J_\beta^{N,*}(\mathbf{x}_0) := \inf_{\gamma \in \Gamma} J_\beta^N(\mathbf{x}_0, \gamma) \tag{4}$$

where Γ denotes the set of all admissible team policies.

We note that this information structure (3) will be our benchmark for evaluating the performance of the approximate solutions using simpler information structures presented in the paper. In other words, the value function that is achieved when the agents share full information and full history will be taken to be our reference point for simpler information structures.

For example, one immediate observation is that the problem under full information sharing can be reformulated as a centralized control problem where the state and action spaces are $\mathbb{X}^N$ and $\mathbb{U}^N$. Therefore, one can consider Markov policies such that $I_t = \{\mathbf{x}_t\}$ without loss of optimality.

However, if the problem is modeled as an MDP with state space $\mathbb{X}^N$ and action space $\mathbb{U}^N$, we face some computational challenges:

(i) the curse of dimensionality when N is large, since $\mathbb{X}^N$ and $\mathbb{U}^N$ might be too large even when $\mathbb{X}, \mathbb{U}$ are of manageable size,

(ii) the curse of coordination: even if the optimal team policy is found, its execution at the agent level requires coordination among the agents. In particular, the agents may need to follow asymmetric policies to achieve optimality, even though we assume full symmetry for the dynamics and the cost models. The following simple example from Bayraktar and Kara (2024) shows that the agents may need to follow asymmetric policies to achieve optimality which requires coordination among the agents.

Example 1 *Consider a team control problem with two agents, i.e.* $N = 2$*. We assume that* $\mathbb{X} = \mathbb{U} = \{0, 1\}$*. The stage wise cost function of the agents is defined as*

$$c(x, u, \mu_{\mathbf{x}}) = \|\mu_{\mathbf{x}} - \bar{\mu})\|$$

where

$$\bar{\mu} = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1.$$

In words, the state distribution should be distributed equally over the state space $\{0, 1\}$ *for minimal stage-wise cost. For the dynamics we assume a deterministic model such that*

$$x_{t+1} = u_t.$$

In words, the action of an agent purely determines the next state of the same agent. The goal of the agents is to minimize

$$\sum_{t=0}^{\infty} \beta^t E^{g^1, g^2} \left[\frac{c(x_t^1, u_t^1, \mu_{\mathbf{x_t}}) + c(x_t^2, u_t^2, \mu_{\mathbf{x_t}})}{2} \right]$$

for some initial state values $\mathbf{x}_0 = [x_0^1, x_0^2]$*, by choosing policies* g^1, g^2*. The expectation is over the possible randomization of the policies. We assume full information sharing such that every agent has access to the state and action information of the other agent.*

We let the initial states be $x_0^1 = x_0^2 = 0$*. An optimal policy for the agents for the problem is given by*

$$\begin{aligned} g^1(0,0) &= 0, \qquad & g^2(0,0) &= 1 \\ g^1(0,1) &= 0, \qquad & g^2(0,1) &= 1 \\ g^1(1,0) &= 1, \qquad & g^2(1,0) &= 0 \\ g^1(1,1) &= 1, \qquad & g^2(1,1) &= 0 \end{aligned}$$

which always spreads the agents equally over the state space. One can realize that, when the agents are positioned at either $(0, 0)$ *or* $(1, 1)$*, they have to use personalized policies to decide on which one to be placed at* 0 *or* 1*.*

For any symmetric policy $g^1(x^1, x^2) = g^2(x^1, x^2) = g(x^1, x^2)$*, including the randomized ones, there will always be cases with strict positive probability, where the agents are positioned at the same state, and thus the performance will be strictly worse than the optimal performance.*

A standard approach to deal with mean-field control problems when N is large is to consider the infinite population problem, i.e. taking the limit $N \to \infty$. A propagation of chaos argument can be used to show that in the limit, the agents become asymptotically independent. Hence, the problem can be formulated from the perspective of a representative single agent. This approach is suitable to deal with coordination challenges, as the correlation between the agents vanish in the limit, and thus the symmetric policies can achieve optimal performance for the infinite population control problem. In particular, for Example 1 in the infinite population setting, the optimal policy is to follow a randomized policy such that $Pr(u = 1) = Pr(u = 0) = \frac{1}{2}$. We will introduce the limit problem in Section 1.5 and make the connections between the limit problem and the finite population problem rigorous.

1.3 Preliminaries.

Recall that we assume that the state and action spaces of agents $\mathbb{X}, \mathbb{U}$ are finite (see Bayraktar et al. (2025) for finite approximations of continuous space mean-field control problems).

Note. Even though we assume that $\mathbb{X}$ and $\mathbb{U}$ are finite, we will continue using integral signs instead of summation signs for expectation computations due to notation consistency, by simply considering Dirac delta measures.

We metrize $\mathbb{X}$ and $\mathbb{U}$ so that $d(x, x') = 1$ if $x \neq x'$ and $d(x, x') = 0$ otherwise. Note that with this metric, for any $\mu, \nu \in \mathcal{P}(\mathbb{X})$ and for any coupling Q of μ, ν, we have that

$$E_Q\left[|X - Y|\right] = P_Q(X \neq Y)$$

which in particular implies via the optimal coupling that

$$W_1(\mu, \nu) = \|\mu - \nu\|_{TV}$$

where W_1 denotes the first order Wasserstein distance, or the Kantorovich–Rubinstein metric, and $\|\cdot\|_{TV}$ denotes the total variation norm for signed measures.

Note further that for measures defined on finite spaces, we have that

$$\|\mu - \nu\|_{TV} = \frac{1}{2}\|\mu - \nu\|_1 = \frac{1}{2}\sum_x |\mu(x) - \nu(x)|\,. \tag{5}$$

Hence, in what follows we will simply write $\|\mu - \nu\|$ to refer to the distance between μ and ν, which may correspond to the total variation distance, the first order Wasserstein metric, or the normalized L_1 distance.

We also define the following Dobrushin coefficient for the kernel $\mathcal{T}$:

$$\sup_{\mu,\gamma,x,\hat{x}} \| \int_{\mathbb{U}} \mathcal{T}(\cdot|x, u, \mu)\gamma(du|x) - \int_{\mathbb{U}} \mathcal{T}(\cdot|\hat{x}, u, \mu)\gamma(du|\hat{x})\| =: \delta_T \tag{6}$$

Realize that we always have $\delta_T \leq 1$. In certain cases, we can also have strict inequality, e.g. if there exists some $x^* \in \mathbb{X}$ such that

$$\mathcal{T}(x^*|x, u, \mu) < 1 - \alpha, \quad \forall x, u, \mu$$

then one can show that $\delta_T \leq 1 - \alpha < 1$.

1.4 Measure Valued Formulation of the Finite Population Control Problem

For the remaining part of the paper, we will often consider an alternative formulation of the control problem for the finitely many agent case where the controlled process is the state distribution of the agents, rather than the state vector of the agents. We refer the reader to Bayraktar et al. (2025) for the full construction; in this section, we will give an overview.

We define an MDP for the distribution of the agents, where the control actions are the joint distribution of the state and action vectors of the agents.

We let the state space be $\mathcal{Z} = \mathcal{P}_N(\mathbb{X})$ which is the set of all empirical measures on $\mathbb{X}$ that can be constructed using the state vectors of N-agents. In other words, for a given state vector

$\mathbf{x} = \{x^1, \dots, x^N\}$, we consider $\mu_{\mathbf{x}} \in \mathcal{P}_N(\mathbb{X})$ to be the new state variable of the team control problem.

The admissible set of actions for some state $\mu \in \mathcal{Z}$, is denoted by $U(\mu)$, where

$$U(\mu) = \{\Theta \in \mathcal{P}_N(\mathbb{U} \times \mathbb{X}) | \Theta(\mathbb{U}, \cdot) = \mu(\cdot)\}, \tag{7}$$

that is, the set of actions for a state μ, is the set of all joint empirical measures on $\mathbb{X} \times \mathbb{U}$ whose marginal on $\mathbb{X}$ coincides with μ.

We equip the state space $\mathcal{Z}$, and the action sets $U(\mu)$, with the norm $\| \cdot \|_{TV}$ (see (5)) .

One can show that Bayraktar et al. (2025); Bayraktar and Kara (2024) the empirical distributions of the states of agents μ_t, and of the joint state and actions Θ_t define a controlled Markov chain such that

$$\begin{aligned} &Pr(\mu_{t+1} \in B | \mu_t, \dots \mu_0, \Theta_t, \dots, \Theta_0) = Pr(\mu_{t+1} \in B | \mu_t, \Theta_t) \\ &:= \eta(B | \mu_t, \Theta_t) \end{aligned} \tag{8}$$

where $\eta(\cdot|\mu, \Theta) \in \mathcal{P}(\mathcal{P}_N(\mathbb{X}))$ is the transition kernel of the centralized measure valued MDP, which is induced by the dynamics of the team problem.

We define the stage-wise cost function $k(\mu, \Theta)$ by

$$k(\mu, \Theta) := \int c(x, u, \mu) \Theta(du, dx) = \frac{1}{N} \sum_{i=1}^{N} c(x^i, u^i, \mu). \tag{9}$$

Thus, we have an MDP with state space $\mathcal{Z}$, action space $\cup_{\mu \in \mathcal{Z}} U(\mu)$, transition kernel η and the stage-wise cost function k.

We define the set of admissible policies for this measured valued MDP as a sequence of functions $g = \{g_0, g_1, g_2, \dots\}$ such that at every time t, g_t is measurable with respect to the σ-algebra generated by the information variables

$$\hat{I}_t = \{\mu_0, \dots, \mu_t, \Theta_0, \dots, \Theta_{t-1}\}.$$

We denote the set of all admissible control policies by G for the measure valued MDP.

In particular, we define the infinite horizon discounted expected cost function under a policy g by

$$K_\beta^N(\mu_0, g) = E_{\mu_0}^{\eta} \left[\sum_{t=0}^{\infty} \beta^t k(\mu_t, \Theta_t) \right].$$

We also define the optimal cost by

$$K_\beta^{N,*}(\mu_0) = \inf_{g \in G} K_\beta^N(\mu_0, g). \tag{10}$$

The following result shows that this formulation is without loss of optimality:

Theorem 1 (Bayraktar et al. (2025)) *Under Assumption 1.1, for any $\mathbf{x}_0$ that satisfies $\mu_{\mathbf{x}_0} = \mu_0$, that is for any $\mathbf{x}_0$ with distribution μ_0, we have that*

i.)

$$K_\beta^{N,*}(\mu_0) = J_\beta^{N,*}(\mathbf{x}_0).$$

ii.) There exists a stationary and Markov optimal policy g^* for the measure valued MDP, and using g^*, every agent can construct a policy $\gamma^i : \mathbb{X} \times \mathcal{P}_N(\mathbb{X}) \to \mathbb{U}$ such that for $\gamma := \{\gamma^1, \gamma^2, \ldots, \gamma^N\}$, we have that

$$J_\beta^N(\mathbf{x_0}, \gamma) = J_\beta^{N,*}(\mathbf{x_0}).$$

That is, the policy obtained from the measure valued formulation attains the optimal performance for the original team control problem.

1.5 Mean-field Limit Problem

We now introduce the control problem for infinite population teams, i.e. for $N \to \infty$. For some agent $i \in \mathbb{N}$, we define the dynamics as

$$x_{t+1}^i = f(x_t^i, u_t^i, \mu_t^i, w_t^i)$$

where $x_0 \sim \mu_0$ and $\mu_t^i = \mathcal{L}(X_t^i)$ is the law of the state at time t. The agent tries to minimize the following cost function:

$$J_\beta^\infty(\mu_0, \gamma) = \sum_{t=0}^{\infty} \beta^t E\left[c(X_t^i, U_t^i, \mu_t^i)\right]$$

where $\gamma = \{\gamma_t\}_t$ is an admissible policy such that γ_t is measurable with respect to the information variables

$$I_t^i = \left\{x_0^i, \ldots, x_t^i, u_0^i, \ldots, u_{t-1}^i, \mu_0^i, \ldots, \mu_t^i\right\}.$$

Note that the agents are no longer correlated and they are indistinguishable. Hence, in what follows we will drop the dependence on i when we refer to the infinite population problem.

The problem is now a single agent control problem; however, the state variable is not Markovian. However, we can reformulate the problem as an MDP by viewing the state variable as the measure valued μ_t.

We let the state space to be $\mathcal{P}(\mathbb{X})$. Different from the measure valued construction we have introduced in Section 1.4, we let the action space to be $\Gamma = \mathcal{P}(\mathbb{U})^{|\mathbb{X}|}$. In particular, an action $\gamma(\cdot|x) \in \Gamma$ for the team is a randomized policy at the agent level. We equip Γ with the product topology, where we use the weak convergence for each coordinate. We note that each action $\gamma(du|x)$ and state $\mu(dx)$ induce a distribution on $\mathbb{X} \times \mathbb{U}$, which we denote by $\Theta(du, dx) = \gamma(du|x)\mu(dx)$.

Recall the notation in (2); at time t, we can use the following stochastic kernel for the dynamics:

$$x_{t+1} \sim \mathcal{T}(\cdot|x_t, u_t, \mu_t)$$

which is induced by the idiosyncratic noise w_t^i. Hence, we can define

$$\mu_{t+1} = F(\mu_t, \gamma_t) := \int \mathcal{T}(\cdot|x, u, \mu_t)\gamma_t(du|x)\mu_t(dx). \tag{11}$$

Note that the dynamics are deterministic for the infinite population measure valued problem. Furthermore, we can define the stage-wise cost function as

$$k(\mu, \gamma) := \int c(x, u, \mu)\gamma(du|x)\mu(dx). \tag{12}$$

Hence, the problem is a deterministic MDP for the measure valued state process μ_t. A policy, say $g : \mathcal{P}(\mathbb{X}) \to \Gamma$ for the measure-valued MDP, can be written as

$$g(\mu) = \gamma(du|x)$$

for some $\mu \in \mathcal{P}(\mathbb{X})$. That is, an agent observes μ and chooses their actions as an agent-level randomized policy $\gamma(du|x)$.

We reintroduce the infinite horizon discounted cost of the agents for the measure valued formulation:

$$K_\beta(\mu_0, g) = \sum_{t=0}^{\infty} \beta^t k(\mu_t, \gamma_t)$$

for some initial mean-field term μ_0 and under some policy g. Furthermore, the optimal policy is denoted by

$$K^*_\beta(\mu_0) = \inf_g K_\beta(\mu_0, g).$$

At a given time t, the pair (x_t, μ_t) can be used as sufficient information for decision making by the agent i. Furthermore, if the model is fully known by the agents, then the mean-field flow μ_t can be perfectly estimated if every agent agrees and follows the same policy $g(\mu)$, since the dynamics of μ_t is deterministic.

We note that for the infinite population control problem, the coordination requirement between the agents may be relaxed, though cannot be fully abandoned in general (see Section 1.6). In particular, if the agents agree on a common policy $g(\mu) = \gamma(du|x, \mu)$, then for the execution of this policy, no coordination or communication is needed since every agent can estimate the mean-field term μ_t independently and perfectly. Furthermore, every agent can use the same agent-level policy $\gamma(du|x, \mu)$ symmetrically, without any coordination with the other agents.

The following result makes the connection between the finite population and the infinite population control problem rigorous Motte and Pham (2022); Bäuerle (2023); Bayraktar and Kara (2024).

Assumption 1.1 *i. For the transition kernel $\mathcal{T}(\cdot|x, u, \mu)$ (see (2))*

$$\|\mathcal{T}(\cdot|x, u, \mu) - \mathcal{T}(\cdot|x, u, \mu')\| \leq K_f\|\mu - \mu'\|$$

for some $K_f < \infty$, for each x, u and for every $\mu, \mu' \in \mathcal{P}(\mathbb{X})$.

ii c is Lipschitz in μ such that

$$|c(x, u, \mu) - c(x, u, \mu')| \leq K_c\|\mu - \mu'\|$$

for some $K_c < \infty$.

Theorem 2 *Under Assumption 1.1, the following holds:*

i. For any $\mu_0^N \to \mu_0$,

$$\lim_{N\to\infty} K_\beta^{N,*}(\mu_0^N) = K_\beta^{\infty,*}(\mu_0).$$

That is, the optimal value function of the finite population control problem converges to that of the infinite population control problem as $N \to \infty$.

ii. Suppose each agent solves the infinite population control problem given in (11) and (12), and constructs their policies, say

$$g_\infty(\mu) = \gamma_\infty(du|x,\mu).$$

If they follow the infinite population solution in the finite population control problem, for any $\mu_0^N \to \mu_0$ we then have

$$\lim_{N\to\infty} K_\beta^{N}(\mu_0^N, g_\infty) = K_\beta^{\infty,*}(\mu_0).$$

That is, the symmetric policy constructed using the infinite population problem is near optimal for finite but sufficiently large populations.

Remark 3 *The result has significant implications for the computational challenges we have mentioned earlier. Firstly, the second part of the result states that if the number of agents is large enough, then the symmetric policy obtained from the limit problem is near optimal. Hence, the agents can use symmetric policies without coordination, solving their control problems as long as they have access to the mean-field term and their local state. Secondly, note that the flow of the mean-field term μ_t (11) is deterministic if there is no common noise affecting the dynamics. Thus, agents can estimate the marginal distribution of their local state variables x_t^i, without observing the mean-field term if they know the dynamics. In particular, without the common noise, the local state of the agents and the initial mean-field term μ_0 are sufficient information for near optimality.*

However, as we will see in what follows, to achieve near optimal performance, agents must agree on a particular policy $g(\mu) = \gamma(du|x,\mu)\mu(dx)$. In particular, if the optimal infinite population policy is not unique, and the agents apply different optimal policies without coordination, the results of the previous results might fail. Hence, coordination cannot be fully ignored.

1.6 Limitations of Full Decentralization

We have argued in the previous section that the team control problem can be solved near optimally by using the infinite population control solution. Furthermore, if the agents agree on the application of a common optimal policy, the resulting team policy can be executed independently in a decentralized way and achieves near-optimal performance.

The following example shows that if the agents do not coordinate on which policy to follow, i.e. if they are fully decentralized, then the resulting team policy will not achieve the desired outcome.

Example 2 *Consider a team control problem with infinite population where $\mathbb{X} = \mathbb{U} = \{0,1\}$. The stage wise cost function of the agents is defined as*

$$c(x,u,\mu) = \begin{cases} \|\mu - \bar{\mu}_1\| & \text{if } \mu(0) \le \frac{3}{4} \\ \|\mu - \bar{\mu}_2\| & \text{otherwise} \end{cases}$$

where

$$\bar{\mu}_1 = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$$
$$\bar{\mu}_2 = \delta_0.$$

In words, the state distribution should be either be distributed equally over the state space $\{0,1\}$ *or it should fully concentrate in state* 0 *for minimal stage-wise cost. One can check that the cost function satisfies Assumption 1.1 for some* $K_c < \infty$ *(e.g.* $K_c = 1$*). For the dynamics we assume a deterministic model such that*

$$x_{t+1} = u_t.$$

In words, the action of an agent purely determines the next state of the same agent. The goal of the agents is to minimize

$$K_\beta(\mu_0, g) = \limsup_{N\to\infty} \sum_{t=0}^{\infty} \beta^t E[\frac{1}{N}\sum_{i=1}^{N} c(x_t^i, u_t^i, \mu_{\mathbf{x_t}})]$$

where the initial distribution is given by $\mu_0 = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$.

It is easy to see that there are two possible optimal policies for the agents $g_1(\mu) = \gamma_1(du|x,\mu)\mu(dx)$ *and* $g_2(\mu) = \gamma_2(du|x,\mu)\mu(dx)$ *where*

$$\gamma_1(\cdot|x) = \frac{1}{2}\delta_0(\cdot) + \frac{1}{2}\delta_1(\cdot)$$
$$\gamma_2(\cdot|x) = \delta_0(\cdot).$$

If all the agents coordinate and apply either g_1 *or* g_2 *all together, the realized costs will be* 0*, i.e.*

$$K_\beta(\mu_0, g_1) = K_\beta(\mu_0, g_2) = 0.$$

However, if the agents do not coordinate and pick their policies from g_1, g_2 *randomly, the cost incurred will be strictly greater than* 0*. For example, assume that any given agent decides to use* g_1 *with probability* 0.5 *and the policy* g_2 *with probability* 0.5*. Then the resulting policy, say* $\hat{g}$ *will be such that*

$$\hat{g}(\mu) = \left(\frac{1}{2}\gamma_1(du|x) + \frac{1}{2}\gamma_2(du|x)\right)\mu(dx)$$

Thus, at every time step $t \geq 1$*,* $\frac{1}{4}$ *of the agents will be in state* 1 *and* $\frac{3}{4}$ *of the agents will be in state* 0*, hence the total accumulated cost of the resulting policy* $\hat{g}$ *will be*

$$K_\beta(\mu_0, \hat{g}) = \sum_{t=1}^{\infty} \beta^t \frac{1}{4} = \frac{\beta}{1-4\beta} > 0.$$

Thus, we see that if the optimal policy for the mean-field control problem is not unique, the agents cannot follow fully decentralized policies, and they need to coordinate at some level. For this problem, if they agree initially on which policy to follow, then no other communication is needed afterwards for the execution of the decided policy. Nonetheless, an initial agreement and coordination is needed to achieve the optimal performance.

We note that the issue with the previous example results from the fact that the optimal policy is not unique. If the optimal policy can be guaranteed to be unique, then the agents can act fully independently.

2. Learning for Mean-field Control with Linear Approximations

We have seen in the previous sections that in general there are limitations for full decentralization, and that a certain level of coordination is required for optimal or near optimal performance during control. In this section, we will study the learning problem in which neither the agents nor the coordinator know the dynamics and aims to learn the model or optimal decision strategies from the data.

We have observed that the limit problem introduced in Section 1.5 can be seen as a deterministic centralized control problem. In particular, if the model is known, and once it is coordinated which control strategy to follow, the agents do not need further communication or coordination to execute the optimal control. Each agent can simply apply an open-loop policy using only their local state information, and the mean-field term can be estimated perfectly, if every agent is following the same policy. However, to estimate the deterministic mean-field flow μ_t, the model must be known. For problems where the model is not fully known, the open-loop policies will not be applicable.

Our goal in this section is to present various learning algorithms to learn the dynamics and cost model of the control problem. We will first focus on the idealized scenario, where we assume that there exist infinitely many agents on the team. For this case, we provide two methods; (i) the first one where a coordinator has access to all information of every agent, and decides on the exploration policy, and (ii) the second one where each agent learns the model on their own by tracking their local state and the mean-field term. However, the agents need to coordinate for the exploration policy through a common randomness variable to induce stochastic dynamics for better exploration. Next, we study the realistic setting where the team has large but finitely many agents. For this case, we only consider an independent learning method where the agents learn the model on their own using their local information variables.

Before we present our learning algorithms, we note that the space $\mathcal{P}(\mathbb{X})$ is uncountable even under the assumption that $\mathbb{X}$ is finite. Therefore, we will focus on finite representations of the cost function $c(x,u,\mu)$ and the kernel $\mathcal{T}(\cdot|x,u,\mu)$. In particular, we will try to learn the functions of the following form

$$\begin{aligned} c(x,u,\mu) &= \mathbf{\Phi}_{(x,u)}^{\intercal}(\mu)\theta_{(x,u)} \\ \mathcal{T}(\cdot|x,u,\mu) &= \mathbf{\Phi}_{(x,u)}^{\intercal}(\mu)\mathbf{Q}_{(x,u)}(\cdot) \end{aligned} \tag{13}$$

where $\mathbf{\Phi}_{(x,u)}(\mu) = [\Phi_{(x,u)}^1(\mu), \dots, \Phi_{(x,u)}^d(\mu)]^{\intercal}$, for a set of linearly independent functions $\Phi_{(x,u)}^j(\mu) : \mathcal{P}(\mathbb{X}) \to \mathbb{R}$ for each pair (x,u), for some $d < \infty$. We assume that the basis functions $\mathbf{\Phi}_{(x,u)}(\mu)$ are known and the goal is to learn the parameters $\theta_{(x,u)}$ and $\mathbf{Q}_{(x,u)}(\cdot)$. We assume $\theta_{(x,u)} \in \mathbb{R}^d$, and $\mathbf{Q}_{(x,u)}(\cdot) = [Q_{(x,u)}^1(\cdot), \dots, Q_{(x,u)}^d(\cdot)]$ is a vector of unknown signed measures on $\mathbb{X}$.

In what follows, we will assume that the basis functions, $\Phi_{(x,u)}^j(\cdot)$ are uniformly bounded. Note that this is without loss of generality.

Assumption 2.1 *We assume that*

$$\|\Phi_{(x,u)}^j(\cdot)\|_\infty \leq 1$$

for every (x,u) pair, and for all $j \in \{1,\dots,d\}$.

For the rest of the paper, we will use $\theta_{(x,u)}$ and $\theta(x,u)$ interchangeably; similarly we will use $\mathbf{Q}_{(x,u)}$ and $\mathbf{Q}(x,u)$ interchangeably.

Remark 4 *We note that we* **do not** *assume that the model and the cost function have the linear form given in (13). However, we will aim to learn and estimate models among the class of linear functions presented in (13). We will later analyze error bounds for the case where the actual model is not linear and thus the learned model does not perfectly match the true model and study the performance loss when we apply policies that are learned for the linear model.*

2.1 Coordinated Learning with Linear Function Approximation for Infinitely Many Players

In this section, we will consider an idealized scenario, where there are infinitely many agents, and a coordinator learns the model by linear function approximation.

Data collection. For this section, we assume that there exists a training set T that consists of a time sequence of length M. The training set is assumed to be coming from an arbitrary sequence of data. The data at each time stage contains

$$x^i, u^i, X_1^i, c(x^i, u^i, \mu), \mu$$

for all the agents present in the team, $i \in \{1, \dots, N, \dots\}$, where the agents' states are distributed according to μ at the given time step. That is, every data point includes the current state and action, the one-step ahead state, the stage-wise cost realization, and the mean-field term for every agent. Furthermore, we assume the ideal scenario where there are infinitely many agents. Hence, at every time step, the coordinator has access to infinitely many data points where the spaces $\mathbb{X}, \mathbb{U}$ are finite. The coordinator then has access to infinitely many sample transitions observed under (x,u,μ), and thus, the kernel $\mathcal{T}(\cdot|x,u,\mu)$ can be perfectly estimated for every x such that $\mu(x) > 0$ and $\gamma(u|x) > 0$, via empirical measures. Here, γ represents the exploration policy of the agents. We assume the following:

Assumption 2.2 *For any $x \in \mathbb{X}$, the exploration policy for every agent puts positive probability on to every control action such that*

$$\gamma(u|x) > 0, \textit{ for all } (x,u) \in \mathbb{X} \times \mathbb{U}.$$

We define the following sets for which the model and the cost functions can be learned perfectly within the training data: let $x \in \mathbb{X}$, we define

$$P_x := \{\mu \in T : \mu(x) > 0\}. \tag{14}$$

$P_x \subset \mathcal{P}(\mathbb{X})$ denotes the set of probability measures which assign positive measure to a particular state $x \in \mathbb{X}$ that are also in the training data for the mean-field terms. In particular, for a given (x,u) pair, the kernel $\mathcal{T}(\cdot|x,u,\mu)$ and the cost $c(x,u,\mu)$ can be learned perfectly for every $\mu \in P_x$ with Assumption 2.2.

For a given $x \in \mathbb{X}$, we denote by M_x the number of mean-field terms within the set P_x (see (14)). For every $(x,u) \in \mathbb{X} \times \mathbb{U}$ pair, the coordinator aims to find $\theta_{(x,u)}$ and $\mathbf{Q}_{(x,u)}$ such that

$$\frac{1}{M_x} \sum_{j=1}^{M_x} \left| c(x,u,\mu_j) - \mathbf{\Phi}^{\intercal}_{(\mathbf{x},\mathbf{u})}(\mu_j)\theta_{(\mathbf{x},\mathbf{u})} \right|^2$$

$$\frac{1}{M_x}\sum_{j=1}^{M_x}\left\|\mathcal{T}(\cdot|x,u,\mu_j)-\mathbf{\Phi}^{\intercal}_{(\mathbf{x},\mathbf{u})}(\mu_j)\mathbf{Q}_{(\mathbf{x},\mathbf{u})}(\cdot)\right\|$$

is minimized.

The least squares linear models can be estimated in closed form for $\theta_{(x,u)}$ and $\mathbf{Q}_{(x,u)}(\cdot)$ using the training data. We define the following vector and matrices to present the closed form solution in a more compact form: for each $(x,u)\in\mathbb{X}\times\mathbb{U}$ we introduce $\mathbf{b}_{(x,u)}\in\mathbb{R}^{M_x}$, and $\mathbf{d}_{(x,u)}\in\mathbb{R}^{M_x\times|\mathbb{X}|}$,

$$\mathbf{b}_{(x,u)}=\begin{bmatrix}c(x,u,\mu_1)\\ c(x,u,\mu_2)\\ \vdots\\ c(x,u,\mu_{M_x})\end{bmatrix},\quad \mathbf{d}_{(x,u)}=\begin{bmatrix}\mathcal{T}(x^1|x,u,\mu_1) & \mathcal{T}(x^2|x,u,\mu_1) & \dots & \mathcal{T}(x^{|\mathbb{X}|}|x,u,\mu_1)\\ \mathcal{T}(x^1|x,u,\mu_2) & \mathcal{T}(x^2|x,u,\mu_2) & \dots & \mathcal{T}(x^{|\mathbb{X}|}|x,u,\mu_2)\\ \vdots & & & \\ \mathcal{T}(x^1|x,u,\mu_{M_x}) & \mathcal{T}(x^2|x,u,\mu_{M_x}) & \dots & \mathcal{T}(x^{|\mathbb{X}|}|x,u,\mu_{M_x})\end{bmatrix}. \tag{15}$$

Furthermore, we also define $\mathbf{A}_{(x,u)}\in\mathbb{R}^{d\times M_x}$

$$\mathbf{A}_{(x,u)}=\left[\mathbf{\Phi}_{(x,u)}(\mu_1),\dots,\mathbf{\Phi}_{(x,u)}(\mu_{M_x})\right]. \tag{16}$$

Assuming that $\mathbf{A}_{(x,u)}$ has linearly independent columns, i.e. $\mathbf{\Phi}_{(x,u)}(\mu_i)$ and $\mathbf{\Phi}_{(x,u)}(\mu_j)$ are linearly independent for $\mu_i\neq\mu_j$, the estimates for $\theta_{(x,u)}$ and $\mathbf{Q}_{(x,u)}$ can be written as follows

$$\begin{aligned}\theta_{(x,u)}&=\left(\mathbf{A}^{\intercal}_{(x,u)}\mathbf{A}_{(x,u)}\right)^{-1}\mathbf{A}^{\intercal}_{(x,u)}\mathbf{b}_{(x,u)}\\ \mathbf{Q}_{(x,u)}&=\left(\mathbf{A}^{\intercal}_{(x,u)}\mathbf{A}_{(x,u)}\right)^{-1}\mathbf{A}^{\intercal}_{(x,u)}\mathbf{d}_{(x,u)}.\end{aligned} \tag{17}$$

Note that above, each row of $\mathbf{Q}_{(x,u)}$ represents a signed measure on $\mathbb{X}$.

2.2 Independent Learning with Linear Function Approximation for Infinitely Many Players

In this section, we will introduce a learning method where the agents perform independent learning to some extent. Here, rather than using a training set, we will focus on an online learning algorithm where at every time step, agent i observes $x^i,u^i,X_1^i,c(x^i,u^i,\mu),\mu$. That is, each agent has access to their local state, action, cost realizations, one-step ahead state, and the mean-field term. However, they do not have access to local information about the other agents.

We first argue that full decentralization is usually not possible in the context of learning either. Recall that the mean-field flow is deterministic if every agent follows the same independently randomized agent-level policy. Furthermore, the flow of the mean-field terms remains deterministic even when the agents choose different exploration policies if the randomization is independent. To see this, assume that each agent picks some policy $\gamma_w(du|x)$ randomly by choosing $w\in\mathbb{W}$ from some arbitrary distribution, where the mapping $w\to\gamma_w(du|x)$ is predetermined. If the agents pick $w\in\mathbb{W}$ independently, the mean-field dynamics is given by

$$\mu_{t+1}(\cdot)=\int\mathcal{T}(\cdot|x,u,\mu_t)\gamma_w(du|x)P_w(dw)\mu_t(dx)$$

where $P_w(dw)$ is the distribution by which the agents perform their independent randomization for the policy selection. Hence, the mean-field term dynamics follow a deterministic flow. Note

that for the above example, for simplicity, we assume that the agents pick according to the same distribution. In general, even if the agents follow different distributions for $w \in \mathbb{W}$, the dynamics of the mean-field flow would remain deterministic according to a mixture distribution.

Deterministic behavior might cause poor exploration performance. There might be cases where the mean-field flow gets stuck at a fixed distribution without learning or exploring the 'important' parts of the space $\mathcal{P}(\mathbb{X})$ sufficiently. To overcome this issue, and to make sure that the system is stirred sufficiently well during the exploration, one option is to introduce a common randomness for the selection of the exploration policies. In particular, each agent follows a randomized policy $\gamma_w(du|x)$ where the common randomness $w \in \mathbb{W}$ is mutual information. Then the dynamics of the mean-field flow can be written as

$$\mu_{t+1}(\cdot) = F(\mu_t, w) := \int \mathcal{T}(\cdot|x, u, \mu_t)\gamma_w(du|x)\mu_t(dx). \tag{18}$$

The common noise variable ensures a level of coordination among the agents. However, this is still a significant relaxation compared to full coordination where the agents share their full state or control data. In this section, we show that agents can construct independent learning iterates that converge by coordinating through an arbitrary common source of randomness.

We assume the following for the mean-field flow during the exploration:

Assumption 2.3 *Consider the Markov chain $\{\mu_t\}_t \subset \mathcal{P}(\mathbb{X})$ whose dynamics are given by (18) We assume that μ_t has geometric ergodicity with a unique invariant probability measure $P(\cdot) \in \mathcal{P}(\mathcal{P}(\mathbb{X}))$ such that*

$$\|Pr(\mu_t \in \cdot) - P(\cdot)\|_{TV} \leq K\rho^t$$

for some $K < \infty$ and some $\rho < 1$.

Remark 5 *We can establish some sufficient conditions on the transition kernel of the system to test the ergodicity. We note that ((Hernandez-Lerma, 2012, p 56, 3.1.1)) a sufficient condition is the following: there exists a mean-field state, say $\mu^* \in \mathcal{P}(\mathbb{X})$ such that*

$$Pr\left(w : \int \mathcal{T}(\cdot|x, u, \mu)\gamma_w(du|x)\mu(dx) = \mu^*(\cdot)\right) > 0, \text{ for all } \mu \in \mathcal{P}(\mathbb{X}). \tag{19}$$

That is, we need to be able to find a set of common noise realizations whose induced randomized exploration policies can take the state distribution to μ^* independent of the starting distribution μ.

The assumption stated in this form indicates that the condition is of the stochastic reachability or controllability type. It requires that from any initial distribution μ, there exists a control policy that can steer the distribution of the system to some target measure μ^*. We also note that the above can be generalized to a k-step transition requirement. Analyzing this stochastic controllability behavior for the mean field systems is beyond the scope of the current paper, however, we give some examples in what follows.

We look further into (19). We fix some $x_1 \in \mathbb{X}$, and we define the following values:

$$U(x, \mu) = \max_u \mathcal{T}(x_1|x, u, \mu)$$
$$L(x, \mu) = \min_u \mathcal{T}(x_1|x, u, \mu).$$

Note that $\mu^*(x_1)$ is the average of the probabilities $\mathcal{T}^\gamma(x_1|x,\mu) := \int \mathcal{T}(x_1|x,u,\mu)\gamma_w(du|x)$ under the measure μ for the x values. Furthermore, by selecting an appropriate randomized policy, we can control these values in the interval

$$I(x,\mu) := [L(x,\mu), U(x,\mu)]\,.$$

Thus, if the intersection of these intervals is nonempty, i.e. if

$$\bigcap_{x,\mu} I(x,\mu) \neq \emptyset \tag{20}$$

then one can set $\mu^*(x_1)$ to be a value in this intersection independent of μ. By doing this for all x_1, we can set a reachable μ^* from any $\mu \in \mathcal{P}(\mathbb{X})$. As a result, if (20) holds for every x_1, then (19), and thus Assumption 2.3 can be shown to hold.

A somewhat restrictive example for (20) is the following: assume that there exists a control action u^* which can reset the state to some x^* from any state x and any mean-field term μ, that is

$$\mathcal{T}(x^*|x,u^*,\mu) = 1 \text{ for all } x,\mu.$$

This means that $1 \in I(x,\mu)$ for all x,μ when $x_1 = x^*$, and $0 \in I(x,\mu)$ for all x,μ when $x_1 \neq x^*$. Hence, $\mu_1(\cdot) = \delta_{x^*}(\cdot)$ can be reached from any starting point by applying the policy $\gamma(x) = u^*$ for all x. If this policy is among the set of exploration policies, the ergodicity assumption for the mean-field flow would be satisfied.

We note again that a general result would require more in depth analysis, however, the above gives some idea on the implications of this assumption on the controllability of the mean-field model.

We now define the trained measures, $P_x \in \mathcal{P}(\mathcal{P}(\mathbb{X}))$ for each (x,u) pair based on the invariant measure for the mean-field flow. Under Assumption 2.2, that is assuming $Pr(u|x) = \int \gamma_w(u|x)P_w(dw) > 0$ for all (x,u) pairs, we can write

$$\begin{aligned} P(\mu \in A|x,u) &= \frac{Pr(x,u,\mu \in A)}{Pr(x,u)} \\ &= \frac{\int_{\mu\in A}\int_w \gamma_w(u|x)P_w(dw)\mu(x)P(d\mu)}{\int_{\mu\in\mathcal{P}(\mathbb{X})}\int_w \gamma_w(u|x)P_w(dw)\mu(x)P(d\mu)} \\ &= \frac{\int_{\mu\in A}\mu(x)P(d\mu)}{\int_{\mu\in\mathcal{P}(\mathbb{X})}\mu(x)P(d\mu)} =: P_x(\mu \in A) \end{aligned} \tag{21}$$

Note that the trained sets of mean-field terms are independent of the control action u, as the exploration policies are independent of the mean-field terms given the state x. These sets have similar implications as the sets defined in (14). In particular, they indicate for which mean-field terms, one can estimate the kernel $\mathcal{T}(\cdot|x,u,\mu)$ and the cost function $c(x,u,\mu)$ via the training process.

We now summarize the algorithm used for each agent. We drop the dependence on agent identity i, and summarize the steps for a generic agent. At every time step t, every agent performs the following steps:

- Observe the common randomness w given by the coordinator, and pick an action such that $u_t \sim \gamma_w(\cdot|x_t)$

- Collect $x_t, u_t, x_{t+1}, \mu_t, c$ where $c = c(x_t, u_t, \mu_t)$
- For all $(x, u) \in \mathbb{X} \times \mathbb{U}$

$$\theta_{t+1}(x,u) = \theta_t(x,u) + \alpha_t(x,u)\mathbf{\Phi}(x,u,\mu_t)\left[c - \mathbf{\Phi}^\intercal(x,u,\mu_t)\theta_t(x,u)\right] \tag{22}$$

- Note that the signed measure vector $\mathbf{Q}_t(\cdot|x,u)$ consists d signed measures defined on $\mathbb{X}$, we denote by $\mathbf{Q}_t^j(x,u)$ the vector values of $\mathbf{Q}_t(x^j|x,u)$ for $x^j \in \mathbb{X}$ where $j \in \{1\ldots,|\mathbb{X}|\}$. For all (x,u) and $j \in \{1\ldots,|\mathbb{X}|\}$

$$\mathbf{Q}_{t+1}^j(x,u) = \mathbf{Q}_t^j(x,u) + \alpha_t(x,u)\mathbf{\Phi}(x,u,\mu_t)\left[\mathbb{1}_{\{x_{t+1}=x^j\}} - \mathbf{\Phi}^\intercal(x,u,\mu_t)\mathbf{Q}_t^j(x,u)\right]. \tag{23}$$

We next show that the above algorithm converges if the learning rates are chosen properly. To show the convergence, we first present a convergence result for stochastic gradient descent algorithms with quadratic cost, where the error is Markov and stationary. We note that similar results have been established in the literature for the stochastic gradient iterations under Markovian noise processes under various assumptions; however, verifying most of these assumptions, such as the boundedness of the gradient, the boundedness of the iterates, or uniformly bounded variance, is not straightforward. Hence, we provide a proof in the appendix for completeness.

Proposition 2.1 *Let $\{S_t\} \subset \mathbb{S}$ denote a Markov chain with the invariant probability measure $\pi(\cdot)$ where $\mathbb{S}$ is a standard Borel space. We assume that $\{S_t\}$ has geometric ergodicity such that $\|Pr(S_t \in \cdot) - \pi(\cdot)\|_{TV} \leq K\rho^t$ for some $K < \infty$ and some $\rho < 1$. Let $g(s,v)$ be such that*

$$g(s,v) = \left(k(s)^\intercal v - h(s)\right)^2$$

for some $k : \mathbb{S} \to \mathbb{R}^d$, $h : \mathbb{S} \to \mathbb{R}$ and for $v \in \mathbb{R}^d$. We assume that k, h are uniformly bounded. We denote by

$$G_t(v) = E[g(S_t, v)]$$
$$G(v) = \int g(s,v)\pi(ds).$$

Consider the iterations

$$v_{t+1} = v_t - \alpha_t \nabla g(S_t, v_t)$$

where the gradient is with respect to v_t. If the learning rates are such that $\sum_t \alpha_t = \infty$ and $\sum_t \alpha_t^2 < \infty$ with probability one, we then have that $G_t(v_t) \to \min_v G(v) = G(v^*)$ almost surely.

Proof The proof can be found in Appendix A. ■

Corollary 6 *Let Assumption 2.3 and Assumption 2.2 hold and let the learning rates be chosen such that $\alpha_t(x,u) = 0$ unless $(X_t, U_t) = (x,u)$. Furthermore, $\sum_t \alpha_t(x,u) = \infty$ and $\sum_t \alpha_t^2(x,u) < \infty$*

with probability one for all $(x,u) \in \mathbb{X} \times \mathbb{U}$. Then, the iterations given in (22) and (23) converge with probability 1. Furthermore, the limit points, say $\theta^*(x,u)$ and $\mathbf{Q}^*_{(x,u)}(\cdot)$ are such that

$$\theta^*(x,u) = \underset{\theta(x,u)\in\mathbb{R}^d}{\arg\min} \int \left|c(x,u,\mu) - \mathbf{\Phi}^\intercal(x,u,\mu)\theta(x,u)\right|^2 P_x(d\mu)$$

$$\mathbf{Q}^{*,j}(x,u) = \underset{\mathbf{Q}^j(x,u)\in\mathbb{R}^d}{\arg\min} \int \left|\mathcal{T}(j|x,u,\mu) - \mathbf{\Phi}^\intercal_{x,u}(\mu)\mathbf{Q}^j(x,u)\right|^2 P_x(d\mu)$$

for every (x,u) pair and for every j, where $\mathbf{Q}^{j,*}(x,u)$ is the jth column of $\mathbf{Q}^*_{(x,u)}(\cdot)$. Furthermore, $P_x(\cdot)$ denotes the trained set based on the invariant measure of the mean-field flow under the exploration policy with common randomness (see (21)).

Proof We define the following stopping times

$$\tau_{k+1} = \min\{t > \tau_k : (X_t, U_t) = (x,u)\}$$

such that τ_k indicates the k-th time the (x,u) pair is visited.

For the iterations (22), Proposition 2.1 applies such that for each (x,u), $v_k \equiv \theta_{\tau_k}(x,u)$, $k(\mu) \equiv \mathbf{\Phi}^\intercal(x,u,\mu)$ and $h(\mu) \equiv c(x,u,\mu)$ and finally the noise process $s_k \equiv \mu_{\tau_k}$. Note that $\mathbf{\Phi}$ and c are assumed to be uniformly bounded which also agrees with the assumptions in Propositions 2.1. Furthermore, with the strong Markov property, μ_{τ_k} is also a Markov chain which is sampled when the state-action pair is (x,u). Thus, the invariant measure for the sampled process is P_x as defined in (21).

For the iterations (23), Proposition 2.1 applies such that for each (x,u) and each j, $v_t \equiv \mathbf{Q}^j_{\tau_k}(x,u)$, $k(\mu) \equiv \mathbf{\Phi}^\intercal(x,u,\mu)$ and $h(\mu) \equiv \mathbb{1}_{\{X_1 = x^j\}}$. We note that $X_1 = f(x,u,\mu,w)$ (see (1)) where w is the i.i.d. noise for the dynamics of agents. Thus, the noise process for iterations (23) can be taken to be the joint process (μ_t, w_t) where μ_t is an ergodic Markov process, and w_t is an i.i.d. process. In particular, for every (x,u) pair and for every x^j, if we consider the expectation over (μ, w) where $\mu \sim P_x(\cdot)$, we get

$$E\left[\mathbb{1}_{\{X_1 = x^j\}}\right] = E\left[\mathbb{1}_{\{f(x,u,\mu,w)=x^j\}}\right] = \int \mathcal{T}(x^j|x,u,\mu)P_x(d\mu)$$

for every $x^j \in \mathbb{X}$.

The algorithm in (23) minimizes

$$\int \left|\mathbb{1}_{\{f(x,u,\mu,w)=x^j\}} - \mathbf{\Phi}^\intercal_{x,u}(\mu)\mathbf{Q}^j(x,u)\right|^2 P_x(d\mu)P_w(dw)$$

for each j where P_w is the distribution of the noise term. We can then open up the above term to write:

$$\begin{aligned}
&\underset{\mathbf{Q}^j(x,u)}{\arg\min} \int \left|\mathbb{1}_{\{f(x,u,\mu,w)=x^j\}} - \mathbf{\Phi}^\intercal_{x,u}(\mu)\mathbf{Q}^j(x,u)\right|^2 P_x(d\mu)P_w(dw) \\
&= \underset{\mathbf{Q}^j(x,u)}{\arg\min} \int (\mathbb{1}_{\{f(x,u,\mu,w)=x^j\}})^2 - 2\mathbb{1}_{\{f(x,u,\mu,w)=x^j\}}\mathbf{\Phi}^\intercal_{x,u}(\mu)\mathbf{Q}^j(x,u) \\
&\qquad\qquad + (\mathbf{\Phi}^\intercal_{x,u}(\mu)\mathbf{Q}^j(x,u))^2 P_x(d\mu)P_w(dw)
\end{aligned}$$

$$= \arg\min_{\mathbf{Q}^j(x,u)} \int -2\mathbb{1}_{\{f(x,u,\mu,w)=x^j\}} \mathbf{\Phi}^{\intercal}_{x,u}(\mu)\mathbf{Q}^j(x,u) + (\mathbf{\Phi}^{\intercal}_{x,u}(\mu)\mathbf{Q}^j(x,u))^2 P_x(d\mu)P_w(dw)$$

$$= \arg\min_{\mathbf{Q}^j(x,u)} \int -2\mathcal{T}(x^j|x,u,\mu)\mathbf{\Phi}^{\intercal}_{x,u}(\mu)\mathbf{Q}^j(x,u) + (\mathbf{\Phi}^{\intercal}_{x,u}(\mu)\mathbf{Q}^j(x,u))^2 P_x(d\mu)$$

$$= \arg\min_{\mathbf{Q}^j(x,u)} \int (\mathcal{T}(x^j|x,u,\mu))^2 - 2\mathcal{T}(x^j|x,u,\mu)\mathbf{\Phi}^{\intercal}_{x,u}(\mu)\mathbf{Q}^j(x,u) + (\mathbf{\Phi}^{\intercal}_{x,u}(\mu)\mathbf{Q}^j(x,u))^2 P_x(d\mu)$$

$$= \arg\min_{\mathbf{Q}^j(x,u)} \int \left(\mathcal{T}(x^j|x,u,\mu)) - \mathbf{\Phi}^{\intercal}_{x,u}(\mu)\mathbf{Q}^j(x,u)\right)^2 P_x(d\mu).$$

Hence, the algorithm minimizes

$$\int \left(\mathcal{T}(x^j|x,u,\mu)) - \mathbf{\Phi}^{\intercal}_{x,u}(\mu)\mathbf{Q}^j(x,u)\right)^2 P_x(d\mu)$$

for each j. ■

2.3 Learning for Finitely Many Players

In this section, we will study the more realistic scenario in which the number of agents is large but finite. The learning methods presented in the previous sections have focused on the ideal case where the system has infinitely many players. Although the setting with the infinitely many agents helps us to fix the ideas for the learning in the mean-field control setup, we should note that it is only an artificial setup, and the infinite population setup is only used as an approximation for large population control problems. Hence, we need to study the actual setup for which the limit problem is argued to be a well approximation, that is the problem with very large but finitely many agents.

We will apply the independent learning algorithm presented for the infinite population case, and study the performance of the learned solutions for the finitely many player setting. In particular, we will assume that the agents follow the iterations given in (22) and (23). We note, however, that the agents will not need to use common randomness during exploration as the flow of the mean-field term is stochastic for finite populations without common randomness. The method remains valid under common randomness as well; in fact, the common randomness, in general, encourages the exploration of the state space. The method is identical to the one presented in Section 2.2. However, we present the method again since it has some subtle differences.

At every time step t, agent i performs the following steps:

- Pick an action such that $u_t^i \sim \gamma^i(\cdot|x_t^i)$
- Collect $x_t^i, u_t^i, x_{t+1}^i, \mu_t^N, c$ where $c = c(x_t^i, u_t^i, \mu_t^N)$ and $\mu_t^N = \mu_{\mathbf{x_t}}$
- For all $(x,u) \in \mathbb{X} \times \mathbb{U}$

$$\theta_{t+1}(x,u) = \theta_t(x,u) + \alpha_t(x,u)\mathbf{\Phi}_{x,u}(\mu_t^N)\left[c - \mathbf{\Phi}^{\intercal}_{x,u}(\mu_t^N)\theta_t(x,u)\right] \tag{24}$$

- Denoting by $\mathbf{Q}_t^j(x,u)$ the vector values of $\mathbf{Q}_t(x^j|x,u)$ for all $x^j \in \mathbb{X}$ where $j \in \{1\ldots,|\mathbb{X}|\}$. For all (x,u) and $j \in \{1\ldots,|\mathbb{X}|\}$

$$\mathbf{Q}_{t+1}^j(x,u) = \mathbf{Q}_t^j(x,u) + \alpha_t(x,u)\mathbf{\Phi}_{x,u}(\mu_t^N)\left[\mathbb{1}_{\{x_{t+1}^i=x^j\}} - \mathbf{\Phi}^{\intercal}_{x,u}(\mu_t^N)\mathbf{Q}_t^j(x,u)\right] \tag{25}$$

Remark 7 *We note that the iterates θ_t and $\mathbf{Q}_t$ depend on the agent identity i, in this case, as each agent can learn the model independently. Moreover, the learning rates $\alpha_t(x,u)$ and the basis functions $\mathbf{\Phi}_{x,u}$ might depend on the agent identity as well. However, we omit the dependence in the notation to reduce notational clutter.*

Assumption 2.4 *Under the exploration team policy $\gamma(\cdot|\mathbf{x}) = \left[\gamma^1(\cdot|x^1), \ldots, \gamma^N(\cdot|x^N)\right]^\intercal$, the state vector process $\mathbf{x_t} = [x_t^1, \ldots, x_t^N]$ of the agents is irreducible and aperiodic and in particular admits a unique invariant measure, and thus the mean-field flow $\mu_t^N = \mu_{\mathbf{x_t}}$ admits a unique invariant measure, say $P^N(\cdot) \in \mathcal{P}_N(\mathbb{X})$, as well.*

Remark 8 *We note that a sufficient condition for the above assumption to hold is that there exists some $x' \in \mathbb{X}$ such that $\mathcal{T}(x'|x,u,\mu^N) \geq \epsilon > 0$ for any x, u and for any μ^N. In particular, this implies that*

$$Pr\left(\mathbf{X}_{t+1} = [x', \ldots, x']|\mathbf{x}_t, \gamma\right) = \prod_{i=1}^{N} \sum_{u} \mathcal{T}(x'|x_t^i, u, \mu_t^N)\gamma^i(u|x_t^i) \geq \epsilon^N > 0$$

and thus (Hernandez-Lerma, 2012, p 56, 3.1.1) implies that the process $\mathbf{X}_t$ is geometrically ergodic.

The next result shows the convergence of the algorithm. Similar to the previous section, we first define the trained sets of mean-field terms for every (x,u) pair using the stationary distribution of the mean-field terms. We assume that Assumption 2.2 holds for every policy γ^i such that $\gamma^i(u|x) > 0$ for every (x,u) pair. Denoting the invariant distribution of the joint state process by $P(\mathbf{x})$, for some $\mu^N \in \mathcal{P}_N(\mathbb{X})$, for agent i, we can write

$$\begin{aligned}
P(\mu^N|(x^i,u^i) = (x,u)) &= \int_{\mathbf{x}\in\mathbb{X}^N} Pr(\mu^N|\mathbf{x}) P(\mathbf{x}|(x^i,u^i) = (x,u)) \\
&= \int_{\mathbf{x}\in\mathbb{X}^N} \mathbb{1}_{\{\mu^N = \mu_{\mathbf{x}}\}} \frac{\gamma^i(u|x)\mathbb{1}_{\{x=\mathbf{x}[i]\}}}{P(x^i = x, u^i = u)} P(\mathbf{x}) \\
&= \int_{\mathbf{x}\in\mathbb{X}^N} \mathbb{1}_{\{\mu^N = \mu_{\mathbf{x}}\}} \frac{\gamma^i(u|x)\mathbb{1}_{\{x=\mathbf{x}[i]\}}}{\gamma^i(u|x)P(x^i = x)} P(\mathbf{x}) \\
&= \int_{\mathbf{x}\in\mathbb{X}^N} \mathbb{1}_{\{\mu^N = \mu_{\mathbf{x}}\}} \frac{\mathbb{1}_{\{x=\mathbf{x}[i]\}}}{P(x^i = x)} P(\mathbf{x}) \\
&= P(\mu^N|x^i = x) =: P_x^i(\mu^N). \qquad (26)
\end{aligned}$$

Note that the trained measure of mean-fields is independent of the control actions; however, it does depend on the agent identity as the agents follow distinct exploration policies.

Corollary 9 *Let Assumption 2.4 hold, and let Assumption 2.2 hold for each policy γ^i. Assume further that the learning rates of every agent satisfy the assumption of Corollary 6 such that $\sum_t \alpha_t(x,u) = \infty$ and $\sum_t \alpha_t^2(x,u) < \infty$ with probability one for all $(x,u) \in \mathbb{X} \times \mathbb{U}$. Then, the iterations given in (24) and (25) converge with probability one. Furthermore, for agent i, the limit points, say $\theta^*_{(x,u)}$ and $\mathbf{Q}^*_{(x,u)}(\cdot)$ are such that*

$$\theta^*(x,u) = \underset{\theta(x,u)\in\mathbb{R}^d}{\arg\min} \int \left|c(x,u,\mu) - \mathbf{\Phi}^\intercal(x,u,\mu)\theta(x,u)\right|^2 P_x^i(d\mu)$$

$$\mathbf{Q}^{*,j}(x,u) = \underset{\mathbf{Q}^j(x,u)\in\mathbb{R}^d}{\arg\min} \int \left|\mathcal{T}(j|x,u,\mu) - \mathbf{\Phi}^{\intercal}_{x,u}(\mu)\mathbf{Q}^j(x,u)\right|^2 P^i_x(d\mu)$$

for every (x,u) pair and for every j, where $\mathbf{Q}^{j,*}(x,u)$ is the jth column of $\mathbf{Q}^*_{(x,u)}(\cdot)$. Furthermore, $P_x(\cdot)$ denotes the trained set based on the invariant measure of the mean-field flow under the exploration policy with common randomness (see (26)).

Proof The proof is identical to the proof of Corollary 6, and is an application of Proposition 2.1. The only difference is the ergodicity of the mean-field process μ^N_t, which does not require common randomness for exploration policies. ■

3. Uniform Error Bounds for Model Approximation

The learning methods we have presented in Section 2 minimize the L_2 distance between the true model and the linear approximate model, under the probability measure induced by the training data. In particular, denoting the learned parameters for a fixed pair $(x,u) \in \mathbb{X}\times\mathbb{U}$ by $\theta^*_{(x,u)}$, and $\mathbf{Q}^*_{(x,u)}(\cdot)$, we have that

$$\begin{aligned}
\theta^*_{(x,u)} &= \underset{\theta\in\mathbb{R}^d}{\arg\min} \int \left|c(x,u,\mu) - \mathbf{\Phi}^{\intercal}_{(x,u)}(\mu)\theta_{(x,u)}\right|^2 P_x(d\mu) \\
\mathbf{Q}^{*,j}(x,u) &= \underset{\mathbf{Q}^j(x,u)\in\mathbb{R}^d}{\arg\min} \int \left|\mathcal{T}(j|x,u,\mu) - \mathbf{\Phi}^{\intercal}_{(x,u)}(\mu)\mathbf{Q}^j(x,u)\right|^2 P_x(d\mu) \qquad (27)
\end{aligned}$$

for some probability measure $P_x(\cdot)\in\mathcal{P}(\mathbb{X})$. The measure $P_x(\cdot)$ depends on the learning method used.

- For the coordinated learning methods presented in Section 2.1, $P_x(\cdot)$ represents the empirical distribution of the mean-field terms in the training data for which the state x has positive measure (see (14)).
- For the individual learning method presented in Section 2.2 for infinite populations, $P_x(\cdot)$ represents the invariant measure of the mean-field flow under the randomized exploration policies given the state x is observed. See (21).
- Finally, for the individual learning method for finite populations in Section 2.3, P_x depends on the agent identity i, and thus denoted by P^i_x. Similar to the infinite population setting, it represents the invariant measure of the process $\mu_{\mathbf{x_t}}$ conditioned on the event $(x^i = x)$, for agent i where $\mathbf{x_t}$ is the N dimensional vector state of the team of N agents. We note that each agent might have different trained sets of mean-field terms in this setting, since the policies may be distinct.

When the learned policy is executed, the flow of the mean-field is not guaranteed to stay in the support of the training measure $P_x(\cdot)$. Hence, in what follows, we aim to generalize the L_2 performance of the learned models over the space $\mathcal{P}(\mathbb{X})$.

In what follows, we will sometimes refer to $P_x(\cdot)$ as the *training measure*.

3.1 Ideal Case: Perfectly Linear Model

If the cost and the kernel are fully linear for a given set of basis functions $\mathbf{\Phi}_{(x,u)}(\mu) = [\Phi^1_{(x,u)}(\mu), \dots \Phi^d_{(x,u)}(\mu)]^\intercal$ then the linear model can be learned perfectly. That is, for the given basis functions $\mathbf{\Phi}_{(x,u)}(\mu)$, there exist $\theta^*_{(x,u)}$ and $\mathbf{Q}^*_{(x,u)}(\cdot)$ such that

$$c(x,u,\mu) = \mathbf{\Phi}^\intercal_{(x,u)}(\mu)\theta^*_{(x,u)}$$
$$\mathcal{T}(\cdot|x,u,\mu) = \mathbf{\Phi}^\intercal_{(x,u)}(\mu)\mathbf{Q}^*_{(x,u)}(\cdot)$$

The model can be learned perfectly with a coordinator under the method presented in Section 2.1 if

- the training set T is such that for each pair (x,u), there exist at least d different data points. Furthermore, for a given data point of the form $(x^i, u^i, X^i_1, \mu, c^i)_{i=1}^\infty$, the state-action distribution for this point is such that $Pr(x,u) > 0$
- and if the basis functions $\mathbf{\Phi}_{(x,u)}(\mu)$ and $\mathbf{\Phi}_{(x,u)}(\mu')$ are linearly independent for every $\mu \neq \mu'$ that is if $\mathbf{A}_{x,u}$ (see (2.1)) has independent columns.

For the independent learning methods given in Section 2.2 and Section 2.3, the learned model will be the true model with no error if the iterations converge.

3.2 Nearly Linear Models

In this section, we provide a result that states that if the true model can be approximated ϵ close to a linear model, then the models learned with the least square method can approximate the true model uniformly in the order of ϵ if the training set is informative enough.

The following assumption states that the true model is nearly linear.

Assumption 3.1 *We assume the existence of $\bar{\theta}_{x,u} \in \mathbb{R}^d$ and $\bar{\mathbf{Q}}_{x,u}(\cdot) \in \mathbb{R}^{d \times |\mathbb{X}|}$ with the following property: denoting by $\bar{\mathbf{Q}}^j(x,u) \in \mathbb{R}^d$ the jth column of $\bar{\mathbf{Q}}_{x,u}(\cdot)$, for some $\epsilon > 0$, and $\epsilon_j > 0$*

$$\sup_{x,u,\mu} \left|\mathcal{T}(j|x,u,\mu) - \mathbf{\Phi}^\intercal_{x,u}(\mu)\bar{\mathbf{Q}}^j(x,u)\right| \leq \epsilon_j$$
$$\sup_{x,u,\mu} \left|c(x,u,\mu) - \mathbf{\Phi}^\intercal_{x,u}(\mu)\bar{\theta}_{x,u}(\cdot)\right| \leq \epsilon. \tag{28}$$

In particular, further assuming $\sum_j \epsilon_j \leq \epsilon$, the above implies that

$$\left\|\mathcal{T}(\cdot|x,u,\mu) - \mathbf{\Phi}^\intercal_{x,u}(\mu)\bar{\mathbf{Q}}_{(x,u)}(\cdot)\right\| \leq \frac{\epsilon}{2}$$

for all x, u, μ.

We note that, in general, there is no guarantee that the learned dynamics constitute a proper stochastic kernel. This can be guaranteed when the model is fully linear as discussed in Section 3.1 or when we consider a discretization based approximation as described in Section 3.3. However, for general linear approximations, as in this section, we project the learned model $\mathbf{\Phi}^\intercal_{(x,u)}(\mu)\mathbf{Q}^*_{(x,u)}$ onto the set of probability measures $P(\mathbb{X})$, i.e. the simplex over $\mathbb{X}$.

In particular, we use the following notation:

$$\hat{c}(x,u,\mu) := \mathbf{\Phi}^\intercal_{x,u}(\mu)\theta^*_{(x,u)}$$

$$\hat{\mathcal{T}}(\cdot|x,u,\mu) := \underset{\mu\in\mathcal{P}(\mathbb{X})}{\arg\min} \|\mu - \mathbf{\Phi}^{\intercal}_{x,u}(\mu)\mathbf{Q}^{*}_{(x,u)}(\cdot)\| \tag{29}$$

where $\theta^{*}_{(x,u)}$ and $\mathbf{Q}^{*}_{(x,u)}(\cdot)$ denote the learned models based on the least square method, see (27).

Proposition 3.1 *Let $P_x(\cdot) \subset \mathcal{P}(\mathcal{P}(\mathbb{X}))$ denote the training distribution of the mean-field terms for the state $x \in \mathbb{X}$. Let Assumption 3.1 hold. For the estimate models, $\hat{c}(x,u,\mu)$ and $\hat{\mathcal{T}}(\cdot|x,u,\mu)$ defined based on the least square method (see (29)), we have that*

$$|c(x,u,\mu) - \hat{c}(x,u,\mu)| \leq \epsilon\left(1 + \frac{2\sqrt{d}}{\sqrt{\lambda_{\min}}}\right)$$
$$\left\|\mathcal{T}(\cdot|x,u,\mu) - \hat{\mathcal{T}}(\cdot|x,u,\mu)\right\| \leq \epsilon\left(1 + \frac{2\sqrt{d}}{\sqrt{\lambda_{\min}}}\right)$$

where λ_{min} is the minimum eigenvalue of $\int \phi_{(x,u)}(\mu)\phi^{\intercal}_{(x,u)}(\mu)P_x(d\mu)$.

Proof We first note that since the learned $\theta^{*}_{(x,u)}$ minimizes the L_2 distance to the true model under the training measure P_x, (28) implies that

$$\int_{\mathcal{P}(\mathbb{X})} \left|c(x,u,\mu) - \mathbf{\Phi}^{\intercal}_{(x,u)}(\mu)\theta^{*}_{(\mathbf{x},\mathbf{u})}\right|^2 P_x(d\mu) \leq \int_{\mathcal{P}(\mathbb{X})} \left|c(x,u,\mu) - \mathbf{\Phi}^{\intercal}_{(x,u)}(\mu)\bar{\theta}_{(\mathbf{x},\mathbf{u})}\right|^2 P_x(d\mu) \leq \epsilon^2.$$

In particular, via the triangle inequality (under the L_2 norm) we also have that

$$\int_{\mathcal{P}(\mathbb{X})} \left|\mathbf{\Phi}^{\intercal}_{(x,u)}(\mu)\bar{\theta}_{(x,u)} - \mathbf{\Phi}^{\intercal}_{(x,u)}(\mu)\theta^{*}_{(\mathbf{x},\mathbf{u})}\right|^2 P_x(d\mu) \leq 4\epsilon^2.$$

We can further write that

$$\begin{aligned}
&\int_{\mathcal{P}(\mathbb{X})} \left|\mathbf{\Phi}^{\intercal}_{(x,u)}(\mu)\bar{\theta}_{(x,u)} - \mathbf{\Phi}^{\intercal}_{(x,u)}(\mu)\theta^{*}_{(\mathbf{x},\mathbf{u})}\right|^2 P_x(d\mu)\\
&= \int_{\mathcal{P}(\mathbb{X})} \left|\mathbf{\Phi}^{\intercal}_{(x,u)}(\mu)\left(\bar{\theta}_{(x,u)} - \theta^{*}_{(x,u)}\right)\right|^2 P_x(d\mu)\\
&= \left(\bar{\theta}_{(x,u)} - \theta^{*}_{(x,u)}\right)^{\intercal} \int_{\mathcal{P}(\mathbb{X})} \mathbf{\Phi}_{(x,u)}(\mu)\mathbf{\Phi}^{\intercal}_{(x,u)}(\mu)P_x(d\mu)\left(\bar{\theta}_{(x,u)} - \theta^{*}_{(x,u)}\right)\\
&\geq \left\|\bar{\theta}_{(x,u)} - \theta^{*}_{(x,u)}\right\|_2^2 \lambda_{\min}
\end{aligned}$$

where λ_{min} is the minimum eigenvalue of $\int \phi_{(x,u)}(\mu)\phi^{\intercal}_{(x,u)}(\mu)P_x(d\mu)$. Thus, we have that

$$\left\|\bar{\theta}_{(x,u)} - \theta^{*}_{(x,u)}\right\|_2 \leq \frac{2\epsilon}{\sqrt{\lambda_{\min}}}.$$

Finally, using the triangle inequality with the fact that $\hat{c}(x,u,\mu) = \mathbf{\Phi}^{\intercal}_{x,u}(\mu)\theta^{*}_{(x,u)}$

$$|c(x,u,\mu) - \hat{c}(x,u,\mu)| \leq \left|c(x,u,\mu) - \mathbf{\Phi}^{\intercal}_{(x,u)}(\mu)\bar{\theta}_{(x,u)}\right| + \left|\mathbf{\Phi}^{\intercal}_{(x,u)}(\mu)\bar{\theta}_{(x,u)} - \mathbf{\Phi}^{\intercal}_{x,u}(\mu)\theta^{*}_{(x,u)}\right|$$

$$\leq \epsilon + \|\mathbf{\Phi}_{(x,u)}(\mu)\|_2 \left\|\bar{\theta}_{(x,u)} - \theta^*_{(x,u)}\right\|_2 \leq \epsilon + \frac{2\epsilon\sqrt{d}}{\sqrt{\lambda_{\min}}}$$

where we used $\|\mathbf{\Phi}_{(x,u)}(\mu)\|_2 \leq \sqrt{d}$ since we assume that $\|\Phi^j_{(x,u)}(\cdot)\|_\infty \leq 1$ via Assumption 2.1.

For the proof of the error bound of the estimate kernel $\hat{\mathcal{T}}(\cdot|x,u,\mu)$ we follow identical steps, recalling that by construction $\mathbf{\Phi}^\intercal_{(x,u)}(\mu)\mathbf{Q}^{*,j}(x,u)$ minimizes the L_2 distance to $\mathcal{T}(j|x,u,\mu)$ and using Assumption 3.1, we write that

$$\lambda_{\min}\|\bar{\mathbf{Q}}^j(x,u) - \mathbf{Q}^{*,j}(x,u)\|_2^2 \leq \int \left|\mathbf{\Phi}^\intercal_{(\mathbf{x},\mathbf{u})}(\mu)\bar{\mathbf{Q}}^j(x,u) - \mathbf{\Phi}^\intercal_{(x,u)}(\mu)\mathbf{Q}^{*,j}(x,u)\right|^2 P_x(d\mu) \leq 4\epsilon_j^2,$$

which yields

$$\|\bar{\mathbf{Q}}^j(x,u) - \mathbf{Q}^{*,j}(x,u)\|_2 \leq \frac{2\epsilon_j}{\sqrt{\lambda_{\min}}}$$

where λ_{min} is the minimum eigenvalue of $\int \phi_{(x,u)}(\mu)\phi^\intercal_{(x,u)}(\mu)P_x(d\mu)$.

Since $\hat{\mathcal{T}}(\cdot|x,u,\mu)$ is the projection of $\mathbf{\Phi}^\intercal_{x,u}(\mu)\mathbf{Q}^*_{(x,u)}(\cdot)$ onto the space of probability measures, we have, by the definition of the projection, that

$$\|\hat{\mathcal{T}}(\cdot|x,u,\mu) - \mathbf{\Phi}^\intercal_{x,u}(\mu)\mathbf{Q}^*_{(x,u)}(\cdot)\| \leq \|\mathcal{T}(\cdot|x,u,\mu) - \mathbf{\Phi}^\intercal_{x,u}(\mu)\mathbf{Q}^*_{(x,u)}(\cdot)\|.$$

We then write the following using the triangle inequality

$$\begin{aligned}
&\left\|\mathcal{T}(\cdot|x,u,\mu) - \hat{\mathcal{T}}(\cdot|x,u,\mu)\right\| \\
&\leq \left\|\mathcal{T}(\cdot|x,u,\mu) - \mathbf{\Phi}^\intercal_{x,u}(\mu)\mathbf{Q}^*_{(x,u)}(\cdot)\right\| + \left\|\mathbf{\Phi}^\intercal_{x,u}(\mu)\mathbf{Q}^*_{(x,u)}(\cdot) - \hat{\mathcal{T}}(\cdot|x,u,\mu)\right\| \\
&\leq 2\|\mathcal{T}(\cdot|x,u,\mu) - \mathbf{\Phi}^\intercal_{x,u}(\mu)\mathbf{Q}^*_{(x,u)}(\cdot)\| \\
&\leq 2\left\|\mathcal{T}(\cdot|x,u,\mu) - \mathbf{\Phi}^\intercal_{x,u}(\mu)\bar{\mathbf{Q}}_{(x,u)}(\cdot)\right\| + 2\left\|\mathbf{\Phi}^\intercal_{x,u}(\mu)\bar{\mathbf{Q}}_{(x,u)}(\cdot) - \mathbf{\Phi}^\intercal_{x,u}(\mu)\mathbf{Q}^*_{(x,u)}(\cdot)\right\| \\
&\leq \epsilon + \sum_j \left|\mathbf{\Phi}^\intercal_{x,u}(\mu)\bar{\mathbf{Q}}^j(x,u) - \mathbf{\Phi}^\intercal_{x,u}(\mu)\mathbf{Q}^{*,j}(x,u)\right| \\
&\leq \epsilon + \sum_j \|\mathbf{\Phi}^\intercal_{x,u}(\mu)\|_2\|\bar{\mathbf{Q}}^j(x,u) - \mathbf{Q}^{*,j}(x,u)\|_2 \\
&\leq \epsilon + \frac{\sqrt{d}}{\sqrt{\lambda_{\min}}}\sum_j \epsilon_j \leq \epsilon + 2\frac{\sqrt{d}\epsilon}{\sqrt{\lambda_{\min}}}
\end{aligned}$$

where we used $\sum_j \epsilon_j \leq \epsilon$ with Assumption 3.1. ■

3.3 A Special Case: Linear Approximation via Discretization

In this section, we show that the discretization of the space $\mathcal{P}(\mathbb{X})$ can be seen as a particular case of linear function approximation with a special class of basis functions. In particular, for this case, we can analyze the error bounds of the learned policy with mild conditions on the model.

Let $\{B_i\}_{i=1}^d \subset \mathcal{P}(\mathbb{X})$ be a disjoint set of quantization bins of $\mathcal{P}(\mathbb{X})$ such that $\cup B_i = \mathcal{P}(\mathbb{X})$. We define the basis functions for the linear approximation such that

$$\Phi^i_{x,u}(\cdot) = \mathbb{1}_{B_i}(\cdot)$$

for all (x, u) pairs. Note that in general the quantization bins, B_i's, can be chosen differently for every (x, u); for the simplicity of the analysis, we will work with a discretization scheme which is the same for every (x, u). An important property of the discretization is that the basis functions form an orthonormal basis for any training measure $P(\cdot)$ with $P(B_i) > 0$ for each quantization bin B_i. That is

$$\int \langle \Phi^i_{x,u}(\mu), \Phi^j_{x,u}(\mu) \rangle P(d\mu) = \mathbb{1}_{\{i=j\}}$$

for every (x, u) pair. This property allows us to analyze the uniform error bounds of the discretization method more directly.

The linear fitted model (see (27) with the chosen basis functions becomes

$$\begin{aligned} \theta^i_{(x,u)} &= \frac{\int_{B_i} c(x, u, \mu) P(d\mu)}{P(B_i)} \\ Q^i_{(x,u)}(\cdot) &= \frac{\int_{B_i} \mathcal{T}(\cdot|x, u, \mu) P(d\mu)}{P(B_i)} \end{aligned} \tag{30}$$

In words, the learned coefficients are the averages of cost and transition realizations from the training set of the corresponding quantization bin.

The following then is an immediate result of Assumption 1.1.

Proposition 3.2 *Let $\theta_{x,u} = [\theta^1_{x,u}, \dots, \theta^d_{x,u}]^\intercal$ and $\mathbf{Q}_{x,u}(\cdot) = \left[Q^1_{x,u}(\cdot), \dots, Q^d_{x,u}(\cdot)\right]^\intercal$ be given by (30). If the training measure $P(\cdot)$ is such that $P(B_i) > 0$ for each quantization bin B_i, under Assumption 1.1, we then have that*

$$\begin{aligned} \left|c(x, u, \mu) - \mathbf{\Phi}^\intercal_{x,u}(\mu)\theta_{x,u}\right| &\leq K_c L \\ \left\|\mathcal{T}(\cdot|x, u, \mu) - \mathbf{\Phi}^\intercal_{x,u}(\mu)\mathbf{Q}_{x,u}\right\| &\leq K_f L \end{aligned}$$

where L is the largest diameter of the quantizations bins such that

$$L = \max_i \sup_{\mu,\mu' \in B_i} \|\mu - \mu'\|.$$

4. Error Analysis for Control with Misspecified Models

In the previous section, we have studied the uniform mismatch bounds of the learned models. In this section, we will focus on what happens if the controllers designed for the linear estimates are used for the true dynamics. We will provide error bounds for the performance loss of the control designed for a possibly missepecified model.

We will analyze the infinite population and the finite population settings separately. We note that some of the following results (e.g. Lemma 16) have been studied in the literature to establish the connection between the N-agent control problems and the limit mean-field control problem without

the model mismatch aspect. That is, existing results study what happens if one uses the infinite population solution for the finite population control problem with perfectly known dynamics (see e.g. Motte and Pham (2023); Bäuerle (2023); Motte and Pham (2022)). However, we present the proof of every result for completeness and because of the connections in the analysis we follow throughout the paper. Furthermore, the existing results are often stated under slightly different assumptions and settings such as being stated only for closed loop policies, or only for policies that are open loop in the sense that they are measurable with respect to the noise process.

4.1 Error Bounds for Infinitely Many Agents

As we have observed in Example 2, even when agents agree on the model knowledge, without coordination on which policy to follow, the optimality may not be achieved. Therefore, we assume that after the learning period, the team of agents collectively agrees on the cost and transition models given by $\hat{c}(x,u,\mu)$ and $\hat{\mathcal{T}}(\cdot|x,u,\mu)$ and designs policies for this model. We will assume that

$$\begin{aligned} &|c(x,u,\mu)-\hat{c}(x,u,\mu)|\leq\lambda \\ &\left\|\mathcal{T}(\cdot|x,u,\mu)-\hat{\mathcal{T}}(\cdot|x,u,\mu)\right\|\leq\lambda \end{aligned} \tag{31}$$

for some $\lambda<\infty$ and for all x,u,μ. That is λ represents the uniform model mismatch constant.

We will consider two different cases for the execution of the designed control.

- *Closed loop control:* The team decides on a policy $\hat{g}:\mathcal{P}(\mathbb{X})\to\Gamma$, and uses their local states and the mean-field term to apply the policy $\hat{g}$. That is, an agent i observes the mean-field term μ_t, chooses $\hat{g}(\mu_t)=\hat{\gamma}(\cdot|x^i,\mu_t)$ and applies their control action according to $\hat{\gamma}(\cdot|x^i,\mu_t)$ with the local state x^i. The important distinction is that the mean-field term μ_t is observed by every agent, and they decide on their agent-level policies with the observed mean-field term. Hence, we refer to this execution method to be the closed loop method since the mean-field term is given as a feedback variable.

- *Open loop control:* We have argued earlier that the flow of the mean-field term μ_t is deterministic for the infinitely many agent case, see (11). In particular, the mean-field term μ_t can be estimated with the model information. Hence, for this case, we will assume that the agents only observe their local states, and estimates the mean-field term independent instead of observing it. That is, an agent i estimates the mean-field term $\hat{\mu}_t$, and applies their control action according to $\hat{\gamma}(\cdot|x^i,\hat{\mu}_t)$ with the local state x^i. Note that if the model dynamics were perfectly known, this estimate would coincide with the true flow of the mean-field term. However, when the model is misspecified, the estimate $\hat{\mu}_t$ and the correct mean-field term will deviate from each other, and we will need to study the effects of this deviation on the control performance, in addition to the incorrect computation of the control policy.

In what follows, previously introduced constants K_c, K_f and δ_T will be used often. We refer the reader to Assumption 1.1 for K_c, K_f, and equation (6) for δ_T.

For the results in this section, we will require that $\beta K<1$ where $K=K_f+\delta_T$. We note that this assumption is needed to show the Lipschitz continuity of the value function $K^*_\beta(\mu)$ with respect to μ. The following provides an example where this bound is not satisfied, and the value function is not Lipschitz continuous.

Example 3 *Consider a control-free (without loss of optimality) dynamics, with a binary state space* $\mathbb{X} = \{0, 1\}$. *We assume that*

$$\mathcal{T}(0|x, \mu) = \mu(0)^2$$

that is, the state process moves to 0 *with probability* $\mu(0)^2$ *independent of the value of the state at the current step. We first notice that* $\|\mu - \mu'\| = |\mu(0) - \mu'(0)|$ *for the binary state space. Furthermore, we note that this kernel is Lipschitz continuous in* μ *with Lipschitz constant 2, that is* $K_f = 2$. *To see this, consider the following for* $\mu, \mu' \in \mathcal{P}(\mathbb{X})$

$$\begin{aligned}\|\mathcal{T}(\cdot|x, \mu) - \mathcal{T}(\cdot|x, \mu')\| &= \frac{1}{2}\left(|\mathcal{T}(0|x, \mu) - \mathcal{T}(0|x, \mu')| + |\mathcal{T}(1|x, \mu) - \mathcal{T}(1|x, \mu')|\right) \\ &= |\mu(0)^2 - \mu'(0)^2| = |\mu(0) - \mu'(0)| \times |\mu(0) + \mu'(0)| \leq 2|\mu(0) - \mu'(0)| = 2\|\mu(\cdot) - \mu'(\cdot)\|\end{aligned}$$

where we used the bound that $|\mu(0) + \mu'(0)| \leq 2$ *which is the minimal uniform upper bound for all* μ, μ'.

Hence, the kernel is Lipschitz continuous with constant 2. Furthermore, since the dynamics do not depend x *and* u, *we have that* $\delta_T = 0$, *and thus* $K = K_f + \delta_T = 2$.

The stage-wise cost is given by $c(\mu) = \mu(0)$. *We consider Lipschitz continuity of the value function around* $\mu(0) = 1$ *,i.e. around* $\mu = \delta_0$. *Note that for some initial distribution* $\mu_0(0) = a$, *one can iteratively show that*

$$\mu_t(0) = a^{2^t}.$$

Hence, we can write the value function as

$$K_\beta(a) = \sum_{t=0}^{\infty} \beta^t a^{2^t}.$$

To show that this function is not Lipschitz continuous, we consider two points $a, b \in [0, 1]$, *without loss of generality assume that* $a \geq b$:

$$\frac{K_\beta(a) - K_\beta(b)}{a - b} = \frac{\sum_{t=0}^{\infty} \beta^t (a^{2^t} - b^{2^t})}{a - b} = \sum_{t=0}^{\infty} \beta^t 2^t c^{2^t - 1}$$

for some $c \in [a, b]$ *where we used the mean value theorem for* $\frac{a^{2^t} - b^{2^t}}{a-b}$. *We can see that the above cannot be bounded uniformly when* c *is around* 1 *if* $\beta \geq 1/2$, *i.e. if* $\beta K \geq 1$. *This implies that the value function cannot be Lipschitz continuous if* $\beta K \geq 1$.

4.1.1 ERROR BOUNDS FOR CLOSED LOOP CONTROL

We assume that the agents calculate an optimal policy, say $\hat{g}$, for the incorrect model ($\hat{\mathcal{T}}$ and $\hat{c}$), and observe the *correct* mean-field term say μ_t, at every time step t. The agents then use

$$\hat{g}(\mu_t) = \hat{\gamma}(\cdot|x_t, \mu_t) \tag{32}$$

to select their control actions u_t at time t.

We denote the accumulated cost under this policy $\hat{g}$ by $K_\beta(\mu_0, \hat{g})$, and we will compare this with the optimal cost that can be achieved, which is $K^*_\beta(\mu_0)$ for some initial distribution μ_0.

Theorem 10 *Consider the closed loop policy $\hat{g}$ in (32) designed for an estimate model $\hat{\mathcal{T}}, \hat{c}$ which satisfies (31) for the infinite population dynamics. Under Assumption 1.1, if $\beta K < 1$*

$$K_\beta(\mu_0, \hat{g}) - K^*_\beta(\mu_0) \leq 2\lambda \frac{(\beta C - \beta K + 1)}{(1-\beta)^2(1-\beta K)}$$

where $K = (K_f + \delta_T)$ and $C = (\|c\|_\infty + K_c)$.

Proof We start with the following upper-bound

$$K_\beta(\mu, \hat{g}) - K^*_\beta(\mu) \leq \left| K_\beta(\mu, \hat{g}) - \hat{K}_\beta(\mu) \right| + \left| \hat{K}_\beta(\mu) - K^*_\beta(\mu) \right| \tag{33}$$

where $\hat{K}_\beta(\mu)$ denotes the optimal value function for the mismatched model. We have an upper-bound for the second term by Lemma 12. We write the following Bellman equations for the first term:

$$\begin{aligned} K_\beta(\mu, \hat{g}) &= k(\mu, \hat{\gamma}) + \beta K_\beta \left(F(\mu, \hat{\gamma}), \hat{g} \right) \\ \hat{K}_\beta(\mu) &= \hat{k}(\mu, \hat{\gamma}) + \beta \hat{K}_\beta \left(\hat{F}(\mu, \hat{\gamma}) \right). \end{aligned}$$

We can then write

$$\begin{aligned} \left| K_\beta(\mu, \hat{g}) - \hat{K}_\beta(\mu) \right| \leq & \left| k(\mu, \hat{\gamma}) - \hat{k}(\mu, \hat{\gamma}) \right| \\ & + \beta \left| K_\beta \left(F(\mu, \hat{\gamma}), \hat{g} \right) - \hat{K}_\beta \left(F(\mu, \hat{\gamma}) \right) \right| \\ & + \beta \left| \hat{K}_\beta \left(F(\mu, \hat{\gamma}) \right) - K^*_\beta \left(F(\mu, \hat{\gamma}) \right) \right| \\ & + \beta \left| K^*_\beta \left(F(\mu, \hat{\gamma}) \right) - K^*_\beta (\hat{F}(\mu, \hat{\gamma})) \right| \\ & + \beta \left| K^*_\beta (\hat{F}(\mu, \hat{\gamma})) - \hat{K}_\beta \left(\hat{F}(\mu, \hat{\gamma}) \right) \right| \end{aligned}$$

We note that $\left| k(\mu, \hat{\gamma}) - \hat{k}(\mu, \hat{\gamma}) \right| \leq \lambda$ and $\|F(\mu, \hat{\gamma}) - \hat{F}(\mu, \hat{\gamma})\| \leq \lambda$. Using Lemma 12 for the third and the last terms above, we get

$$\begin{aligned} \left| K_\beta(\mu, \hat{g}) - \hat{K}_\beta(\mu) \right| \leq & \lambda + \beta \sup_\mu \left| K_\beta(\mu, \hat{g}) - \hat{K}_\beta(\mu) \right| \\ & + 2\lambda\beta \left(\frac{\beta C - \beta K + 1}{(1-\beta)(1-\beta K)} \right) + \beta \|K^*_\beta\|_{Lip} \lambda. \end{aligned}$$

Rearranging the terms and taking the supremum on the left hand side over $\mu \in \mathcal{P}(\mathbb{X})$, and noting that $\|K^*_\beta\|_{Lip} \leq \frac{C}{1-\beta K}$ we can then write

$$\begin{aligned} \left| K_\beta(\mu, \hat{g}) - \hat{K}_\beta(\mu) \right| &\leq \frac{\lambda}{(1-\beta)} \left(1 + 2\beta \left(\frac{\beta C - \beta K + 1}{(1-\beta)(1-\beta K)} \right) + \frac{\beta C}{(1-\beta K)} \right) \\ &= \lambda \left(\frac{(1+\beta)(\beta C - \beta K + 1)}{(1-\beta)^2(1-\beta K)} \right) \end{aligned}$$

Combining this bound, and Lemma 12 with (33), we can conclude the proof.

■

4.1.2 ERROR BOUNDS FOR OPEN LOOP CONTROL

We assume that the agents calculate an optimal policy, say $\hat{g}$ for the incorrect model, and estimate the mean-field flow under the incorrect model with the policy $\hat{g}$. That is, at every time step t, the agents use

$$\hat{g}(\hat{\mu}_t) = \hat{\gamma}(\cdot|x_t, \hat{\mu}_t) \tag{34}$$

to select their control actions u_t at time t. Furthermore, $\hat{\mu}_t$ is estimated with

$$\hat{\mu}_{t+1}(\cdot) = \hat{\mathcal{T}}(\cdot|x, u, \hat{\mu}_t)\hat{\gamma}(du|x, \hat{\mu}_t)\hat{\mu}_t(dx) \tag{35}$$

where $\hat{\mathcal{T}}$ is the learned and possibly incorrect model. We are then interested in the optimality gap given by

$$K_\beta(\mu_0, \hat{g}) - K^*_\beta(\mu_0)$$

where $K_\beta(\mu_0, \hat{g})$ denotes the accumulated cost when the agents follow the open loop policy $\hat{g}(\hat{\mu}_t) = \hat{\gamma}(\cdot|x_t, \hat{\mu}_t)$ at every time t. We note that the distinction from the closed loop control is that $\hat{\mu}_t$ is not observed but estimated using the model $\hat{\mathcal{T}}$.

Theorem 11 *Consider the open loop policy $\hat{g}$ in (32) which is designed for an estimate model that satisfies (31) for the infinite population dynamics. Under Assumption 1.1, if $\beta K < 1$,*

$$K_\beta(\mu_0, \hat{g}) - K^*_\beta(\mu_0) \leq 2\lambda \frac{\beta(C-K)+1}{(1-\beta)(1-\beta K)}$$

for any $\mu_0 \in \mathcal{P}(\mathbb{X})$ where $C = \|c\|_\infty + K_c$ and $K = K_f + \delta_T$.

Proof We start with the following upper-bound

$$K_\beta(\mu_0, \hat{g}) - K^*_\beta(\mu_0) \leq \left|K_\beta(\mu_0, \hat{g}) - \hat{K}_\beta(\mu_0)\right| + \left|\hat{K}_\beta(\mu_0) - K^*_\beta(\mu_0)\right| \tag{36}$$

We have an upper-bound for the second term by Lemma 12. We now focus on the first term:

$$\left|K_\beta(\mu_0, \hat{g}) - \hat{K}_\beta(\mu_0)\right| \leq \sum_{t=0}^{\infty} \beta^t \left|k(\mu'_t, \hat{\gamma}_t) - \hat{k}(\hat{\mu}_t, \hat{\gamma}_t)\right|$$

where we write $\hat{\gamma}_t := \hat{\gamma}(\cdot|x, \hat{\mu}_t)$, and μ'_t denotes the measure flow under the true dynamics with the incorrect policy $\hat{\gamma}_t$, that is

$$\mu'_{t+1} = \hat{F}(\mu'_t, \hat{\gamma}_t) := \int \mathcal{T}(\cdot|x, u, \mu'_t)\hat{\gamma}(du|x, \hat{\mu}_t)\mu'_t(dx).$$

We next claim that

$$\|\mu'_t - \hat{\mu}_t\| \leq \lambda \sum_{n=0}^{t-1} (\delta_T + K_f)^n.$$

We show this by induction. For $t = 1$, we have that

$$\begin{aligned}\|\mu_1' - \hat{\mu}_1\| &= \left\| \int \mathcal{T}(\cdot|x,u,\mu_0)\hat{\gamma}(du|x,\mu_0)\mu_0(dx) - \int \hat{\mathcal{T}}(\cdot|x,u,\mu_0)\hat{\gamma}(du|x,\mu_0)\mu_0(dx) \right\| \\ &\leq \lambda.\end{aligned}$$

We now assume that the claim is true for t:

$$\begin{aligned}\|\mu_{t+1}' - \hat{\mu}_{t+1}\| &= \left\| \int \mathcal{T}(\cdot|x,u,\mu_t')\hat{\gamma}(du|x,\hat{\mu}_t)\mu_t'(dx) - \int \hat{\mathcal{T}}(\cdot|x,u,\hat{\mu}_t)\hat{\gamma}(du|x,\hat{\mu}_t)\hat{\mu}_t(dx) \right\| \\ &\leq \left\| \int \mathcal{T}(\cdot|x,u,\mu_t')\hat{\gamma}(du|x,\hat{\mu}_t)\mu_t'(dx) - \int \mathcal{T}(\cdot|x,u,\mu_t')\hat{\gamma}(du|x,\hat{\mu}_t)\hat{\mu}_t(dx) \right\| \\ &\quad + \left\| \int \mathcal{T}(\cdot|x,u,\mu_t')\hat{\gamma}(du|x,\hat{\mu}_t)\hat{\mu}_t(dx) - \int \hat{\mathcal{T}}(\cdot|x,u,\hat{\mu}_t)\hat{\gamma}(du|x,\hat{\mu}_t)\hat{\mu}_t(dx) \right\| \\ &\leq \delta_T\|\mu_t' - \hat{\mu}_t\| + \sup_{x,u} \left\| \mathcal{T}(\cdot|x,u,\mu_t') - \hat{\mathcal{T}}(\cdot|x,u,\hat{\mu}_t) \right\| \\ &\leq \delta_T\|\mu_t' - \hat{\mu}_t\| + \sup_{x,u} \left\| \mathcal{T}(\cdot|x,u,\mu_t') - \mathcal{T}(\cdot|x,u,\hat{\mu}_t) \right\| \\ &\qquad + \sup_{x,u} \left\| \mathcal{T}(\cdot|x,u,\hat{\mu}_t) - \hat{\mathcal{T}}(\cdot|x,u,\hat{\mu}_t) \right\| \\ &\leq (\delta_T + K_f)\|\mu_t' - \hat{\mu}_t\| + \lambda \\ &\leq (\delta_T + K_f)\lambda \sum_{n=0}^{t-1} (\delta_T + K_f)^n + \lambda = \lambda \sum_{n=0}^{t} (\delta_T + K_f)^n.\end{aligned}$$

where we used the induction argument at the last inequality. We now go back to:

$$\left| K_\beta(\mu_0, \hat{g}) - \hat{K}_\beta(\mu_0) \right| \leq \sum_{t=0}^{\infty} \beta^t \left| k(\mu_t', \hat{\gamma}_t) - \hat{k}(\hat{\mu}_t, \hat{\gamma}_t) \right|.$$

For the term inside the summation, we write

$$\begin{aligned}\left| k(\mu_t', \hat{\gamma}_t) - \hat{k}(\hat{\mu}_t, \hat{\gamma}_t) \right| &= \left| \int c(x,u,\mu_t')\hat{\gamma}(du|x,\hat{\mu}_t)\mu_t'(dx) - \int \hat{c}(x,u,\hat{\mu}_t)\hat{\gamma}(du|x,\hat{\mu}_t)\hat{\mu}_t(dx) \right| \\ &\leq \left| \int c(x,u,\mu_t')\hat{\gamma}(du|x,\hat{\mu}_t)\mu_t'(dx) - \int c(x,u,\mu_t')\hat{\gamma}(du|x,\hat{\mu}_t)\hat{\mu}_t(dx) \right| \\ &\quad + \left| \int c(x,u,\mu_t')\hat{\gamma}(du|x,\hat{\mu}_t)\hat{\mu}_t(dx) - \int \hat{c}(x,u,\hat{\mu}_t)\hat{\gamma}(du|x,\hat{\mu}_t)\hat{\mu}_t(dx) \right| \\ &\leq \|c\|_\infty \|\mu_t' - \hat{\mu}_t\| + \sup_{x,u} \left| c(x,u,\mu_t') - \hat{c}(x,u,\hat{\mu}_t) \right| \\ &\leq \|c\|_\infty \|\mu_t' - \hat{\mu}_t\| + \sup_{x,u} \left| c(x,u,\mu_t') - c(x,u,\hat{\mu}_t) \right| \\ &\qquad + \sup_{x,u} \left| c(x,u,\hat{\mu}_t) - \hat{c}(x,u,\hat{\mu}_t) \right| \\ &\leq (\|c\|_\infty + K_c)\|\mu_t' - \hat{\mu}_t\| + \lambda.\end{aligned}$$

Using this bound, we finalize our argument. In the following we denote by $K := (K_f + \delta_T)$ and $C := (\|c\|_\infty + K_c)$ to conclude:

$$\begin{aligned}
&\left|K_\beta(\mu_0, \hat{g}) - \hat{K}_\beta(\mu_0)\right| \leq \sum_{t=0}^{\infty} \beta^t \left|k(\mu'_t, \hat{\gamma}_t) - \hat{k}(\hat{\mu}_t, \hat{\gamma}_t)\right| \\
&\leq C \sum_{t=0}^{\infty} \beta^t \|\mu'_t - \hat{\mu}_t\| + \frac{\lambda}{1-\beta} \\
&\leq C\lambda \sum_{t=0}^{\infty} \beta^t \sum_{n=0}^{t-1} K^n + \frac{\lambda}{1-\beta} = C\lambda \sum_{t=0}^{\infty} \beta^t \frac{1-K^t}{1-K} + \frac{\lambda}{1-\beta} \\
&= \frac{C\lambda}{(1-\beta)(1-K)} - \frac{C\lambda}{(1-K)(1-\beta K)} + \frac{\lambda}{1-\beta} \\
&= \frac{C\lambda\beta}{(1-\beta)(1-\beta K)} + \frac{\lambda}{1-\beta} = \lambda \frac{\beta(C-K)+1}{(1-\beta)(1-\beta K)}.
\end{aligned}$$

This is the bound for the first term in (36), combining this with the upper-bound on the second term in (36) by Lemma 12, we can complete the proof. ■

Lemma 12 *Under Assumption 1.1, if* $\beta K < 1$

$$\left|\hat{K}_\beta(\mu_0) - K^*_\beta(\mu_0)\right| \leq \lambda \left(\frac{\beta C - \beta K + 1}{(1-\beta)(1-\beta K)}\right)$$

for any initial distribution $\mu_0 \in \mathcal{P}(\mathbb{X})$ *where* $C = \|c\|_\infty + K_c$ *and* $K = K_f + \delta_T$.

Proof The proof can be found in the appendix B. ■

4.2 Error Bounds for Finitely Many Agents

We introduce the following constant to denote the expected distance of an empirical measure to its true distribution:

$$M_N := \sup_{\mu \in \mathcal{P}(\mathbb{X})} E\left[\left\|\mu^N - \mu\right\|\right] \tag{37}$$

$$\bar{M}_N = \sup_{\mu \in \mathcal{P}(\mathbb{X}\times\mathbb{U})} E\left[\left\|\mu^N - \mu\right\|\right] \tag{38}$$

where μ^N is an empirical measure of the distribution μ, and the expectation is with respect to the randomness over the realizations of μ^N.

Remark 13 *We note that the constants can be bounded in terms of the population size* N. *In particular, for the finite space* $\mathbb{X}$ *and* $\mathbb{U}$

$$M_N \leq \frac{K}{\sqrt{N}}$$

where $K < \infty$ *in general depends on the underlying space* $\mathbb{X}$ *(or the space* $\mathbb{X} \times \mathbb{U}$ *for* $\bar{M}_N$*). Furthermore, for continuous state and action spaces, e.g. for* $\mathbb{X} \subset \mathbb{R}^d$*, the empirical error term is in the order of* $O(N^{\frac{-1}{2d}})$.

4.2.1 ERROR BOUNDS FOR OPEN LOOP CONTROL

In this section, we will study the case where each agent in an N-agent control system follows the open-loop control given by the solution of the infinite population control problem with mismatched model estimation. We summarize this for some agent i as follows:

- Collectively agree on a policy $\hat{g}$ as in (34) according to the agreed upon estimate model $\hat{\mathcal{T}}, \hat{c}$ that satisfies (31). Note that this policy is an optimal policy for the infinite population dynamics under the estimate model.
- Estimate the mean-field term $\hat{\mu}_t$ at time t according to (35) using the approximate model $\hat{\mathcal{T}}$
- Find the randomized agent level policy $\hat{\gamma}$ using $\hat{g}(\hat{\mu}_t) = \hat{\gamma}(\cdot|x_t^i, \hat{\mu}_t)$
- Observe local state x_t^i, and apply action $u_t^t \sim \hat{\gamma}(\cdot|x_t^i, \hat{\mu}_t)$.

If every agent follows this policy, we have the following upperbound for the performance loss compared to the optimal value of the N-population control problem,

Theorem 14 *Under Assumption 1.1, if each agent follows the steps summarized above, we then have that*

$$K_\beta^N(\mu^N, \hat{\gamma}) - K_\beta^{N,*}(\mu^N) \leq 2\lambda \left(\frac{\beta C - \beta K + 1}{(1-\beta)(1-\beta K)} \right) + M_N \frac{4\beta C}{(1-\beta)(1-\beta K)}.$$

where $C = (\|c\|_\infty + K_c)$, and $K = (K_f + \delta_T)$.

Proof For some $\hat{\mu}_0 = \mu_0 = \mu_{\mathbf{x_0}} = \mu^N$

$$\begin{aligned} K_\beta^N(\mu^N, \hat{\gamma}) - K_\beta^{N,*}(\mu^N) \leq & \left| K_\beta^N(\mu^N, \hat{\gamma}) - \hat{K}_\beta^*(\mu^N) \right| + \left| \hat{K}_\beta^*(\mu^N) - K_\beta^*(\mu^N) \right| \\ & + \left| K_\beta^*(\mu^N) - K_\beta^{N,*}(\mu^N) \right|. \end{aligned}$$

The second term above is bounded by Lemma 12, the last term is bounded by Lemma 16, finally for the first term we have

$$\left| K_\beta^N(\mu^N, \hat{\gamma}) - \hat{K}_\beta^*(\mu^N) \right| \leq \sum_{t=0}^{\infty} \beta^t E \left[\left| k(\mu_{\mathbf{x_t}}, \hat{\gamma}) - \hat{k}(\hat{\mu}_t, \hat{\gamma}) \right| \right]$$

For the term inside of the expectation, we have

$$\begin{aligned} & \left| k(\mu_{\mathbf{x_t}}, \hat{\gamma}) - \hat{k}(\hat{\mu}_t, \hat{\gamma}) \right| \\ & = \left| \int c(x, u, \mu_{\mathbf{x_t}}) \hat{\gamma}(du|x, \hat{\mu}_t) \mu_{\mathbf{x_t}}(dx) - \int \hat{c}(x, u, \hat{\mu}_t) \hat{\gamma}(du|x, \hat{\mu}_t) \hat{\mu}_t(dx) \right| \\ & \leq \lambda + C \|\mu_{\mathbf{x_t}} - \hat{\mu}_t\| \end{aligned}$$

where $C = (\|c\|_\infty + K_c)$. We can then write

$$\left| K_\beta^N(\mu^N, \hat{\gamma}) - \hat{K}_\beta^*(\mu^N) \right| \leq \sum_{t=0}^{\infty} \beta^t E \left[\left| k(\mu_{\mathbf{x_t}}, \hat{\gamma}) - \hat{k}(\hat{\mu}_t, \hat{\gamma}) \right| \right]$$

$$
\begin{aligned}
&\leq \sum_{t=0}^{\infty} \beta^t \left(\lambda + CE\left[\|\mu_{\mathbf{x_t}} - \hat{\mu}_t\|\right]\right) \\
&\leq \frac{\lambda}{1-\beta} + C\sum_{t=0}^{\infty}\beta^t \sum_{n=0}^{t-1} K^n(\lambda + 2M_N) \\
&= \frac{\lambda}{1-\beta} + \frac{\beta C(\lambda + 2M_N)}{(1-\beta)(1-\beta K)} \\
&= \lambda\left(\frac{\beta C - \beta K + 1}{(1-\beta)(1-\beta K)}\right) + M_N \frac{2\beta C}{(1-\beta)(1-\beta K)}.
\end{aligned}
$$

where we have used Lemma 15 which is presented below. ■

Lemma 15 *Let x_t^i be the state of the agent i at time t when each agent follows the open-loop policy $\hat{\gamma}(\cdot|x_t^i, \hat{\mu}_t)$ in an N-agent control dynamics. We denote by $\mathbf{x}_t$ the vector of the states of N agents at time t. Under Assumption 1.1, we then have that*

$$
E\left[\|\mu_{\mathbf{x_t}} - \hat{\mu}_t\|\right] \leq \sum_{n=0}^{t-1} K^n(\lambda + 2M_N)
$$

where $K = K_f + \delta_t$, and where the expectation is with respect to the random dynamics of the N player control system.

Proof The proof can be found in Appendix C. ■

Lemma 16 *Under Assumption 1.1,*

$$
\left|K_\beta^{N,*}(\mu^N) - K_\beta^*(\mu^N)\right| \leq \frac{2\beta C}{(1-\beta)(1-\beta K)} M_N
$$

where $C = (\|c\|_\infty + K_c)$ and $K = (K_f + \delta_T)$, for any $\mu^N \in \mathcal{P}_N(\mathbb{X}) \subset \mathcal{P}(\mathbb{X})$ that is for any μ^N that can be achieved with an empirical distribution of N agents.

Proof The proof can be found in the appendix D. ■

4.2.2 ERROR BOUNDS FOR CLOSED LOOP CONTROL

In this section, we will assume that the agents find and agree on an optimal policy for the control problem using the agreed-upon mismatched model $\hat{c}, \hat{\mathcal{T}}$ with infinite agent dynamics. However, unlike open-loop control, to execute this policy, they observe the empirical state distribution of the team of N-agents, say μ_t^N at time t and apply $\hat{\gamma}(\cdot|x, \mu_t^N)$. We summarize the application of this policy as follows:

- Collectively agree on a policy $\hat{g}$ as in (32) according to the agreed upon estimate model $\hat{\mathcal{T}}, \hat{c}$ that satisfies (31). Note that this policy is an optimal policy for the infinite population dynamics under the estimate model.

- *Observe* the *correct* mean-field term μ_t.
- Find the randomized agent level policy $\hat{\gamma}$ using $\hat{g}(\mu_t) = \hat{\gamma}(\cdot|x_t^i, \mu_t)$
- Observe local state x_t^i, and apply action $u_t^t \sim \hat{\gamma}(\cdot|x_t^i, \mu_t)$.

We denote the incurred cost under this policy by $K_\beta^N(\mu^N, \hat{\gamma})$ for some initial state distribution μ^N.

Theorem 17 *Under Assumption 1.1, if each agent follows the steps summarized above, we then have that*

$$K_\beta^N(\mu^N, \hat{\gamma}) - K_\beta^{N,*}(\mu^N) \leq \lambda \frac{2(\beta C - \beta K + 1)}{(1-\beta)^2(1-\beta K)} + M_N \frac{4\beta C}{(1-\beta)(1-\beta K)}$$

where $K = (K_f + \delta_T)$ *and* $C = (\|c\|_\infty + K_c)$.

Proof The proof follows very similar steps to the results we have proved earlier. For some $\hat{\mu}_0 = \mu_0 = \mu_{\mathbf{x_0}} = \mu^N$

$$\begin{aligned} K_\beta^N(\mu^N, \hat{\gamma}) - K_\beta^{N,*}(\mu^N) \leq & \left| K_\beta^N(\mu^N, \hat{\gamma}) - \hat{K}_\beta(\mu^N) \right| + \left| \hat{K}_\beta(\mu^N) - K_\beta^*(\mu^N) \right| \\ & + \left| K_\beta^*(\mu^N) - K_\beta^{N,*}(\mu^N) \right|. \end{aligned} \tag{39}$$

The second term above is bounded by Lemma 12, the last term is bounded by Lemma 16. For the first term we write the Bellman equations:

$$\begin{aligned} K_\beta^N(\mu^N, \hat{\gamma}) &= k(\mu^N, \hat{\gamma}) + \beta \int K_\beta^N(\mu_1^N, \hat{\gamma}) \eta(d\mu_1^N | \mu^N, \hat{\gamma}) \\ \hat{K}_\beta(\mu^N) &= \hat{k}(\mu^N, \hat{\gamma}) + \beta \hat{K}_\beta \left(\hat{F}(\mu^N, \hat{\gamma}) \right). \end{aligned}$$

We can then write

$$\begin{aligned} \left| K_\beta^N(\mu^N, \hat{\gamma}) - \hat{K}_\beta^*(\mu^N) \right| \leq & \left| k(\mu^N, \hat{\gamma}) - \hat{k}(\mu^N, \hat{\gamma}) \right| \\ & + \beta \int \left| K_\beta^N(\mu_1^N, \hat{\gamma}) - \hat{K}_\beta(\mu_1^N) \right| \eta(d\mu_1^N | \mu^N, \hat{\gamma}) \\ & + \beta \int \left| \hat{K}_\beta(\mu_1^N) - \hat{K}_\beta \left(\hat{F}(\mu^N, \hat{\gamma}) \right) \right| \eta(d\mu_1^N | \mu^N, \hat{\gamma}) \\ & \leq \lambda + \sup_\mu \left| K_\beta^N(\mu, \hat{\gamma}) - \hat{K}_\beta(\mu) \right| \\ & + 2\beta\lambda \left(\frac{\beta C - \beta K + 1}{(1-\beta)(1-\beta K)} \right) \\ & + \beta \int \left| K_\beta^*(\mu_1^N) - K_\beta^* \left(\hat{F}(\mu^N, \hat{\gamma}) \right) \right| \eta(d\mu_1^N | \mu^N, \hat{\gamma}) \end{aligned}$$

Using almost identical arguments that we have used in the proof of Lemma 16 and Lemma 15, we can bound the last term as

$$\beta \int \left| K_\beta^*(\mu_1^N) - K_\beta^* \left(\hat{F}(\mu^N, \hat{\gamma}) \right) \right| \eta(d\mu_1^N | \mu^N, \hat{\gamma})$$

$$\leq \beta \|K^*_\beta\|_{Lip} \left(M_N + \delta_T M_N + \lambda\right)$$

Re arranging the terms and noting that $\|K^*_\beta\|_{Lip} \leq \frac{C}{1-\beta K}$, we can write that

$$\sup_{\mu \in \mathcal{P}_N(\mathbb{X})} \left| K^N_\beta(\mu, \hat{\gamma}) - \hat{K}_\beta(\mu) \right| \leq M_N \frac{2\beta C}{(1-\beta)(1-\beta K)} + \lambda \frac{(1+\beta)(\beta C - \beta K + 1)}{(1-\beta)^2(1-\beta K)}.$$

Combining this bound with (39), one can show that

$$K^N_\beta(\mu^N, \hat{\gamma}) - K^{N,*}_\beta(\mu^N) \leq \lambda \frac{2(\beta C - \beta K + 1)}{(1-\beta)^2(1-\beta K)} + M_N \frac{4\beta C}{(1-\beta)(1-\beta K)}$$

■

5. Numerical Study

We now present a numerical example to verify the results we have established in the earlier sections.

We consider a multi-agent taxi service model where each agent represents a taxi. The state and action spaces are binary such that $\mathbb{X} = \mathbb{U} = \{0, 1\}$. We assume that at any given time a given zone is in either a surge or a non-surge mode. The state variable X^i_t represents the location of the agent i

- $X^i_t = 0 \rightarrow$ agent is in a surge zone (high demand)
- $X^i_t = 1 \rightarrow$ agent is in a non-surge zone (low demand)

The action variable represents the movement decisions:

- $U^i_t = 0 \rightarrow$ remains where they are
- $U^i_t = 1 \rightarrow$ relocates to another area.

The cost structure is defined as follows:

- If an agent is in a non-surge zone ($X^i_t = 1$), they incur a cost S due to lost earnings
- If an agent relocates ($U^i_t = 1$), they receive a cost R, for movement expenses.
- Furthermore, to encourage a balanced distribution, we penalize deviations from 40%-60% distribution by introducing a cost $10 \times (\mu(0) - 0.4)^2$ where $\mu(0)$ is the fraction of agents in the surge zones.

For the dynamics, we assume that a non-surge area has a fixed probability 0.2 of becoming a surge area in the next time step. Furthermore, we assume that a surge area has a probability $0.7 \times \mu(0) + 0.2$ of becoming a non-surge area, indicating that as more drivers there are in a surge area ($\mu(0)$ is high), the likelihood of it remaining a surge zone decreases (due to increased supply). This then defines the transition probabilities as follows:

$$Pr(X^i_{t+1} = 1 | X^i_t = 0, U^i_t = 0, \mu) = 0.7 \times \mu(0) + 0.2$$
$$Pr(X^i_{t+1} = 1 | X^i_t = 0, U^i_t = 1, \mu) = 0.8$$

$$Pr(X_{t+1}^i = 1 | X_t^i = 1, U_t^i = 0, \mu) = 0.8$$
$$Pr(X_{t+1}^i = 1 | X_t^i = 1, U_t^i = 1, \mu) = 0.7 \times \mu(0) + 0.2$$

We set the parameters as $R = 1$, $S = 7$ and $\beta = 0.7$.

Near optimality of learned models and infinite population approximations. Figure 1 shows the value loss for different values of the number of agents in the system. We graph the loss functions under 3 settings:

- The optimal policy for the infinite population model with perfect knowledge of transition dynamics and costs.
- The estimate policy for the infinite population model, where the transition-cost function is learned using discretization basis functions based on the discretization of the measure space $\mathcal{P}(\mathbb{X})$ into 6 subsets (see Section 3.3).
- The estimate policy for the infinite population model, where the transition-cost function is learned using a class of basis functions:

  $$\phi(\mu) = [1, \mu(0), \mu(0)^2, \mu(0)^3, \sin(\mu(0)), \cos(\mu(0))].$$

  Note that the cost and the transitions are perfectly linear under the basis functions $[1, \mu(0), \mu(0)^2]$.

For the loss, we compare the value of the learned approximate policy with the optimal value in an infinite population environment. Furthermore, we assume that the initial distribution is $\mu_0 = 1/2\delta_0 + 1/2\delta_1$.

In the figure, we also plot a scaled $\frac{1}{\sqrt{N}}$ line which represents the decay rate of the empirical consistency term M_N defined in (37). As verified by the results, the loss in all cases decays at a rate similar to $\frac{1}{\sqrt{N}}$.

We also observe that the policies for the learned model with polynomial basis functions perform as well as the policies under perfect model knowledge, which is expected as the model is perfectly linear for these basis functions.

For the learned model under discretization, there is a small performance gap, which is also expected since the model is not perfectly linear under discretization basis functions. Thus, the learned model does not perfectly match the true model under discretization.

Lack of exploration without common randomness. Another significant observation from the previous sections about the exploration is also verified in this numerical study. In particular, when agents perform learning individually, we observe that the mean-field term tends to get stuck in certain regions without common randomness. However, if agents choose their actions based on a common randomness, then exploration becomes more efficient as seen in Figure 2. In the right graph, the agents follow a policy of the form $\gamma^i(\cdot|x, w^i)$ where w^i is an i.i.d. noise term that is independent across the agents which results in a deterministic flow of the mean-field term, and results in poor exploration. In the graph on the left, the agents follow exploration policies of the form $\gamma(\cdot|x, w^0)$ where w^0 is a common noise that is shared by all agents. As a result, the flow of the mean-field term becomes stochastic and a better exploration is observed.

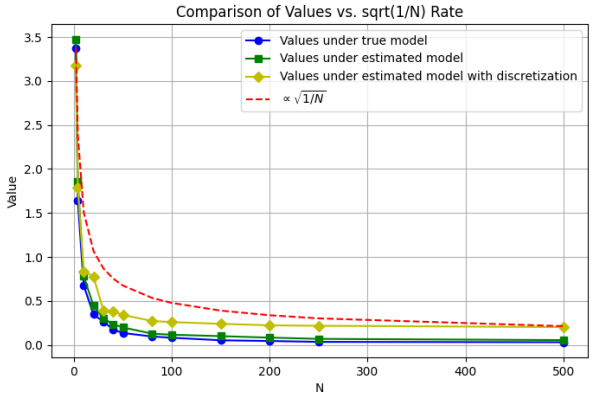

Figure 1: Value comparison under different policies.

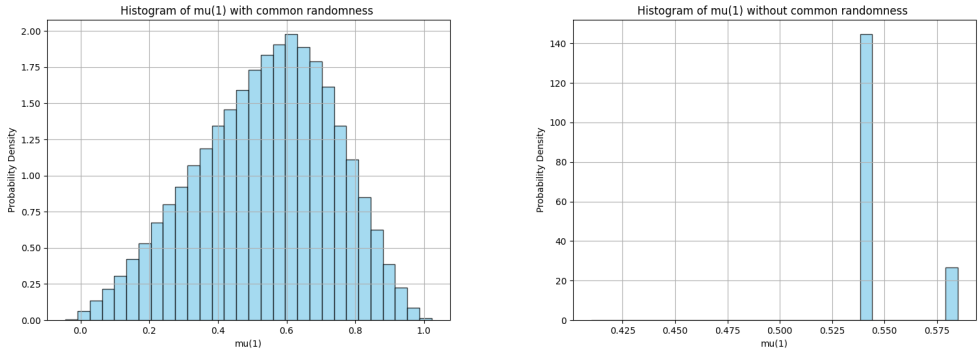

Figure 2: Learned regions with and without common noise.

6. Conclusion

We have studied model-based learning methods for mean-field control problems using linear function approximations, focusing on both fully coordinated and independent learning approaches. We have observed that full decentralization is generally not possible even when the agents agree on a common model. For the independent learning method, although agents do not need to share their local state information for the convergence, a certain level of coordination is inevitable especially for the exploration phase of the control problem which is done using a common noise process. For the learned models, we have provided error analysis which stems from two main sources (i) modeling mismatch due to linear function approximation, (ii) error arising from the infinite population approximation.

We have observed that the exploration is a key challenge in the learning of mean-field control systems. The analysis in the paper suggests that the stochastic controllability of the mean-field systems is closely related to the exploration problem. A natural future direction is then to further analysis of the controllability and exploration properties of the mean-field control system.

Acknowledgments

E. Bayraktar is partially supported by the National Science Foundation under grant DMS-2106556 and by the Susan M. Smith chair.

Appendix A. Proof of Proposition 2.1

Step 1. We first show that $E\left[\|v_t - v^*\|^2\right]$ remains uniformly bounded over t. We write

$$\|v_{t+1} - v^*\|^2 \leq \|v_t - v^*\|^2 - 2\alpha_t \langle \nabla g(s_t, v_t), v_t - v^* \rangle + \alpha_t^2 \nabla^2 g(s_t, v_t) \tag{40}$$

For $E[\nabla^2 g(s_t, v_t)]$ we have that

$$\begin{aligned} &E[\nabla^2 g(s_t, v_t)] = E\left[\left|2k(s_t)(k(s_t)^\intercal v_t - h(s_t))\right|^2\right] \leq KE\left[2\|k(s_t)\|^2\|v_t\|^2 + 2\|h(s_t)\|^2\right] \\ &\leq KE\left[\|v_t\|^2 + 1\right] \leq K\left(E\left[\|v_t - v^*\|^2\right] + 1\right) \end{aligned}$$

where the generic constant $K < \infty$ may represent different values at different steps. Denoting by $A_t := E\left[\|v_t - v^*\|^2\right]$, if we take the expectation on both sides of (40) we can write

$$\begin{aligned} A_{t+1} &\leq A_t - 2\alpha_t E\left[\langle \nabla g(s_t, v_t), v_t - v^* \rangle\right] + \alpha_t^2 K A_t + \alpha_t^2 K \\ &\leq A_t - 2\alpha_t E\left[g(s_t, v_t) - g(s_t, v^*)\right] + \alpha_t^2 K A_t + \alpha_t^2 K \end{aligned} \tag{41}$$

where at the last step we used the convexity of $g(s_t, v_t)$ for every s_t. We now introduce $\hat{s}_t$ which are independent over t and each $\hat{s}_t$ is distributed according to $\pi(\cdot)$. For the middle term above we write

$$\begin{aligned} -2\alpha_t E\left[g(s_t, v_t) - g(s_t, v^*)\right] = &- 2\alpha_t E\left[(g(s_t, v_t) - g(\hat{s}_t, v_t))\right] \\ &- 2\alpha_t E\left[(g(\hat{s}_t, v_t) - g(\hat{s}_t, v^*))\right] \\ &- 2\alpha_t E\left[(g(\hat{s}_t, v^*) - g(s_t, v^*))\right] \end{aligned} \tag{42}$$

where the expectation is with respect to the independent coupling between $s_t, \hat{s}_t$.

We denote by

$$\begin{aligned} b_t^1 &= -2\alpha_t E\left[(g(s_t, v_t) - g(\hat{s}_t, v_t))\right] \\ b_t^2 &= -2\alpha_t E\left[(g(\hat{s}_t, v_t) - g(\hat{s}_t, v^*))\right] \\ b_t^3 &= -2\alpha_t E\left[(g(\hat{s}_t, v^*) - g(s_t, v^*))\right]. \end{aligned}$$

For b_t^1, we consider its absolute value to and write

$$\begin{aligned} &|b_t^1| \leq 2\alpha_t E\left[|g(s_t, v_t) - g(\hat{s}_t, v_t)|\right] \\ &\leq 2\alpha_t E\left[\left|(k(s_t)^\intercal v_t - h(s_t))^2 - (k(\hat{s}_t)^\intercal v_t - h(\hat{s}_t))^2\right|\right] \\ &\leq 2\alpha_t E\left[\left|(k(s_t)^\intercal - k(\hat{s}_t)^\intercal)\, v_t\right| \left|(k(s_t)^\intercal + k(\hat{s}_t)^\intercal)\, v_t - h(s_t) - h(\hat{s}_t)\right|\right] \\ &\leq 2\alpha_t E\left[\left(2\|k\|_\infty \|k(s_t) - k(\hat{s}_t)\| \|v_t\|^2 + 2\|h\|_\infty \|k(s_t) - k(\hat{s}_t)\| \|v_t\|\right)\right] \\ &\leq 2\alpha_t K E\left[\|k(s_t) - k(\hat{s}_t)\|\right] E\left[\|v_t\|^2\right] + 2\alpha_t K E\left[K\|k(s_t) - k(\hat{s}_t)\|\right] E\left[\|v_t\|\right] \end{aligned}$$

$$\leq 2\alpha_t K E\left[\|k(s_t) - k(\hat{s}_t)\|\right] E\left[\|v_t - v^*\|^2\right] \tag{43}$$

where we used generic constant $K < \infty$ for the above analysis that might have different values at different steps. Furthermore, we used the inequality $\|v_t\|^2 \leq 2\|v_t - v^*\|^2 + 2\|v^*\|^2$. We also assume that $\|v_t\| \geq 1$ to use $\|v_t\| \leq \|v_t\|^2$, note that this is without loss of generality as we are trying to show that $E\|v_t - v^*\|^2$ is bounded, and for $\|v_t\| \leq 1$, the boundedness is immediate. For the following analysis, we will denote by

$$\epsilon_t := E\left[\|k(s_t) - k(\hat{s}_t)\|\right].$$

We now consider the series $\sum_{t=1}^{\infty} \alpha_t \epsilon_t$. Since s_t is ergodic with a geometric rate with invariant measure $\pi(\cdot)$ and $\hat{s}_t \sim \pi(\cdot)$, we have that

$$\sum_{t=1}^{\infty} \alpha_t \epsilon_t < \infty \tag{44}$$

We now go back to (41)

$$\begin{aligned} A_{t+1} &\leq A_t - 2\alpha_t E\left[g(s_t, v_t) - g(s_t, v^*)\right] + \alpha_t^2 K A_t + \alpha_t^2 K \\ &\leq A_t + |b_t^1| + b_t^2 + b_t^3 + \alpha_t^2 K A_t \\ &\leq A_t + 2\alpha_t K \epsilon_t A_t + 2\alpha_t K \epsilon_t + b_t^2 + b_t^3 + \alpha_t^2 K A_t + \alpha_t^2 K \\ &\leq (1 + 2\alpha_t K \epsilon_t + \alpha_t^2 K) A_t + b_t^2 + b_t^3 + \alpha_t^2 K. \end{aligned}$$

For the following we denote by

$$c_t = (1 + 2\alpha_t K \epsilon_t + \alpha_t^2 K)$$

Note that one can show the infinite product $\prod_{t=1}^{\infty} c_t$ converges if and only if the sum

$$\sum_{t=1}^{\infty} 2\alpha_t K \epsilon_t + \alpha_t^2 K$$

converges. We have shown that the sum $\sum_{t=1}^{\infty} \alpha_t \epsilon_t$ is convergent due to geometric ergodicity, and we also have that α_t^2 is summable. Thus, we write

$$\prod_{t=1}^{\infty} c_t < C$$

for some $C < \infty$. One can then iteratively show that

$$\begin{aligned} A_{t+1} &\leq \prod_{n=1}^{t} c_n A_0 + C \sum_{n=1}^{t} \left(b_n^2 + b_n^3 + \alpha_n^2 K\right) \\ &\leq C A_0 + C \sum_{n=1}^{t} \left(b_n^2 + b_n^3 + \alpha_n^2 K\right). \end{aligned}$$

Consider $b_n^2 = -2\alpha_n E\left[g(\hat{s}_n, v_n) - g(\hat{s}_n, v^*)\right]$; since $\hat{s}_t \sim \pi(\cdot)$ for all t, and since $v^* = \arg\min_v G(v) = \arg\min_v \int g(s,v)\pi(ds)$, $b_n^2 \leq 0$ for all n. Thus, we can simply remove b_n^2 terms to get a further upperbound. For b_n^3, we have that

$$\sum_{n=1}^{\infty} |b_n^3| \leq \sum_{t=1}^{\infty} 2\alpha_t \left|g(\hat{s}_t, v^*) - g(s_t, v^*)\right| < \infty$$

using an identical argument we used to show $\sum \alpha_t \epsilon_t < \infty$. In particular,

$$\lim_{t\to\infty} A_t \leq \lim_{t\to\infty} CA_0 + C\sum_{n=1}^{t} \left(b_n^3 + \alpha_n^2 K\right) < \infty$$

which shows that $E\|v_t - v^*\|^2$ is bounded uniformly over t, which also implies that $E\|v_t\|^2$ is bounded.

Step 2. Now we have the boundedness, we go back to (41); using the bound on A_t (only for the second A_t in (41)), and summing over the terms, we can write

$$A_{N+1} - A_0 \leq \sum_{t=1}^{N} \left(A_{t+1} - A_t\right) \leq \sum_{t=1}^{N} -2\alpha_t E\left[g(s_t, v_t) - g(s_t, v^*)\right] + \sum_{t=1}^{N} \alpha_t^2 K$$

again using the boundedness of A_t, and the fact that $\sum_{t=1}^{\infty} \alpha_t^2 < \infty$ and sending $N \to \infty$, we get

$$E\left[\sum_{t=1}^{\infty} 2\alpha_t \left(g(s_t, v_t) - g(s_t, v^*)\right)\right] < \infty$$

We now introduce $\hat{s}_t$ which are independent over t and each $\hat{s}_t$ is distributed according to $\pi(\cdot)$. We then write

$$\begin{aligned} &E\left[\sum_{t=1}^{\infty} \underbrace{2\alpha_t \left(g(s_t, v_t) - g(\hat{s}_t, v_t)\right)}_{b_t^1}\right] \\ &+ E\left[\sum_{t=1}^{\infty} \underbrace{2\alpha_t \left(g(\hat{s}_t, v_t) - g(\hat{s}_t, v^*)\right)}_{b_t^2}\right] \\ &+ E\left[\sum_{t=1}^{\infty} \underbrace{2\alpha_t \left(g(\hat{s}_t, v^*) - g(s_t, v^*)\right)}_{b_t^3}\right] < \infty \end{aligned} \tag{45}$$

where we overwrite the definitions of b_t^1, b_t^2, and b_t^3 (only changing the signs of these terms, see (42)).

Recall the analysis for b_t^1 in (43), together with the uniform boundedness of $A_t = E\left\|v_t - v^*\|^2\right]$ over t, we can write that

$$E\left[\sum_{t=1}^{\infty} \left|b_t^1\right|\right] \leq \sum_{t=1}^{\infty} 2\alpha_t K E\left[\|k(s_t) - k(\hat{s}_t)\|\right] < \infty$$

where we exchange the sum and expectation with monotone convergence theorem, and where the last step follows from what we have shown in (44).

For the last term similarly, we have that $E\left[\sum_{t=1}^{\infty} b_t^3\right] < \infty$, from (44), since s_t is geometrically ergodic with invariant measure π and $\hat{s}_t \sim \pi(\cdot)$ and v^* is fixed.

Going back to (45), now that we have shown the last and the first terms are finite, we can write

$$E\left[\sum_{t=1}^{\infty} 2\alpha_t \left(g(\hat{s}_t, v_t) - g(\hat{s}_t, v^*)\right)\right] < \infty.$$

Since $\hat{s}_t$ is i.i.d and distributed according to $\pi(\cdot)$, the above also implies that

$$E\left[\sum_{t=1}^{\infty} 2\alpha_t \left(G(v_t) - G(v^*)\right)\right] < \infty$$

which in turn implies that

$$\sum_{t=1}^{\infty} 2\alpha_t \left(G(v_t) - G(v^*)\right) < \infty$$

almost surely. Furthermore, since α_t is not summable, and $(G(v_t) - G(v^*)) \geq 0$ (as v^* achieves the minimum of $G(v)$), we must have that

$$G(v_t) \to G(v^*), \text{ almost surely.}$$

Appendix B. Proof of Lemma 12

We begin the proof by writing the Bellman equations

$$\hat{K}_\beta(\mu) = \hat{k}(\mu, \hat{\gamma}) + \beta \hat{K}_\beta(\hat{F}(\mu, \hat{\gamma}))$$
$$K^*_\beta(\mu) = k(\mu, \gamma) + \beta K^*_\beta(F(\mu, \gamma))$$

where $\hat{\gamma}$ and γ are optimal agent-level policies that achieves the minimum at the right hand side of the Bellman equations respectively. We can use then use the same agent level policies by exchanging them to get the following upper-bound

$$\begin{aligned}
&\left|\hat{K}_\beta(\mu) - K^*_\beta(\mu)\right| \leq |\hat{k}(\mu, \gamma) - k(\mu, \gamma)| + \beta \left|\hat{K}_\beta(\hat{F}(\mu, \gamma)) - K^*_\beta(F(\mu, \gamma))\right| \\
&\leq |\hat{k}(\mu, \gamma) - k(\mu, \gamma)| + \beta \left|\hat{K}_\beta(\hat{F}(\mu, \gamma)) - K^*_\beta(\hat{F}(\mu, \gamma))\right| + \beta \left|K^*_\beta(\hat{F}(\mu, \gamma)) - K^*_\beta(F(\mu, \gamma))\right| \\
&\leq \lambda + \beta \sup_\mu \left|\hat{K}_\beta(\mu) - K^*_\beta(\mu)\right| + \beta \|K^*_\beta\|_{Lip} \|\hat{F}(\mu, \gamma) - F(\mu, \gamma)\| \qquad (46)
\end{aligned}$$

We have that

$$\begin{aligned}
\hat{F}(\mu, \gamma) - F(\mu, \gamma)\| &\leq \left\| \int \hat{\mathcal{T}}(\cdot|x, u, \mu)\gamma(du|x)\mu(dx) - \int \mathcal{T}(\cdot|x, u, \mu)\gamma(du|x)\mu(dx) \right\| \\
&\leq \lambda.
\end{aligned}$$

Hence, by rearranging the terms in (46), we can write

$$\left|\hat{K}_\beta(\mu) - K^*_\beta(\mu)\right| \leq \frac{\lambda}{1-\beta}\left(1 + \beta\|K_\beta\|_{Lip}\right).$$

Finally, a slight modification of (Bayraktar et al., 2025, Lemma 6) for finite $\mathbb{X}, \mathbb{U}$ can be used to show that

$$\|K^*_\beta\|_{Lip} \leq \frac{C}{1-\beta K}$$

which completes the proof that

$$\left|\hat{K}_\beta(\mu) - K^*_\beta(\mu)\right| \leq \lambda\left(\frac{\beta C - \beta K + 1}{(1-\beta)(1-\beta K)}\right).$$

Appendix C. Proof of Lemma 15

We use the notation $\mu_t = \mu_{\mathbf{x_t}}$ for the following analysis. Note that with stochastic realization results, there exists a random variable v_t uniformly distributed on $[0,1]$, and a measurable function $\hat{\gamma}$ such that

$$\hat{\gamma}(x, v_t)$$

has the same distribution as $\hat{\gamma}(\cdot|x, \hat{\mu}_t)$, where we overwrite the notation for simplicity. Let $\hat{\mathbf{x}}_t$ denote a vector of size N state variables that are distributed according to $\hat{\mu}_t$, i.e. $\hat{\mathbf{x}}_t = [\hat{x}_t^1, \ldots, \hat{x}_t^N]$ such that $\hat{x}_t^i \sim \hat{\mu}_t$ for all $i \in \{1, \ldots, N\}$. Furthermore, let $\mathbf{v}_t$ denote a vector of size N where each element is independent and distributed according to the law of v_t. We then study the following conditional expected difference:

$$E\left[\left\|\mu_{\mathbf{x_{t+1}}} - \hat{\mu}_{t+1}\right\|\right] = E\left[E\left[\left\|\mu_{\mathbf{x_{t+1}}} - \hat{\mu}_{t+1}\right\| |\mathbf{x}_t, \hat{\mathbf{x}}_\mathbf{t}, \mathbf{v_t}\right]\right].$$

Let $\mathbf{w_t}$ denote the vector of size N for the noise variables of the agents at time t. Note that we have $\mathbf{x}_{t+1} = f(\mathbf{x}_t, \mathbf{u}_t, \mathbf{w}_t)$ where $u_t^i = \hat{\gamma}(x_t^i, v_t^i)$ for each i. We also introduce $\hat{\mathbf{u}}_t$ such that $\hat{u}_t^i = \hat{\gamma}(\hat{x}_t^i, v_t^i)$.

We further introduce another vector of noise variables $\hat{\mathbf{w}}_t$ where each element is independently distributed, and the distribution of $\hat{w}_t$ agrees the with the kernel $\hat{\mathcal{T}}(\cdot|x, u, \hat{\mu}_t)$. In other words, we use the functional representation of $\hat{\mathcal{T}}(\cdot|x, u, \hat{\mu}_t)$ where

$$\hat{f}(x, u, \hat{\mu}_t, \hat{w}_t) \sim \hat{\mathcal{T}}(\cdot|x, u, \hat{\mu}_t)$$

for some measurable $\hat{f}$.

We denote by $\mathbf{P}(d\mathbf{w}_t) = P(dw_t^1) \times \cdots \times P(dw_t^N)$ denote the distribution of the vector $\mathbf{w}_t$ where it is assumed that w_t^i and w_t^j are independent for all $i \neq j$. $\hat{\mathbf{w}}_t$ is also distributed according to $\mathbf{P}(\cdot)$. For the joint distribution of $\mathbf{w_t}, \hat{\mathbf{w}}_\mathbf{t}$, we use a coupling of the form

$$\mathbf{\Omega}(d\mathbf{w}_t, d\hat{\mathbf{w}}_t) = \Omega^1(dw_t^1, d\hat{w}_t^1) \times \cdots \times \Omega^N(dw_t^N, d\hat{w}_t^N).$$

That is, we assume independence over $i \in 1, \ldots, N$, however, an arbitrary coupling is assumed between the distribution of $w_t^i, \hat{w}_t^i$. We will later specify the particular selection of coordinate wise

couplings $\Omega^1, \dots, \Omega^N$, however, the following analysis will hold correct for a general selection of $\Omega^1, \dots, \Omega^N$.

For given realizations of $\mathbf{x}_t, \hat{\mathbf{x}}_\mathbf{t}, \mathbf{v}_\mathbf{t}$, we write

$$
\begin{aligned}
&E\left[\left\|\mu_{\mathbf{x_{t+1}}} - \hat{\mu}_{t+1}\right\| |\mathbf{x}_t, \hat{\mathbf{x}}_\mathbf{t}, \mathbf{v_t}\right] = \int \left\|\mu_{f(\mathbf{x}_t,\mathbf{u}_t,\mathbf{w}_t)} - \hat{\mu}_{t+1}\right\| P(d\mathbf{w}_t) \\
&= \int \left\|\mu_{f(\mathbf{x}_t,\mathbf{u}_t,\mathbf{w}_t)} - \hat{\mu}_{t+1}\right\| \mathbf{\Omega}(d\mathbf{w}_t, d\hat{\mathbf{w}}_t) \\
&\leq \int \left\|\mu_{f(\mathbf{x}_t,\mathbf{u}_t,\mathbf{w}_t)} - \mu_{\hat{f}(\hat{\mathbf{x}}_t,\hat{\mathbf{u}}_t,\hat{\mathbf{w}}_t)}\right\| \mathbf{\Omega}(d\mathbf{w}_t, d\hat{\mathbf{w}}_t) + \int \left\|\mu_{\hat{f}(\hat{\mathbf{x}}_t,\hat{\mathbf{u}}_t,\hat{\mathbf{w}}_t)} - \hat{\mu}_{t+1}\right\| \mathbf{\Omega}(d\mathbf{w}_t, d\hat{\mathbf{w}}_t) \quad (47)
\end{aligned}
$$

Note that $\hat{\mathbf{x}}_t$ is a vector of size N where each entry is independent and distributed according to $\hat{\mu}_t$. Furthermore, $\hat{u}_t^i = \hat{\gamma}(\hat{x}_t^i, v_t^i)$, and thus $\hat{u}_t^i \sim \hat{\gamma}(\cdot|\hat{x}_t^i, \hat{\mu}_t)$ for each $i \in \{1, \dots, N\}$. Thus, $\mu_{\hat{f}(\hat{\mathbf{x}}_t,\hat{\mathbf{u}}_t,\hat{\mathbf{w}}_t)}$ is an empirical measure for $\hat{\mu}_{t+1}$ For the second term above, we then have:

$$
\begin{aligned}
&E\left[\int \left\|\mu_{\hat{f}(\hat{\mathbf{x}}_t,\hat{\mathbf{u}}_t,\hat{\mathbf{w}}_t)} - \hat{\mu}_{t+1}\right\| \mathbf{\Omega}(d\mathbf{w}_t, d\hat{\mathbf{w}}_t)\right] = E\left[\int \left\|\mu_{\hat{f}(\hat{\mathbf{x}}_t,\hat{\mathbf{u}}_t,\hat{\mathbf{w}}_t)} - \hat{\mu}_{t+1}\right\| P(d\hat{\mathbf{w}}_t)\right] \\
&= E\left[E\left[\left\|\mu_{\hat{\mathbf{x}}_{t+1}} - \hat{\mu}_{t+1}\right\| |\hat{\mathbf{x}}_t, \hat{\mathbf{u}}_t\right]\right] = E\left[\|\mu_{\hat{\mathbf{x}}_{t+1}} - \hat{\mu}_{t+1}\|\right] \leq M_N \quad (48)
\end{aligned}
$$

see (37) for the definition of M_N.

For the first term in (47); we note that $\mu_{f(\mathbf{x}_t,\mathbf{u}_t,\mathbf{w}_t)}$ and $\mu_{f(\hat{\mathbf{x}}_t,\hat{\mathbf{u}}_t,\hat{\mathbf{w}}_t)}$ are empirical measures, and thus for every given realization of $\mathbf{w}_t$ and $\hat{\mathbf{w}}_t$, the Wasserstein distance is achieved with a particular permutation of $f(\mathbf{x}_t, \mathbf{u}_t, \mathbf{w}_t)$ and $\hat{f}(\hat{\mathbf{x}}_t, \hat{\mathbf{u}}_t, \hat{\mathbf{w}}_t)$ combined together. That is, letting σ denote a permutation map for the vector $\hat{f}(\hat{\mathbf{x}}_t, \hat{\mathbf{u}}_t, \hat{\mathbf{w}}_t)$. we have

$$
\left\|\mu_{f(\mathbf{x}_t,\mathbf{u}_t,\mathbf{w}_t)} - \mu_{\hat{f}(\hat{\mathbf{x}}_t,\hat{\mathbf{u}}_t,\hat{\mathbf{w}}_t)}\right\| = \inf_\sigma \frac{1}{N}\sum_{i=1}^N |f(x_t^i, u_t^i, w_t^i, \mu_{\mathbf{x_t}}) - \sigma(\hat{f}(\hat{x}_t^i, \hat{u}_t^i, \hat{w}_t^i, \hat{\mu}_t))|.
$$

We will however, consider a particular permutation where

$$
\begin{aligned}
&\left\|\int \mathcal{T}(\cdot|x, u, \mu_{\mathbf{x_t}})\mu_{(\mathbf{x_t},\mathbf{u_t})}(du, dx) - \hat{\mathcal{T}}(\cdot|x, u, \hat{\mu}_t)\mu_{(\hat{\mathbf{x}}_\mathbf{t},\hat{\mathbf{u}}_\mathbf{t})}(du, dx)\right\| \\
&= \frac{1}{N}\sum_{i=1}^N \left\|\mathcal{T}(\cdot|x_t^i, u_t^i, \mu_{\mathbf{x}_t}) - \sigma(\hat{\mathcal{T}}(\cdot|\hat{x}_t^i, \hat{u}_t^u, \hat{\mu}_t))\right\| \quad (49)
\end{aligned}
$$

For the following analysis, we will drop the permutation notation σ and assume that the given order of $\hat{f}(\hat{\mathbf{x}}_t, \hat{\mathbf{u}}_t, \hat{\mathbf{w}}_t)$ achieves the Wasserstein distance in (49). Furthermore, the coupling $\mathbf{\Omega}$ is assumed to have the same order of coordinate-wise coupling.

We then write

$$
\begin{aligned}
&\int \left\|\mu_{f(\mathbf{x}_t,\mathbf{u}_t,\mathbf{w}_t)} - \mu_{\hat{f}(\hat{\mathbf{x}}_t,\hat{\mathbf{u}}_t,\hat{\mathbf{w}}_t)}\right\| \mathbf{\Omega}(d\mathbf{w}_t, d\hat{\mathbf{w}}_t) \\
&\leq \int \frac{1}{N}\sum_{i=1}^N \left|f(x_t^i, u_t^i, w_t^i, \mu_{\mathbf{x_t}}) - \hat{f}(\hat{x}_t^i, \hat{u}_t^i, \hat{w}_t^i, \hat{\mu}_t)\right| \mathbf{\Omega}(d\mathbf{w}_t, d\hat{\mathbf{w}}_t) \\
&= \frac{1}{N}\sum_{i=1}^N \int \left|f(x_t^i, u_t^i, w_t^i, \mu_{\mathbf{x_t}}) - \hat{f}(\hat{x}_t^i, \hat{u}_t^i, \hat{w}_t^i, \hat{\mu}_t)\right| \mathbf{\Omega}(d\mathbf{w}_t, d\hat{\mathbf{w}}_t)
\end{aligned}
$$

$$= \frac{1}{N}\sum_{i=1}^{N}\int \left| f(x_t^i, u_t^i, w_t^i, \mu_{\mathbf{x_t}}) - \hat{f}(\hat{x}_t^i, \hat{u}_t^i, \hat{w}_t^i, \hat{\mu}_t)\right| \Omega^i(dw_t^i, d\hat{w}_t^i). \tag{50}$$

The analysis thus far, works for any coupling $\mathbf{\Omega}(d\mathbf{w}_t, d\hat{\mathbf{w}}_t)$. In particular, the analysis holds for the coupling that satisfies

$$\|\mathcal{T}(\cdot|x_t, u_t, \mu_{\mathbf{x_t}}) - \hat{\mathcal{T}}(\cdot|\hat{x}_t^i, \hat{u}_t^i, \hat{\mu}_t)\| = \int \left| f(x_t^i, u_t^i, w_t^i, \mu_{\mathbf{x_t}}) - \hat{f}(\hat{x}_t^i, \hat{u}_t^i, \hat{w}_t^i, \hat{\mu}_t)\right| \Omega^i(dw_t^i, d\hat{w}_t^i).$$

for every i for some coordinate-wise coupling $\Omega^i(dw_t^i, d\hat{w}_t^i)$. Continuing from the term (50), we can then write

$$\begin{aligned}
&\int \left\| \mu_{f(\mathbf{x}_t, \mathbf{u}_t, \mathbf{w}_t)} - \mu_{\hat{f}(\hat{\mathbf{x}}_t, \hat{\mathbf{u}}_t, \hat{\mathbf{w}}_t)}\right\| \mathbf{\Omega}(d\mathbf{w}_t, d\hat{\mathbf{w}}_t) \\
&\leq \frac{1}{N}\sum_{i=1}^{N}\int \left| f(x_t^i, u_t^i, w_t^i, \mu_{\mathbf{x_t}}) - \hat{f}(\hat{x}_t^i, \hat{u}_t^i, \hat{w}_t^i, \hat{\mu}_t)\right| \Omega^i(dw_t^i, d\hat{w}_t^i) \\
&= \int \frac{1}{N}\sum_{i=1}^{N} \left\| \mathcal{T}(\cdot|x_t, u_t, \mu_{\mathbf{x_t}}) - \hat{\mathcal{T}}(\cdot|\hat{x}_t^i, \hat{u}_t^i, \hat{\mu}_t)\right\| \\
&= \left\| \int \mathcal{T}(\cdot|x, u, \mu_{\mathbf{x}_t})\mu_{(\mathbf{x_t}, \mathbf{u_t})}(du, dx) - \int \hat{\mathcal{T}}(\cdot|x, u, \hat{\mu}_t)\mu_{(\hat{\mathbf{x}}_\mathbf{t}, \hat{\mathbf{u}}_\mathbf{t})}(du, dx)\right\|
\end{aligned}$$

where the last step follows from the particular permutation we consider (see (49)).

Furthermore, we also have that

$$\begin{aligned}
&\left\| \int \mathcal{T}(\cdot|x, u, \mu_{\mathbf{x}_t})\mu_{(\mathbf{x_t}, \mathbf{u_t})}(du, dx) - \int \hat{\mathcal{T}}(\cdot|x, u, \hat{\mu}_t)\mu_{(\hat{\mathbf{x}}_\mathbf{t}, \hat{\mathbf{u}}_\mathbf{t})}(du, dx)\right\| \\
&\leq \left\| \int \mathcal{T}(\cdot|x, u, \mu_{\mathbf{x}_t})\mu_{(\mathbf{x_t}, \mathbf{u_t})}(du, dx) - \int \mathcal{T}(\cdot|x, u, \mu_{\mathbf{x}_t})\mu_{(\hat{\mathbf{x}}_\mathbf{t}, \hat{\mathbf{u}}_\mathbf{t})}(du, dx)\right\| \\
&+ \left\| \int \mathcal{T}(\cdot|x, u, \mu_{\mathbf{x}_t})\mu_{(\hat{\mathbf{x}}_\mathbf{t}, \hat{\mathbf{u}}_\mathbf{t})}(du, dx) - \int \mathcal{T}(\cdot|x, u, \hat{\mu}_t)\mu_{(\hat{\mathbf{x}}_\mathbf{t}, \hat{\mathbf{u}}_\mathbf{t})}(du, dx)\right\| \\
&+ \left\| \int \mathcal{T}(\cdot|x, u, \hat{\mu}_t)\mu_{(\hat{\mathbf{x}}_\mathbf{t}, \hat{\mathbf{u}}_\mathbf{t})}(du, dx) - \int \hat{\mathcal{T}}(\cdot|x, u, \hat{\mu}_t)\mu_{(\hat{\mathbf{x}}_\mathbf{t}, \hat{\mathbf{u}}_\mathbf{t})}(du, dx)\right\| \\
&\leq \delta_T \|\mu_{\mathbf{x_t}} - \mu_{\hat{\mathbf{x}}_\mathbf{t}}\| + K_f \|\mu_{\mathbf{x_t}} - \hat{\mu}_t\| + \lambda
\end{aligned} \tag{51}$$

where for the first term we use the following bound:

$$\begin{aligned}
&\left\| \int \mathcal{T}(\cdot|x, u, \mu_{\mathbf{x}_t})\mu_{(\mathbf{x_t}, \mathbf{u_t})}(du, dx) - \int \mathcal{T}(\cdot|x, u, \mu_{\mathbf{x}_t})\mu_{(\hat{\mathbf{x}}_\mathbf{t}, \hat{\mathbf{u}}_\mathbf{t})}(du, dx)\right\| \\
&= \left\| \int \mathcal{T}(\cdot|x, \hat{\gamma}(x, v^i), \mu_{\mathbf{x}_t})\mu_{\mathbf{x_t}}(dx) - \int \mathcal{T}(\cdot|x, \hat{\gamma}(x, v^i), \mu_{\mathbf{x}_t})\mu_{\hat{\mathbf{x}}_\mathbf{t}}(dx)\right\| \\
&\leq \delta_T \|\mu_{\mathbf{x_t}} - \mu_{\hat{\mathbf{x}}_\mathbf{t}}\|
\end{aligned}$$

Combining (47), (48), and (51), we can then write

$$E\left[\left\|\mu_{\mathbf{x_{t+1}}} - \hat{\mu}_{t+1}\right\|\right] \leq M_N + \delta_T E[\|\mu_{\mathbf{x_t}} - \mu_{\hat{\mathbf{x}}_\mathbf{t}}\|] + K_f E[\|\mu_{\mathbf{x_t}} - \hat{\mu}_t\|] + \lambda$$

$$\leq (1+\delta_T)M_N + KE[\|\mu_{\mathbf{x_t}} - \hat{\mu}_t\|] + \lambda$$

where $K = (\delta_T + K_f)$. Noting that we have assumed $\mu_{\mathbf{x_0}} = \hat{\mu}_0$, this bound implies that

$$E\left[\|\mu_{\mathbf{x_t}} - \hat{\mu}_t\|\right] \leq \sum_{n=0}^{t-1} K^n(\lambda + 2M_N)$$

where we have used the fact that $\delta_T \leq 1$ to simplify the notation.

Appendix D. Proof of Lemma 16

We start by writing the Bellman equations:

$$K^*_\beta(\mu^N) = k(\mu^N, \gamma_\infty) + \beta K^*_\beta(F(\mu^N, \gamma_\infty))$$
$$K^{N,*}_\beta(\mu^N) = k(\mu^N, \Theta^N) + \beta \int K^{N,*}_\beta(\mu^N_1)\eta(d\mu^N_1|\mu^N, \Theta^N)$$

where we assume that an optimal selector for the infinite population problem at μ^N is γ_∞ such that the agents should use the randomized agent-level policy $\gamma_\infty(\cdot|x, \mu^N)$. For the N-agent problem, we assume that an optimal state-action distribution at μ^N is given by some $\Theta^N \in \mathcal{P}_N(\mathbb{X} \times \mathbb{U})$, which can be achieved by some $\mathbf{x}, \mathbf{u}$, such that $\mu_{\mathbf{x}} = \mu^N$ and $\mu_{(\mathbf{x},\mathbf{u})} = \Theta^N$.

We first assume that $K^*_\beta(\mu^N) > K^{N,*}_\beta(\mu^N)$. For the infinite population problem, instead of using the optimal selector γ_∞, we use a randomized agent-level policy from the finite population problem by writing $\Theta^N(du, dx) = \gamma^N(du|x)\mu^N(dx)$, and letting the agents use γ^N. We emphasize that the optimal state action distribution for N-agents is not achieved if each agent symmetrically use $\gamma^N(du|x)$, in other words, γ^N is not an optimal agent-level policy for the N-population problem. To have the equality $\Theta^N(du, dx) = \gamma_N(du|x)\mu^N(dx)$ the number of agents needs to tend to infinity. We can then write

$$K^*_\beta(\mu^N) - K^{N,*}_\beta(\mu^N) \leq K_\beta(\mu^N, \gamma^N) - K^{N,*}_\beta(\mu^N)$$
$$= k(\mu^N, \gamma^N) - k(\mu^N, \Theta^N) + \beta K^*_\beta\left(F(\mu^N, \gamma^N)\right) - \beta \int K^{N,*}_\beta(\mu^N_1)\eta(d\mu^N_1|\mu^N, \Theta^N)$$

Note that

$$k(\mu^N, \gamma^N) = \int c(x, u, \mu^N)\gamma^N(du|x)\mu^N(dx) = \int c(x, u, \mu^N)\Theta^N(du, dx) = k(\mu^N, \Theta^N).$$

Hence, we can continue:

$$K^*_\beta(\mu^N) - K^{N,*}_\beta(\mu^N) \leq \beta K^*_\beta\left(F(\mu^N, \gamma^N)\right) - \beta \int K^{N,*}_\beta(\mu^N_1)\eta(d\mu^N_1|\mu^N, \Theta^N)$$
$$\leq \beta \int \left|K^*_\beta\left(F(\mu^N, \gamma^N)\right) - K^*_\beta(\mu^N_1)\right| \eta(d\mu^N_1|\mu^N, \Theta^N)$$
$$+ \beta \int \left|K^*_\beta(\mu^N_1) - K^{N,*}_\beta(\mu^N_1)\right| \eta(d\mu^N_1|\mu^N, \Theta^N)$$
$$\leq \beta\|K^*_\beta\|_{Lip} \int \left\|F(\mu^N, \gamma^N) - \mu^N_1\right\| \eta(d\mu^N_1|\mu^N, \Theta^N) + \beta \sup_{\mu \in \mathcal{P}_N(\mathbb{X})} \left|K^*_\beta(\mu) - K^{N,*}_\beta(\mu)\right|. \quad (52)$$

We now focus on the term $\int \left\|F(\mu^N,\gamma^N)-\mu_1^N\right\| \eta(d\mu_1^N|\mu^N,\Theta^N)$. We will follow a very similar methodoly as we have used in the proof of Lemma 15 with slight differences. We denote by $\mathbf{P}(d\mathbf{w}) = P(dw^1)\times\cdots\times P(dw^N)$ denote the distribution of the vector $\mathbf{w}$ where it is assumed that w^i and w^j are independent for all $i\neq j$. Let $\mathbf{x},\mathbf{u}$ such that $\mu_{\mathbf{x}}=\mu^N$ and $\mu_{(\mathbf{x},\mathbf{u})}=\Theta^N$. We then have that

$$\int \left\|F(\mu^N,\gamma^N)-\mu_1^N\right\| \eta(d\mu_1^N|\mu^N,\Theta^N) = \int \left\|F(\mu^N,\gamma^N)-\mu_{f(\mathbf{x},\mathbf{u},\mathbf{w})}\right\| \mathbf{P}(d\mathbf{w})$$

where $f(\mathbf{x},\mathbf{u},\mathbf{w}) = [f(x^1,u^1,w^1,\mu^N),\dots,f(x^N,u^N,w^N,\mu^N)]$. We now introduce $(\hat{x}^i,\hat{u}^i)\sim\Theta^N(du,dx)$ where $i\in\{1,\dots,N\}$, which are different than the state action vectors $(\mathbf{x},\mathbf{u})$ and $\mu_{(\hat{\mathbf{x}},\hat{\mathbf{u}})}$ forms on empirical measure for Θ^N whereas $\mu_{\hat{\mathbf{x}}}$ forms an empirical measure for μ^N. We further introduce $\hat{\mathbf{w}}=[\hat{w}^1,\dots,\hat{w}^N]$. $\hat{\mathbf{w}}$ is also distributed according to $\mathbf{P}(\cdot)$. For the joint distribution of $\mathbf{w},\hat{\mathbf{w}}$, we use a coupling of the form

$$\mathbf{\Omega}(d\mathbf{w},d\hat{\mathbf{w}}) = \Omega^1(dw^1,d\hat{w}^1)\times\cdots\times\Omega^N(dw^N,d\hat{w}^N).$$

That is, we assume independence over $i\in 1,\dots,N$, however, an arbitrary coupling is assumed between the distribution of $w^i,\hat{w}^i$. We will later specify the particular selection of coordinate wise couplings $\Omega^1,\dots,\Omega^N$. We write

$$\begin{aligned}
&\int \left\|F(\mu^N,\gamma^N)-\mu_{f(\mathbf{x},\mathbf{u},\mathbf{w})}\right\| \mathbf{P}(d\mathbf{w})\\
&\leq E\left[\int \left\|F(\mu^N,\gamma^N)-\mu_{f(\hat{\mathbf{x}},\hat{\mathbf{u}},\hat{\mathbf{w}})}\right\| + \left\|\mu_{f(\hat{\mathbf{x}},\hat{\mathbf{u}},\hat{\mathbf{w}})}-\mu_{f(\mathbf{x},\mathbf{u},\mathbf{w})}\right\| \mathbf{\Omega}(d\mathbf{w},d\hat{\mathbf{w}})\right]
\end{aligned}$$

where the expectation is with respect to the random realizations of $(\hat{x}^i,\hat{u}^i)\sim\Theta^N(du,dx)$. The first term corresponds to the expected difference between the empirical measures of $\mu_1 = F(\mu^N,\gamma^N)$ and μ_1 itself, and thus is bounded by M_N.

For the second term, we note that $\mu_{f(\mathbf{x},\mathbf{u},\mathbf{w})}$ and $\mu_{f(\hat{\mathbf{x}},\hat{\mathbf{u}},\hat{\mathbf{w}})}$ are empirical measures, and thus for every given realization of $\mathbf{w}$ and $\hat{\mathbf{w}}$, the Wasserstein distance is achieved with a particular permutation of $f(\mathbf{x},\mathbf{u},\mathbf{w})$ and $f(\hat{\mathbf{x}},\hat{\mathbf{u}},\hat{\mathbf{w}})$ combined together. That is, letting σ denote a permutation map for the vector $f(\hat{\mathbf{x}},\hat{\mathbf{u}},\hat{\mathbf{w}})$. we have

$$\left\|\mu_{f(\mathbf{x},\mathbf{u},\mathbf{w})}-\mu_{\hat{f}(\hat{\mathbf{x}},\hat{\mathbf{u}},\hat{\mathbf{w}})}\right\| = \inf_{\sigma}\frac{1}{N}\sum_{i=1}^N |f(x^i,u^i,w^i,\mu^N)-\sigma(f(\hat{x}^i,\hat{u}^i,\hat{w}^i,\mu^N))|.$$

We will however, consider a particular permutation where

$$\begin{aligned}
&\left\|\int \mathcal{T}(\cdot|x,u,\mu^N)\mu_{(\mathbf{x},\mathbf{u})}(du,dx) - \mathcal{T}(\cdot|x,u,\mu^N)\mu_{(\hat{\mathbf{x}},\hat{\mathbf{u}})}(du,dx)\right\|\\
&= \frac{1}{N}\sum_{i=1}^N \left\|\mathcal{T}(\cdot|x^i,u^i,\mu^N)-\sigma(\mathcal{T}(\cdot|\hat{x}^i,\hat{u}^u,\mu^N))\right\|
\end{aligned}$$

For the following analysis, we will drop the permutation notation σ and assume that the given order of $f(\hat{\mathbf{x}},\hat{\mathbf{u}},\hat{\mathbf{w}})$ achieves the Wasserstein distance above. Furthermore, the coupling $\mathbf{\Omega}$ is assumed to have the same order of coordinate-wise coupling.

We then write

$$\begin{aligned}
&\int \left\|\mu_{f(\mathbf{x},\mathbf{u},\mathbf{w})} - \mu_{f(\hat{\mathbf{x}},\hat{\mathbf{u}},\hat{\mathbf{w}})}\right\| \mathbf{\Omega}(d\mathbf{w}, d\hat{\mathbf{w}}) \\
&\leq \int \frac{1}{N}\sum_{i=1}^{N} \left|f(x^i,u^i,w^i,\mu^N) - f(\hat{x}^i,\hat{u}^i,\hat{w}^i,\mu^N)\right| \mathbf{\Omega}(d\mathbf{w}, d\hat{\mathbf{w}}) \\
&= \frac{1}{N}\sum_{i=1}^{N} \int \left|f(x^i,u^i,w^i,\mu^N) - f(\hat{x}^i,\hat{u}^i,\hat{w}^i,\mu^N)\right| \mathbf{\Omega}(d\mathbf{w}, d\hat{\mathbf{w}}) \\
&= \frac{1}{N}\sum_{i=1}^{N} \int \left|f(x^i,u^i,w^i,\mu^N) - f(\hat{x}^i,\hat{u}^i,\hat{w}^i,\mu^N)\right| \Omega^i(dw^i, d\hat{w}^i).
\end{aligned}$$

The analysis thus far, works for any coupling $\mathbf{\Omega}(d\mathbf{w}, d\hat{\mathbf{w}})$. In particular, the analysis holds for the coupling that satisfies

$$\|\mathcal{T}(\cdot|x,u,\mu^N) - \mathcal{T}(\cdot|\hat{x}^i,\hat{u}^i,\mu^N)\| = \int \left|f(x^i,u^i,w^i,\mu^N) - f(\hat{x}^i,\hat{u}^i,\hat{w}^i,\mu^N)\right| \Omega^i(dw^i, d\hat{w}^i).$$

for every i for some coordinate-wise coupling $\Omega^i(dw^i, d\hat{w}^i)$. We can then write

$$\begin{aligned}
&\int \left\|\mu_{f(\mathbf{x},\mathbf{u},\mathbf{w})} - \mu_{f(\hat{\mathbf{x}},\hat{\mathbf{u}},\hat{\mathbf{w}})}\right\| \mathbf{\Omega}(d\mathbf{w}, d\hat{\mathbf{w}}) \\
&\leq \frac{1}{N}\sum_{i=1}^{N} \int \left|f(x^i,u^i,w^i,\mu^N) - f(\hat{x}^i,\hat{u}^i,\hat{w}^i,\mu^N)\right| \Omega^i(dw^i, d\hat{w}^i) \\
&= \int \frac{1}{N}\sum_{i=1}^{N} \left\|\mathcal{T}(\cdot|x,u,\mu^N) - \mathcal{T}(\cdot|\hat{x}^i,\hat{u}^i,\mu^N)\right\| \\
&= \left\| \int \mathcal{T}(\cdot|x,u,\mu^N)\mu_{(\mathbf{x},\mathbf{u})}(du,dx) - \int \mathcal{T}(\cdot|x,u,\mu^N)\mu_{(\hat{\mathbf{x}},\hat{\mathbf{u}})}(du,dx)\right\|.
\end{aligned}$$

We can then write that

$$\begin{aligned}
&\int \left\|F(\mu^N,\gamma^N) - \mu_{f(\mathbf{x},\mathbf{u},\mathbf{w})}\right\| \mathbf{P}(d\mathbf{w}) \\
&\leq E\left[\int \left\|F(\mu^N,\gamma^N) - \mu_{f(\hat{\mathbf{x}},\hat{\mathbf{u}},\hat{\mathbf{w}})}\right\| + \left\|\mu_{f(\hat{\mathbf{x}},\hat{\mathbf{u}},\hat{\mathbf{w}})} - \mu_{f(\mathbf{x},\mathbf{u},\mathbf{w})}\right\| \mathbf{\Omega}(d\mathbf{w}, d\hat{\mathbf{w}})\right] \\
&\leq M_N + E\left[\left\| \int \mathcal{T}(\cdot|x,u,\mu^N)\mu_{(\mathbf{x},\mathbf{u})}(du,dx) - \int \mathcal{T}(\cdot|x,u,\mu^N)\mu_{(\hat{\mathbf{x}},\hat{\mathbf{u}})}(du,dx)\right\|.\right] \\
&\leq M_N + E\left[\delta_T \left\|\mu_{(\hat{\mathbf{x}},\hat{\mathbf{u}})} - \mu_{(\mathbf{x},\mathbf{u})}\right\|\right] \leq M_N + \delta_T \bar{M}_N
\end{aligned}$$

where in the last step we used the fact that $\mu_{(\hat{\mathbf{x}},\hat{\mathbf{u}})}$ is an empirical measure for $\mu_{(\mathbf{x},\mathbf{u})} = \Theta^N$.

We then conclude that for the term (52):

$$\begin{aligned}
&K^*_\beta(\mu^N) - K^{N,*}_\beta(\mu^N) \\
&\leq \beta\|K^*_\beta\|_{Lip} \int \left\|F(\mu^N,\gamma^N) - \mu^N_1\right\| \eta(d\mu^N_1|\mu^N,\Theta^N) + \beta \sup_{\mu\in\mathcal{P}_N(\mathbb{X})} \left|K^*_\beta(\mu) - K^{N,*}_\beta(\mu)\right|
\end{aligned}$$

$$\leq \beta \|K_\beta^*\|_{Lip} \left(M_N + \delta_T \bar{M}_N\right) + \beta \sup_{\mu \in \mathcal{P}_N(\mathbb{X})} \left| K_\beta^*(\mu) - K_\beta^{N,*}(\mu) \right| \tag{53}$$

We now assume that $K_\beta^*(\mu^N) < K_\beta^{N,*}(\mu^N)$. To get an upper bound similar to (52), for the finite population problem, we let agents to use the randomized policy γ_∞ that is optimal for the infinite population problem, instead of choosing actions that achieves Θ^N which is the optimal selection for the N population problem for the state distribution μ^N. Let $\mathbf{x}$ be such that $\mu_{\mathbf{x}} = \mu^N$, we introduce $\mathbf{u} = [u^1, \ldots, u^N]$ where $u^i = \gamma_\infty(x^i, v^i)$ for some i.i.d. v^i. Denoting by $\hat{\Theta}^N = \mu_{(\mathbf{x},\mathbf{u})}$, and following the steps leading to (52), we now write

$$\begin{aligned}
&K_\beta^{N,*}(\mu^N) - K_\beta^*(\mu^N) \leq \beta \int K_\beta^{N,*}(\mu_1^N) \eta(d\mu_1^N | \mu^N, \hat{\Theta}^N) - \beta K_\beta^*(F(\mu^N, \gamma_\infty)) \\
&\leq \beta \int \left| K_\beta^{N,*}\left(\mu_1^N\right) - K_\beta^*(\mu_1^N) \right| \eta(d\mu_1^N | \mu^N, \hat{\Theta}^N) \\
&\quad + \beta \int \left| K_\beta^*(\mu_1^N) - K_\beta^*\left(F(\mu^N, \gamma_\infty)\right) \right| \eta(d\mu_1^N | \mu^N, \hat{\Theta}^N) \\
&\leq \beta \sup_{\mu \in \mathcal{P}_N(\mathbb{X})} \left| K_\beta^*(\mu) - K_\beta^{N,*}(\mu) \right| + \beta \|K_\beta^*\|_{Lip} \int \left\| F(\mu^N, \gamma_\infty) - \mu_1^N \right\| \eta(d\mu_1^N | \mu^N, \hat{\Theta}^N)
\end{aligned} \tag{54}$$

Following almost identical steps as the first case, one can show that

$$\begin{aligned}
&\int \left\| F(\mu^N, \gamma_\infty) - \mu_1^N \right\| \eta(d\mu_1^N | \mu^N, \hat{\Theta}^N) \\
&\leq M_N + \delta_T E\left[\left\| \mu_{(\hat{\mathbf{x}},\hat{\mathbf{u}})} - \mu_{(\mathbf{x},\mathbf{u})} \right\| \right]
\end{aligned}$$

where $\hat{x}^i \sim \mu^N$, $\mu_{\mathbf{x}} = \mu^N$ and $u^i = \gamma_\infty(x^i, v^i)$, $\hat{u}^i = \gamma_\infty(\hat{x}^i, v^i)$, and the expectation above is with respect to the random selections of $\hat{x}^i$ and v^i. Note that u^i and $\hat{u}^i$ uses the same randomization v^i, hence averaging over the distribution of v^i, we can write that

$$\begin{aligned}
E\left[\left\| \mu_{(\hat{\mathbf{x}},\hat{\mathbf{u}})} - \mu_{(\mathbf{x},\mathbf{u})} \right\| \right] \leq & E\left[\| \gamma_\infty(du|x) \mu_{\mathbf{x}}(dx) - \gamma_\infty(du|x) \mu_{\hat{\mathbf{x}}}(dx) \| \right] \\
& \leq E\left[\| \mu_{\mathbf{x}} - \mu_{\hat{\mathbf{x}}} \| \right] \leq M_N.
\end{aligned}$$

In particular, we can conclude that the bound (54) can be concluded as:

$$K_\beta^*(\mu^N) - K_\beta^{N,*}(\mu^N) \leq \beta \|K_\beta^*\|_{Lip} \left(M_N + \delta_T M_N\right) + \beta \sup_{\mu \in \mathcal{P}_N(\mathbb{X})} \left| K_\beta^*(\mu) - K_\beta^{N,*}(\mu) \right|. \tag{55}$$

Thus, noting that $M_N \leq \bar{M}_N$, and combining (53) and (55), we can write

$$|K_\beta^*(\mu^N) - K_\beta^{N,*}(\mu^N)| \leq \beta \|K_\beta^*\|_{Lip} \left(\bar{M}_N + \delta_T \bar{M}_N\right) + \beta \sup_{\mu \in \mathcal{P}_N(\mathbb{X})} \left| K_\beta^*(\mu) - K_\beta^{N,*}(\mu) \right|.$$

Rearranging the terms and taking the supremum on the left hand side over $\mu^N \in \mathcal{P}_N(\mathbb{X})$, we can write

$$\sup_{\mu \in \mathcal{P}_N(\mathbb{X})} |K_\beta^*(\mu) - K_\beta^{N,*}(\mu)| \leq \frac{\beta \|K_\beta^*\|_{Lip} (1 - \delta_T) \bar{M}_N}{1 - \beta}$$

which proves the result together with $\|K_\beta^*\|_{Lip} \leq \frac{C}{1 - \beta K}$ and $\delta_T \leq 1$.

References

Yves Achdou, Pierre Cardaliaguet, François Delarue, Alessio Porretta, Filippo Santambrogio, Yves Achdou, and Mathieu Laurière. Mean field games and applications: Numerical aspects. *Mean Field Games: Cetraro, Italy 2019*, pages 249–307, 2020.

Berkay Anahtarci, Can Deha Kariksiz, and Naci Saldi. Q-learning in regularized mean-field games. *Dynamic Games and Applications*, pages 1–29, 2022.

Andrea Angiuli, Jean-Pierre Fouque, and Mathieu Laurière. Unified reinforcement q-learning for mean field game and control problems. *Mathematics of Control, Signals, and Systems*, 34(2): 217–271, 2022.

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

Nicole Bäuerle. Mean field Markov decision processes. *Applied Mathematics & Optimization*, 88 (1):12, 2023.

Erhan Bayraktar and Ali Devran Kara. Infinite horizon average cost optimality criteria for mean-field control. *SIAM Journal on Control and Optimization*, 62(5):2776–2806, 2024.

Erhan Bayraktar and Xin Zhang. Solvability of infinite horizon Mckean–Vlasov FBSDEs in mean field control problems and games. *Applied Mathematics & Optimization*, 87(1):13, 2023.

Erhan Bayraktar, Andrea Cosso, and Huyên Pham. Randomized dynamic programming principle and Feynman-Kac representation for optimal control of Mckean-Vlasov dynamics. *Transactions of the American Mathematical Society*, 370(3):2115–2160, 2018.

Erhan Bayraktar, Alekos Cecchin, and Prakash Chakraborty. Mean field control and finite agent approximation for regime-switching jump diffusions. *Applied Mathematics & Optimization*, 88 (2):36, 2023.

Erhan Bayraktar, Nicole Bäuerle, and Ali Devran Kara. Finite approximations for mean-field type multi-agent control and their near optimality. *Applied Mathematics & Optimization*, 92(1):1–46, 2025.

Alain Bensoussan, Jens Frehse, and Phillip Yam. *Mean field games and mean field type control theory*, volume 101. Springer, 2013.

René Carmona and François Delarue. Probabilistic analysis of mean-field games. *SIAM Journal on Control and Optimization*, 51(4):2705–2734, 2013.

René Carmona and Mathieu Laurière. Convergence analysis of machine learning algorithms for the numerical solution of mean field control and games i: The ergodic case. *SIAM Journal on Numerical Analysis*, 59(3):1455–1485, 2021.

René Carmona and Mathieu Laurière. Convergence analysis of machine learning algorithms for the numerical solution of mean field control and games: II the finite horizon case. *The Annals of Applied Probability*, 32(6):4065–4105, 2022.

René Carmona, Mathieu Laurière, and Zongjun Tan. Model-free mean-field reinforcement learning: mean-field mdp and mean-field Q-learning. *The Annals of Applied Probability*, 33(6B):5334–5381, 2023.

Diogo S. Carvalho, Francisco S. Melo, and Pedro A. Santos. A new convergent variant of q-learning with linear function approximation. *Advances in Neural Information Processing Systems*, 33: 19412–19421, 2020.

Mao Fabrice Djete, Dylan Possamaï, and Xiaolu Tan. McKean–Vlasov optimal control: the dynamic programming principle. *The Annals of Probability*, 50(2):791–833, 2022.

Romuald Elie, Julien Perolat, Mathieu Laurière, Matthieu Geist, and Olivier Pietquin. On the convergence of model free learning in mean field games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7143–7150, 2020.

Zuyue Fu, Zhuoran Yang, Yongxin Chen, and Zhaoran Wang. Actor-critic provably finds Nash equilibria of linear-quadratic mean-field games. 2020. URL `https://openreview.net/forum?id=H1lhqpEYPr`.

Maximilien Germain, Joseph Mikael, and Xavier Warin. Numerical resolution of Mckean-Vlasov FBSDEs using neural networks. *Methodology and Computing in Applied Probability*, pages 1–30, 2022.

Diogo A Gomes and João Saúde. Mean field games models—a brief survey. *Dynamic Games and Applications*, 4(2):110–154, 2014.

Haotian Gu, Xin Guo, Xiaoli Wei, and Renyuan Xu. Mean-field controls with q-learning for cooperative marl: convergence and complexity analysis. *SIAM Journal on Mathematics of Data Science*, 3(4):1168–1196, 2021.

Haotian Gu, Xin Guo, Xiaoli Wei, and Renyuan Xu. Dynamic programming principles for mean-field controls with learning. *Operations Research*, 2023.

Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. Learning mean-field games. *Advances in Neural Information Processing Systems*, 32, 2019.

O. Hernandez-Lerma. *Adaptive Markov control processes*, volume 79. Springer Science & Business Media, 2012.

Minyi Huang, Peter E Caines, and Roland P Malhamé. Large-population cost-coupled LQG problems with nonuniform agents: individual-mass behavior and decentralized epsilon -Nash equilibria. *IEEE transactions on automatic control*, 52(9):1560–1571, 2007.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on learning theory*, pages 2137–2143. PMLR, 2020.

Daniel Lacker. Limit theory for controlled McKean–Vlasov dynamics. *SIAM Journal on Control and Optimization*, 55(3):1641–1672, 2017.

Mathieu Laurière and Olivier Pironneau. Dynamic programming for mean-field type control. *Comptes Rendus Mathematique*, 352(9):707–713, 2014.

Mathieu Lauriere, Sarah Perrin, Sertan Girgin, Paul Muller, Ayush Jain, Theophile Cabannes, Georgios Piliouras, Julien Pérolat, Romuald Elie, Olivier Pietquin, et al. Scalable deep reinforcement learning algorithms for mean field games. In *International conference on machine learning*, pages 12078–12095. PMLR, 2022.

Francisco C. Melo, Sean P. Meyn, and Isabel M. Ribeiro. An analysis of reinforcement learning with function approximation. In *Proceedings of the 25th international conference on Machine learning*, pages 664–671, 2008.

Sean Meyn. The projected Bellman equation in reinforcement learning. *IEEE Transactions on Automatic Control*, 69(12), 2024.

Médéric Motte and Huyên Pham. Mean-field Markov decision processes with common noise and open-loop controls. *The Annals of Applied Probability*, 32(2):1421–1458, 2022.

Médéric Motte and Huyên Pham. Quantitative propagation of chaos for mean field Markov decision process with common noise. *Electronic Journal of Probability*, 28:1–24, 2023.

Barna Pásztor, Andreas Krause, and Ilija Bogunovic. Efficient model-based multi-agent mean-field reinforcement learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=gvcDSDYUZx`.

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.

Huyên Pham and Xiaoli Wei. Dynamic programming for optimal control of stochastic McKean–Vlasov dynamics. *SIAM Journal on Control and Optimization*, 55(2):1069–1101, 2017.

Lars Ruthotto, Stanley J Osher, Wuchen Li, Levon Nurbekyan, and Samy Wu Fung. A machine learning framework for solving high-dimensional mean field game and mean field control problems. *Proceedings of the National Academy of Sciences*, 117(17):9183–9193, 2020.

Naci Saldi, Tamer Basar, and Maxim Raginsky. Markov–nash equilibria in mean-field games with discounted cost. *SIAM Journal on Control and Optimization*, 56(6):4256–4287, 2018.

Naci Saldi, Tamer Başar, and Maxim Raginsky. Approximate nash equilibria in partially observed stochastic games with mean-field interactions. *Mathematics of Operations Research*, 44(3):1006–1033, 2019.

Jayakumar Subramanian and Aditya Mahajan. Reinforcement learning in stationary mean-field games. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 251–259, 2019.

Csaba Szepesvári and William D Smart. Interpolation-based q-learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 100, 2004.

Hamidou Tembine, Quanyan Zhu, and Tamer Başar. Risk-sensitive mean-field games. *IEEE Transactions on Automatic Control*, 59(4):835–850, 2013.

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.