Efficient Model-Based Multi-Agent Mean-Field Reinforcement Learning

Barna Pásztor *barna.pasztor@ai.ethz.ch*
ETH Zürich

Ilija Bogunovic *i.bogunovic@ucl.ac.uk*
University College London

Andreas Krause *krausea@ethz.ch*
ETH Zürich

Reviewed on OpenReview: *https://openreview.net/forum?id=gvcDSDYUZx*

Abstract

Learning in multi-agent systems is highly challenging due to several factors including the non-stationarity introduced by agents' interactions and the combinatorial nature of their state and action spaces. In particular, we consider the *Mean-Field Control* (MFC) problem which assumes an asymptotically infinite population of identical agents that aim to collaboratively maximize the collective reward. In many cases, solutions of an MFC problem are good approximations for large systems, hence, efficient learning for MFC is valuable for the analogous discrete agent setting with many agents. Specifically, we focus on the case of *unknown* system dynamics where the goal is to simultaneously optimize for the rewards and learn from experience. We propose an efficient *model-based* reinforcement learning algorithm, M^3–UCRL, that runs in episodes, balances between exploration and exploitation during policy learning, and *provably* solves this problem. Our main theoretical contributions are the first general regret bounds for model-based reinforcement learning for MFC, obtained via a novel mean-field type analysis. To learn the system's dynamics, M^3–UCRL can be instantiated with various statistical models, e.g., neural networks or Gaussian Processes. Moreover, we provide a practical parametrization of the core optimization problem that facilitates gradient-based optimization techniques when combined with differentiable dynamics approximation methods such as neural networks.

1 Introduction

Multi-Agent Reinforcement Learning (MARL) extends the scope of reinforcement learning (RL) to multiple agents acting in a shared system. It is receiving considerable attention given a great number of real-world applications including autonomous driving (Shalev-Shwartz et al., 2016), finance (Lee et al., 2002; Lee et al., 2007; Lehalle & Mouzouni, 2019), social science (Castelfranchi, 2001; Leibo et al., 2017), swarm motion (Almulla et al., 2017), e-sports (Vinyals et al., 2019; Pachocki et al., 2018), and traffic routing (El-Tantawy et al., 2013), to name a few. Despite the recent popularity, analyzing agents' performance in these systems remains a particularly challenging task for several reasons including non-stationarity, scalability, competing learning goals, and varying information availability of agents. In this work, we target the challenges of *non-stationarity* and *scalability.* From the perspective of an individual agent, the dynamics of the system are non-stationary, since other agents update their behaviour, which alters over time how the environment reacts to certain actions of the agent. Multi-agent systems also suffer from a *combinatorial nature* due to the exponential growth of the state and action space as the number of agents in the system increases.

To tackle the previous challenges, *Mean-Field Control* (MFC) exploits the insight that many relevant MARL problems involve a large number of *very similar* agents that are coordinated via a centralized controller.

As a motivating application, consider a ride-hailing service in which a central dispatcher coordinates the routes of many drivers around the city. MFC considers the limiting regime of controlling *infinitely many identical collaborative* agents, and utilizes the notion of mean-field approximation from statistical physics (Weiss, 1907) to model their interactions. Specifically, instead of focusing on the individual agents and their interactions, the state of the system is described via the agents' state distribution and the dynamics and rewards are formulated accordingly. Despite the introduced approximations, the solution to the MFC problem often remains adequate for the finite agent equivalent (Lacker, 2017).

The focus of this work is on learning the optimal policies in MFC when the underlying dynamics are *unknown*. In complex domains such as fleets of autonomous vehicles or robot swarms, this is often the case, and the models consequently need to be learnt from data, giving rise to a well-known exploration–exploitation dilemma. In this paper, we propose the *Model-Based Multi-Agent Mean-Field Upper-Confidence RL* algorithm (M^3–UCRL) that is provably efficient and can effectively trade-off between exploration and exploitation during policy learning. Similarly to other Model-Based Reinforcement Learning (MBRL) algorithms, M^3–UCRL collects data about the unknown dynamics online (i.e., by proposing and executing a policy in the true system) and estimates the possible dynamics the system might follow. To optimize the agent's actions, it simulates the unknown system by using the estimated dynamics. Chua et al. (2018) has shown that MBRL can solve challenging single-agent RL problems with low sample complexity. We transfer this property to multi-agent problems and show that M^3–UCRL can efficiently solve the MFC problem with unknown dynamics.

Related work. Our M^3–UCRL extends the class of Model-Based Reinforcement Learning (MBRL) algorithms that rely on the *optimism-in-the-face-of-uncertainty* principle. Algorithms based on this approach achieve optimal regret for tabular MDPs (Brafman & Tennenholtz, 2002; Auer et al., 2009; Efroni et al., 2019; Domingues et al., 2020), while in the continuous setting, Abbasi-Yadkori & Szepesvári (2011) and Jin et al. (2020) show $\tilde{O}(\sqrt{T})$ regret bounds for LQR and linear MDPs, respectively. The recently designed episodic model-based algorithms (Chowdhury & Gopalan, 2019; Kakade et al., 2020; Curi et al., 2020; Sessa et al., 2022) achieve similar regret in the kernelized setting. Sessa et al. (2022) consider the general multi-agent setting, and prove a regret bound with $O(N^{H/2})$ dependence on the number N of agents. In comparison, M^3–UCRL assumes an asymptotically large population of identical agents. Crucially, as we will show, its performance does *not* depend on the number of agents, making it more suitable for large systems.

Multi-Agent Reinforcement Learning (MARL) has seen tremendous progress in recent years as many real-world applications involve interactions among a number of agents. Buşoniu et al. (2008) provide an overview of classical and early MARL results. Several other MARL surveys focus on: non-stationarity induced by the agent interactions (Hernandez-Leal et al., 2017), deep MARL (Hernandez-Leal et al., 2019; Nguyen et al., 2020), theoretical foundations in Markov/stochastic/extensive-form games (Zhang et al., 2019) and cooperative MARL (OroojlooyJadid & Hajinezhad, 2019). Recently, Zhang et al. (2019) and Nguyen et al. (2020) recognize the lack of practical model-based MARL algorithms (with an early exception of the R-Max algorithm proposed by Brafman & Tennenholtz (2000; 2002) for two-player zero-sum Markov Games). We fill this gap by designing a practical model-based algorithm that considers a large population of agents, and that is compatible with deep models.

Modeling interaction via the mean-field approach in MARL has recently gained popularity due to the introduction of Mean-Field Games (MFG) (Lasry & Lions, 2006a;b; Huang et al., 2006; 2007). MFG consider the limiting regime of a competitive game among identical agents with the objective of reaching a Nash Equilibrium. Various works (Yin et al., 2010; 2013; Cardaliaguet & Hadikhanloo, 2017; Guéant et al., 2011; Bensoussan et al., 2013; Carmona & Delarue, 2013; Carmona et al., 2018; Gomes et al., 2014) mainly consider solving MFGs in continuous-time, while some very recent ones (Yang et al., 2017; 2018; Fu et al., 2020; Guo et al., 2019; 2020; Elie et al., 2020; Agarwal et al., 2019) analyze the problem from the discrete-time RL perspective. The progress of the field has been recently summarized by Laurière et al. (2022).

At the same time, cooperative discrete-time *Mean-Field Control*, the main focus of this paper, has received significantly less attention. Gast et al. (2012); Motte & Pham (2019); Gu et al. (2020; 2021) and Bäuerle (2021) consider planning with known dynamics. In the learning setting, Wang et al. (2020); Carmona et al. (2019b); Gu et al. (2021); Angiuli et al. (2022; 2021) extend the Q-Learning algorithm and show convergence for specific variants of MFC focusing on discrete state and action spaces. The mean-field extension of the linear-quadratic problem has been studied by Carmona et al. (2019a); Wang et al. (2021); Carmona et al. (2021) with policy optimisation methods. Subramanian & Mahajan (2019) also proposes a policy-gradient based algorithm that solves Mean-Field Games and Control problems locally. Additionally, Chen et al. (2021); Li et al. (2021); Wang et al. (2020) consider the approximation error between the Mean-Field and the related Multi-Agent Reinforcement Learning problems. These previous model-free methods are either sample inefficient, assume access to a simulator or consider finite state and action spaces. On the contrary, our algorithm is model-based, sample-efficient, works for continuous spaces and learns the policy online by performing exploration on a real system.

Main contributions. We design a novel *Model-based Multi-agent Mean-field Upper Confidence* RL (M^3–UCRL) algorithm for the centralized control problem of a large population of collaborative agents. M^3–UCRL is practical, performs exploration on a real system (requires no simulator and efficient in terms of samples), and is compatible with neural network policy learning. Our main contributions are the general theoretical regret bounds obtained via a novel mean-field-based analysis, that we also specialize to Gaussian Process models. In the main step of our theoretical analysis, we bound the distance between the mean-field distributions under the true and the approximated dynamics via the model's total estimated uncertainty. Finally, we pose the exploration via entropy maximization problem used in a recent MFG survey (Laurière et al., 2022) in the continuous state and action domain and demonstrate the performance of our algorithm in it. Our results show that M^3–UCRL is capable of finding close-to-optimal policies within a few episodes, in contrast to model-free algorithms that require at least six orders of magnitude more samples.

2 Problem Statement

We consider the episodic Mean-Field Control (MFC) problem with time-horizon H, compact state space $\mathcal{S} \subseteq \mathbb{R}^p$ and compact action space $\mathcal{A} \subseteq \mathbb{R}^q$ that are common for all the agents in the system. We index episodes with the variable t and use T to denote the number of episodes passed. In standard N-agent settings, the system's state in episode t at time $h \in \{0, \dots, H-1\}$ is described with individual agents' states $(s_{t,h}^{(1)}, \dots, s_{t,h}^{(N)}) \in \mathcal{S}^N$ which grows exponentially in the number of agents N and represents a common issue in scaling MARL algorithms. In MFCs, we assume that all agents are *identical* and the population is *asymptotically infinite*, i.e., $N \to \infty$, therefore, the agents' states can be described by the *mean-field distribution*:

$$\mu_{t,h}(s) = \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{\{s_{t,h}^{(i)} = s\}},$$

where $\mu_{t,h}$ belongs to the space of probability measures $\mathcal{P}(\mathcal{S})$ over $\mathcal{S}$. Due to the assumption of *identical* agents in MFC, we can focus on a *representative agent* from the population that interacts with the distribution of agents instead of individual agents and interactions.

System Dynamics. Before every episode t, the representative agent selects a policy profile $\boldsymbol{\pi}_t = (\pi_{t,0}, \dots, \pi_{t,H-1})$ where the H individual policies $\pi_{t,h} : \mathcal{S} \times \mathcal{P}(\mathcal{S}) \to \mathcal{A}$ are chosen from a set of admissible policies Π (e.g., we parametrize our polices via neural networks; see below). At every time h, the representative agent selects an action $a_{t,h} = \pi_{t,h}(s_{t,h}, \mu_{t,h})$. The environment then returns reward $r(s_{t,h}, a_{t,h}, \mu_{t,h})$, and the agent observes its new state $s_{t,h+1}$ and the mean-field distribution $\mu_{t,h+1}$.

In MFCs, the system's dynamics are typically given by a McKean-Vlasov type of stochastic processes and depend on the agent's state, action and the mean-field distribution:

$$s_{t,h+1} = f(s_{t,h}, a_{t,h}, \mu_{t,h}) + \omega_{t,h}, \tag{1}$$

where $\omega_{t,h}$ is an *i.i.d.* additive noise vector. This is in contrast to the standard single-agent model-based RL (cf. Chowdhury & Gopalan (2019); Curi et al. (2020); Kakade et al. (2020)), where the dynamics only depend on agent's action and state. Crucially, since the focus of this work is on *learning* in MFC, we assume that the true dynamics are *unknown*, and the goal is to explore the space and learn about f over a number of episodes. [1] To do so, the representative agent relies on the collected random observations,

[1] We expect our results to easily extend and account for unknown rewards by using the same modeling assumptions for the reward function (e.g., as in Chowdhury & Gopalan (2019)).

$\mathcal{D}_t = \{((s_{t,h}, a_{t,h}, \mu_{t,h}), s_{t,h+1})\}_{h=0}^{H-1}$, which come from the interaction with the true system (i.e., from policy rollouts) during episode t.

Finally, after every episode, we assume that the whole system is reset and the agent's initial state s_0 is drawn from a known initial distribution μ_0, i.e., $s_0 \sim \mu_0$, that remains the same in every episode.

Mean-field Flow. Since, in MFC, all agents are identical, interact in a common environment, and follow the *same policy* $\boldsymbol{\pi}_t = (\pi_{t,0}, \dots, \pi_{t,H-1})$, the subsequent mean-field distributions satisfy the following mean-field flow property as shown by Gu et al. (2020):

Lemma 1. *For a given initial distribution μ_0, dynamics f and policy $\boldsymbol{\pi} = (\pi_0, \dots, \pi_{h-1})$, the mean-field distribution trajectory $\{\mu_h\}_{h=0}^{H-1}$, follows*

$$\mu_{h+1}(ds') = \int_{s \in \mathcal{S}} \mu_h(ds)\mathbb{P}[s_{h+1} \in ds'], \tag{2}$$

where $\mu_h \in \mathcal{P}(\mathcal{S})$ for all $h \geq 0$, $a_h = \pi_h(s_h, \mu_h)$, s_{h+1} is given by Eq. (1), and $\mu_h(ds) = \mathbb{P}[s_h \in ds]$ under $\boldsymbol{\pi}$.

We provide the proof in Appendix B.2. To shorten the notation, we use $\Phi(\mu_{t,h}, \pi_{t,h}, f)$ to denote the mean-field transition function from Eq. (2), i.e., we have $\mu_{t,h+1} = \Phi(\mu_{t,h}, \pi_{t,h}, f)$. The lemma shows that the next mean-field distribution explicitly depends on the unknown f (since $s_{t,h+1} = f(s_{t,h}, a_{t,h}, \mu_{t,h}) + \omega_{t,h}$), policy played by the agents $\pi_{t,h}$, and previous state distribution $\mu_{t,h}$.

Performance metric. Given a policy profile $\boldsymbol{\pi} = (\pi_1, \dots, \pi_{H-1})$, the performance of the representative agent is measured via the expected cumulative reward:

$$\begin{aligned} \mathsf{J}(\boldsymbol{\pi}) &= \mathbb{E}\left[\sum_{h=0}^{H-1} r(s_h, a_h, \mu_h)\right] && (3a) \\ \text{s.t.} \quad a_h &= \pi_h(s_h, \mu_h), && (3b) \\ s_{h+1} &= f(s_h, a_h, \mu_h) + \omega_h, && (3c) \\ \mu_{h+1} &= \Phi(\mu_h, \pi_h, f), && (3d) \end{aligned}$$

where the expectation is taken over the noise in the transitions and initial distribution $s_0 \sim \mu_0$. By considering the representative agent's perspective, the goal is to discover a *socially* optimal policy $\boldsymbol{\pi}^*$ that maximizes the expected total reward, i.e.,

$$\boldsymbol{\pi}^* \in \arg\max_{\boldsymbol{\pi}} \mathsf{J}(\boldsymbol{\pi}) \tag{4}$$

In our theoretical analysis, we make the following assumptions regarding the system's true dynamics, reward function and the set of admissible policies. We use $z, z' \in \mathcal{Z} = \mathcal{S} \times \mathcal{A} \times \mathcal{P}(\mathcal{S})$ to denote (s, a, μ) and (s', a', μ'), respectively, and we make use of the Wasserstein-1 metric: $W_1(\mu, \mu') = \inf_\nu \int_{\mathcal{S}\times\mathcal{S}} \|x - y\|_2 d\nu(x, y)$, where ν is any probability measure on $\mathcal{S} \times \mathcal{S}$ with marginals μ and μ' (see Appendix A for useful properties of $W_1(\mu, \mu')$ that are used in our analysis).

Assumption 1. *The transition function f is L_f–Lipschitz-continuous, i.e., $\|f(z) - f(z')\|_2 \leq L_f d(z, z')$ where $d(z, z') := \|s - s'\|_2 + \|a - a'\|_2 + W_1(\mu, \mu')$ and $\omega_{t,h}$ are* i.i.d. *additive σ-sub-Gaussian noise vectors for all $t \geq 1$ and $h \in \{0, \dots, H-1\}$.*

Assumption 2. *The set of admissible policies Π consists of L_π–Lipschitz-continuous policies such that for any $\pi \in \Pi$: $\|\pi(s, \mu) - \pi(s', \mu')\|_2 \leq L_\pi(\|s - s'\|_2 + W_1(\mu, \mu'))$, and the reward function is L_r–Lipschitz-continuous, i.e., $|r(z) - r(z')| \leq L_r d(z, z')$.*

Regularity assumptions like these are standard in the single-agent model-based reinforcement learning literature (Jaksch et al., 2010; Curi et al., 2020; Chowdhury & Gopalan, 2019; Sessa et al., 2022) and mild since, in practice, the policy and reward function classes are typically chosen to satisfy the previous smoothness assumptions. In our experiments (see Section 5), we parametrize our policies with Neural Networks with Lipschitz continuous activations (e.g., tanh, ReLU and linear). We note that the Lipschitzness of such policies then follows from that of the activations when the network's weights are bounded (in practice, this can be done by directly bounding them or via regularization).

Remark 1. *The episodic Mean-Field Control problem relates closely to the finite-horizon case of the evolutive Mean-Field Game problem described in Laurière et al. (2022). Both settings consider mean-field distributions evolving over time depending on the agents' policies, however, the MFC problem focuses on collaborative agents while agents in the evolutive MFG are competitive. This distinction leads to different solution concepts of the problems; social welfare in MFC and Nash Equilibrium in MFG.*

3 The M^3–UCRL Algorithm

In this section, we tackle the Mean-Field Control Problem with unknown dynamics from Eq. (4) by relying on a *model-based* learning scheme. Our *Model-based Multi-agent Mean-field Upper Confidence* RL (M^3–UCRL) algorithm uses a statistical model to estimate the system's dynamics and to effectively trade-off between exploration and exploitation during policy learning. Before stating our algorithm, we consider the main properties of the considered dynamics models.

Statistical Model. Our representative agent learns about the unknown dynamics from the data collected during the interactions with the environment, i.e., the policy rollouts. In particular, we take a model-based perspective (see Algorithm 1), where the agent models the unknown dynamics and sequentially, after every episode t, updates and improves its model estimates based on the previous transition observations, i.e., $\mathcal{D}_1 \quad \cdots \quad \mathcal{D}_t$ where $\mathcal{D}_t = \{((s_{t,h}, a_{t,h}, \mu_{t,h}), s_{t,h+1})\}_{h=0}^{H-1}$ is the set of state transition observations in episode t. To reason about plausible models at every episode t, the representative agent can take a frequentist or Bayesian perspective. In the first case, it estimates the mean $\boldsymbol{m}_t : \mathcal{Z} \to \mathcal{S}$ and confidence $\boldsymbol{\Sigma}_t : \mathcal{Z} \to \mathsf{R}^{p \times p}$ functions. In the second case, the agent estimates the posterior distribution over dynamical models $p(\tilde{f}|\mathcal{D}_t)$, that leads to $\boldsymbol{m}_t(z) = \mathsf{E}_{\tilde{f} \quad p(\tilde{f}|\mathcal{D}_t)}[\tilde{f}(z)]$, and $\boldsymbol{\Sigma}_t^2(z) = \mathrm{Var}[\tilde{f}(z)]$. We denote $\boldsymbol{\sigma}_t(\cdot) = \mathrm{diag}(\boldsymbol{\Sigma}_t(\cdot))$ and make the following assumptions regarding the considered statistical model irrespective of the taken perspective:

Assumption 3 (Calibrated model)**.** *The statistical model is calibrated w.r.t.* f*, i.e., there is a known non-decreasing sequence of confidence parameters* $\{\beta_t\}_{t \geq 0}$*, each* $\beta_t \quad \mathsf{R}_{>0}$ *and depending on* δ*, such that with probability at least* $1 - \delta$*, it holds jointly for all* t *and* $z \quad \mathcal{Z}$ *that* $|f(z) - \boldsymbol{m}_t(z)| \leq \beta_t \boldsymbol{\sigma}_t(z)$ *elementwise.*

Assumption 4. *The function* $\boldsymbol{\sigma}_t(\cdot)$ *is* $L_{\boldsymbol{\sigma}}$*-Lipschitz-continuous for all* $t \geq 1$*.*

The calibrated model assumption (or its equivalents) is standard in online model-based learning (Srinivas et al., 2010; Chowdhury & Gopalan, 2017; 2019; Curi et al., 2020; Sessa et al., 2022). It states that the agent can build high probability confidence bounds around the unknown dynamics at the beginning of every episode. As more observations become available, we expect the *epistemic* uncertainty of the model (that arises due to the lack of data and is encoded in $\boldsymbol{\sigma}(\cdot)$) to shrink, and consequently allow the representative agent to make better decisions as it becomes more confident about the true dynamics. Our theoretical results obtained in Section 4 hold for general model classes as long as Assumption 3 and Assumption 4 are satisfied. Below and in Section 4, we provide concrete conditions (about f) and models for which this assumption provably holds.

Algorithm. At the beginning of episode t, the representative agent constructs the confidence set of dynamics functions, denoted by $\mathcal{F}_{t-1}$, satisfying the elementwise confidence interval in Assumption 3 with $\boldsymbol{m}_{t-1}(\cdot)$ and $\boldsymbol{\sigma}_{t-1}(\cdot)$ estimated based on the observations up until the end of the previous episode $t-1$, i.e.,

$$\mathcal{F}_{t-1} = \left\{ \tilde{f} : |\tilde{f}(z) - \boldsymbol{m}_{t-1}(z)| \leq \beta_{t-1}\boldsymbol{\sigma}_{t-1}(z) \text{ elementwise and } \quad z \quad \mathcal{Z} \right\}.$$

Then, the agent selects the optimistic policy $\boldsymbol{\pi}_t$ which achieves the highest possible cumulative reward over the set of admissible policies, Π, and plausible system dynamics $\mathcal{F}_{t-1}$. In particular, the representative agent solves the following problem:

$$\boldsymbol{\pi}_t = \arg\max_{\boldsymbol{\pi} \quad \Pi} \max_{\tilde{f} \quad \mathcal{F}_{t-1}} \mathsf{E}\left[\sum_{h=0}^{H-1} r(\tilde{s}_{t,h}, \tilde{a}_{t,h}, \tilde{\mu}_{t,h})\right] \tag{5a}$$

$$\text{s.t.} \quad \tilde{a}_{t,h} = \pi_{t,h}(\tilde{s}_{t,h}, \tilde{\mu}_{t,h}), \tag{5b}$$

$$\tilde{s}_{t,h+1} = \tilde{f}(\tilde{s}_{t,h}, \tilde{a}_{t,h}, \tilde{\mu}_{t,h}) + \tilde{\omega}_{t,h}, \tag{5c}$$

$$\tilde{\mu}_{t,h+1} = \Phi(\tilde{\mu}_{t,h}, \pi_{t,h}, \tilde{f}), \tag{5d}$$

Algorithm 1 Model-based RL for Mean-field Control

```
Input: Calibrated dynamical model, reward function $r(s_{t,h}, a_{t,h}, \mu_{t,h})$, horizon $H$, initial state $s_{1,0} \quad \mu_0$
for $t = 1, 2, \dots$ do
    Use M$^3$–UCRL to select policy profile $\boldsymbol{\pi}_t = (\pi_{t,0}, \dots, \pi_{t,h-1})$ by using the current dynamics model and reward function, i.e., solve Eq. (5)
    for $h = 0, \dots, H - 1$ do
        $a_{t,h} = \pi_{t,h}(s_{t,h}, \mu_{t,h})$,
        $s_{t,h+1} = f(s_{t,h}, a_{t,h}, \mu_{t,h}) + \omega_{t,h}$
        $\mu_{t,h+1} = \Phi(\mu_{t,h}, \pi_{t,h}, f)$
    end for
    Update agent's statistical model with new observations $\{(s_{t,h}, a_{t,h}, \mu_{t,h}), s_{t,h+1}\}_{h=0}^{H-1}$
    Reset the system to $\mu_{t+1,0} \leftarrow \mu_0$ and $s_{t+1,0} \quad \mu_{t+1,0}$
end for
```

where the expectation in Eq. (5a) is taken w.r.t. the initial state distribution μ_0 and noise in transitions.

The policy $\boldsymbol{\pi}_t$ found by solving the previous problem is then used in the true multi-agent system (e.g., used by every driver in a ride-hailing service) for one episode. It induces the observed mean-field flow, and aims to improve the total collective reward. After the episode ends, the representative agent augments its observed data, i.e., $\boldsymbol{D}_{1:t} = \boldsymbol{D}_{1:t-1} \quad \boldsymbol{D}_t$, and improves its model estimates $\boldsymbol{m}_t(\cdot)$ and $\boldsymbol{\sigma}_t(\cdot)$. The algorithm is summarized in Algorithm 1. Since solving Eq. (5) is a challenging task even for known dynamics and the focus of this work is on the statistical complexity of learning in MFC (similarly to Kakade et al. (2020); Chowdhury & Gopalan (2019); Curi et al. (2020); Sessa et al. (2022)), we assume that we can solve Eq. (5) for a given $\boldsymbol{F}_{t-1}$ and Π. We propose a practical implementation below with details on solving Eq. (5). The algorithm implements the *upper-confidence bound* principle, since the system's true dynamics f belong to the confidence set $\boldsymbol{F}_{t-1}$ with high-probability (by Assumption 3). Consequently, the reward achieved by $\boldsymbol{\pi}_t$ under the best possible dynamics $\tilde{f}$ upper bounds the performance of $\boldsymbol{\pi}_t$ under the true dynamics f.

Practical Implementation. In general, optimizing over the set $\boldsymbol{F}_{t-1}$ is not tractable. However, a practical problem reformulation has been designed (Moldovan et al., 2015; Curi et al., 2020) for single-agent problems, which defines an auxiliary policy to select the dynamics function $\tilde{f} \quad \boldsymbol{F}_{t-1}$. We generalize this approach to the MFC setting to implement M^3–UCRL for our experiments (see Appendix D.1 for details).

M^3–UCRL can be combined with any statistical model that satisfies Assumption 3. For example, under mild assumptions on the unknown dynamics, Gaussian Processes (GP) models (Rasmussen, 2003) can be provably calibrated under some regularity assumptions on the dynamics (Srinivas et al., 2010; Abbasi-Yadkori & Szepesvaŕi, 2011). Neural Network (NN) models (Anthony & Bartlett, 2009) such as Probabilistic and Deterministic Ensembles (Lakshminarayanan et al., 2017; Chua et al., 2018) are more scalable, and even though they are not calibrated by default, their empirical recalibration is possible (Kuleshov et al., 2018). Finally, using such differentiable dynamics models and parameterizing $\boldsymbol{\pi}_t(\cdot)$ and the auxiliary policy selecting $\tilde{f}$ via separate neural networks, allows for optimizing Eq. (5) by using gradient-based approaches (see Appendix D).

Remark 2. *M^3–UCRL uses an optimistic upper-confidence bound strategy similarly to single-agent model-based upper-confidence algorithms (Chowdhury & Gopalan, 2019; Kakade et al., 2020; Curi et al., 2020; Sessa et al., 2022), however, it addresses several important differences: (i) dynamics and rewards explicitly depend on the states of all agents through the mean-field distribution; (ii) each agent's action depends on the mean-field distribution and the evolution of the mean-field trajectory is induced by the collection of actions taken by the agents; (iii) crucially, M^3–UCRL employs the principle that each agent in the system executes the same policy selected by the representative agent; Moreover, we note that single-agent MBRL algorithms are inherently unsuitable for solving the Mean Field Control problem.*

4 Theoretical Analysis

We analyze our approach using the standard notion of *cumulative regret.* It measures the difference in the cumulative expected reward between the socially optimal policy $\boldsymbol{\pi}$ (from Eq. (4)) and the individual policies selected by M^3–UCRL in every episode: $R_T = \sum_{t=1}^{T}(\mathsf{J}(\boldsymbol{\pi}\ \ ) - \mathsf{J}(\boldsymbol{\pi}_t))$. We also use r_t to denote the regret incurred during episode t, i.e., $r_t := \mathsf{J}(\boldsymbol{\pi}\ \ ) - \mathsf{J}(\boldsymbol{\pi}_t)$. We aim to show that M^3–UCRL achieves *sublinear* cumulative regret, i.e., $\lim_{T \to \infty} \frac{1}{T} R_T = 0$. This implies that as the number of episodes increases, the performance of the selected policies converges to that of the optimal policy: $\mathsf{J}(\boldsymbol{\pi}_t) \to \mathsf{J}(\boldsymbol{\pi}\ \ )$. We defer all the proofs from this section to Appendix B and only highlight the main contributions of our work.

In general, the speed of convergence of M^3–UCRL will depend on the difficulty of estimating f. For more complex functions, we expect our models to require more observations to achieve close approximation of the underlying dynamics. Consequently, we use the following model-based complexity measure to quantify this aspect of the learning problem:

$$I_T = \max_{\tilde{D}_1,\ldots,\tilde{D}_T} \sum_{t=1}^{T} \sum_{z \quad \tilde{D}_1 \ \cdots \ \tilde{D}_t} \quad \boldsymbol{\sigma}_{t-1}(z) \ {}_2^2, \tag{6}$$

where $\tilde{D}_t$ is a set of possible observations in an episode for each $t = 1, \ldots, T$, and $\boldsymbol{\sigma}_{t-1}(\cdot)$ is the confidence estimate of the selected statistical model computed based on $\tilde{D}_1 \quad \cdots \quad \tilde{D}_{t-1}$. We note that the analogous notion of model complexity has been recently used in single agent model-based RL (Chowdhury & Gopalan, 2019; Curi et al., 2020) and the multi-agent setting (Sessa et al., 2022). Intuitively, I_T measures the chosen model's total confidence in estimating the dynamics over T episodes in the worst-case, i.e., when the sets of observations $\tilde{D}_1, \ldots, \tilde{D}_T$ are the least informative about the unknown dynamics f. For statistical models that fail to learn and generalize, I_T may grow linearly with T. On the other hand, for models that learn from a "small" number of samples, we expect $\boldsymbol{\sigma}_{t-1}(\cdot)$ to shrink fast.

General Model-based Regret Bound. The following theorem bounds the cumulative regret after T episodes in terms of the model complexity measure I_T defined in Eq. (6).

Theorem 1. *Under Assumptions 1 to 4, let* $\overline{L}_{T-1} = 1 + 2(1 + L_\pi)\,[L_f + 2\beta_{T-1} L_{\boldsymbol{\sigma}}]$*, and let* $\mu_{t,h} \quad \mathcal{P}(\mathcal{S})$ *for all* $t, h > 0$*. Then for all* $T \geq 1$ *and fixed constant* $H > 0$*, with probability at least* $1 - \delta$*, the regret of M*^{3}*–UCRL is at most:*

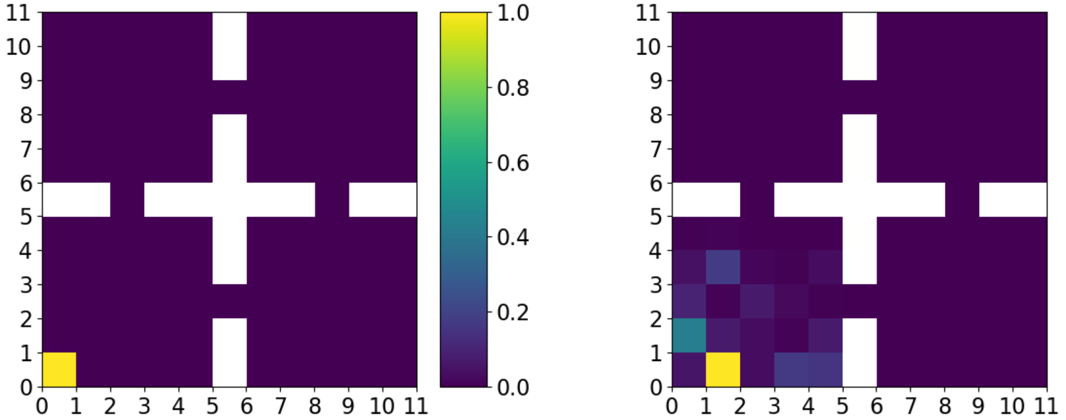
$$R_T = \mathcal{O}\Big(\beta_T L_r (1 + L_\pi) \overline{L}_{T-1}^{H-1} \sqrt{H^3 T I_T}\Big).$$

The obtained regret bound explicitly depends on constant factors that describe the environment such as the episode horizon H and the Lipschitz constants. The dependence on the size of the problem, i.e., p and q, is implicit in β_T and I_T and specific to the statistical model used in the algorithm. We provide more explicit dependencies for Gaussian Processes below and in Appendix C.1. We also note that β_T is a function of δ (see Assumption 3). When polices are selected according to Eq. (5) at every episode, our result implies that the obtained performance $\mathsf{J}(\boldsymbol{\pi}_t)$ eventually converges to the optimal one $\mathsf{J}(\boldsymbol{\pi}\ \ )$ if the joint dependence on the model-dependent quantities I_T and β_T scales as $\mathcal{O}(\sqrt{T})$.

In comparison to other model-based algorithms, both M^3–UCRL and H-UCRL (Curi et al., 2020) achieve $\mathcal{O}(\sqrt{H^3 T I_T})$ regret while H-MARL (Sessa et al., 2022) achieves $\mathcal{O}(N^{H/2}\sqrt{H^3 T I_T})$. The additional factor of $\mathcal{O}(N^{H/2})$ comes from the fact that H-MARL considers separate agents with individual action spaces while M^3–UCRL optimises for a representative agent instead. In terms of tightness of our results, we expect that the dependency on $\mathcal{O}(\sqrt{T I_T})$ in the regret bound is unavoidable. Scarlett et al. (2017) showed that this is a lower bound for the kernelized bandit problem for specific kernels and using a Gaussian Process. This result relates to the special case of MFC when $H = 1$ and using Gaussian Process statistical model to estimate the transition dynamics. The dependency on H^3, however, could be improved under more restrictive assumptions.

In our theoretical analysis, we address non-trivial challenges that arise from considering an infinite number of agents. In particular, we first translate the Mean Field Control to a non-standard MDP with the distributional state space $\mathcal{P}(\mathcal{S})$ and action space Π of all admissible policies (for details, see Appendix B.2). This enables us to use the upper-confidence bound principle with the assumption of well-calibrated models to show the following bound on the episodic regret:

$$r_t \leq 2L_r(1 + L_\pi) \sum_{h=1}^{H-1} W_1(\mu_{t,h}, \tilde{\mu}_{t,h}),$$



(a)

(b)

Figure 1: Fig. 1a shows the initial mean-field distribution $\mu_{h,0}$ for each episode h. Fig. 1b shows the mean-field distribution at the end of an episode using a policy that selects actions uniformly at random.

where $\mu_{t,h}$ and $\tilde{\mu}_{t,h}$ are the mean-field distributions when the policy $\boldsymbol{\pi}_t$ is deployed in (i) the true multi-agent system and (ii) the system with dynamics given by the transition function $\tilde{f}_t$ selected by the oracle corresponding to $\boldsymbol{\pi}_t$. The main theoretical contribution of our work is provided in Lemma 5 (see Appendix B.3) which shows the following bound on the Wasserstein-1 distance between the two mean-field distributions:

$$W_1(\tilde{\mu}_{t,h}, \mu_{t,h}) \leq 2\beta_{t-1}\overline{L}_{t-1}^{h-1} \sum_{i=1}^{h-1} \int_{\mathcal{S}} \ \boldsymbol{\sigma}_{t-1}(s, \pi_t(s, \mu_{t,i}), \mu_{t,i}) \ {}_2\, \mu_{t,i}(ds).$$

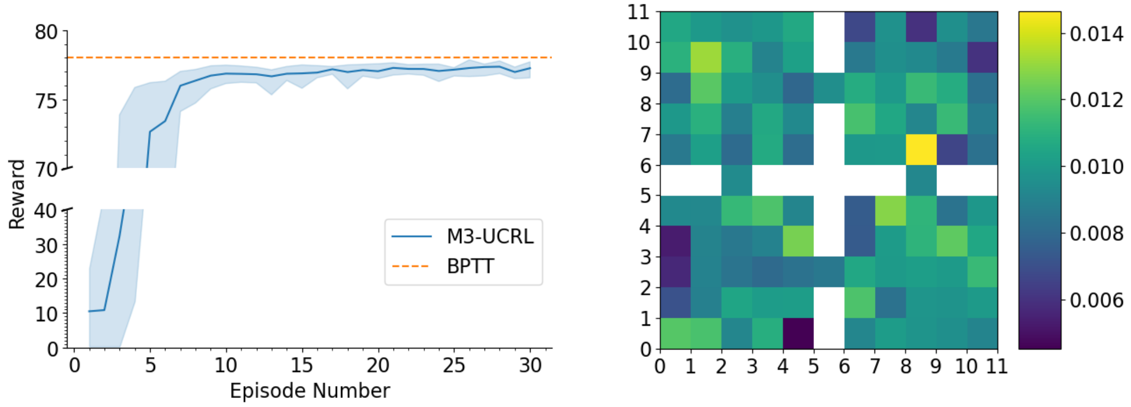
This shows that the deviation between the used $\tilde{\mu}_{t,h}$ and $\mu_{t,h}$ is bounded by the uncertainty of the statistical model integrated over $\mu_{t,h}$. As the model's uncertainty decreases, the set $\mathcal{F}_t$ shrinks towards the true dynamics f and the dynamics $\tilde{f}$ selected in $Eq.$ (5) is restricted to be close to f.

Next, we show an example of how the previously obtained regret bounds can be instantiated for particular models. To do so, we consider the case of Gaussian Process models.

Bounds for Gaussian Process (GP) Models. GP models are frequently used to model unknown dynamics in the model-based literature (see, e.g., Srinivas et al. (2010); Chowdhury & Gopalan (2019); Curi et al. (2020); Sessa et al. (2022)). They can successfully distinguish between model's and noise uncertainty, while their expressiveness follows from different kernel functions that can be used to model different types of correlation in data.

Under some regularity assumptions on f, i.e., when the true f has a bounded norm in the Reproducing Kernel Hilbert Space (RKHS) induced by the GP kernel function, these models provably satisfy Assumption 3. Moreover, we formally show in Appendix C that for such a GP statistical model, we can bound I_T by the *GP maximum mutual information (MMI)*. Similarly, we can express β_T via the same quantity, and conclude that sublinearity of our regret depends on the MMI rates. Srinivas et al. (2010) obtain sublinear (in the number of episodes T) upper bounds for this quantity in case of the most commonly-used kernels, such as the linear, squared-exponential, or Matérn kernels, while Krause & Ong (2011) prove sublinear bounds for certain composite kernels.

Finally, we note that the aforementioned works consider kernels defined on Euclidean spaces while our input space for f, namely $\mathcal{Z} = \mathcal{S} \times \mathcal{A} \times \mathcal{P}(\mathcal{S})$, includes the space of probability densities $\mathcal{P}(\mathcal{S})$. Providing bounds on the GP maximum mutual information capacity in case of kernels defined on probability spaces is an interesting direction for future work.



(a)

(b)

Figure 2: Convergence of the M^3–UCRL algorithm under system dynamics given by Eq. (7). Fig. 2a shows the rewards defined in Eq. (4) achieved by M^3–UCRL over 30 episodes. Confidence bounds show the minimum and maximum over 10 independent experiments. The BPTT line corresponds to the rewards achieved under known dynamics and considered to be optimal in the environment. Fig. 2b shows the close to uniform mean-field distribution at the end of the episode, $h = 20$, for a policy optimised by M^3–UCRL.

5 Experiments

In this section, we demonstrate the performance of the M^3–UCRL algorithm on the *exploration via entropy maximization* problem introduced by Geist et al. (2021) and used as a benchmark problem in a recent survey on Mean-Field Games (Laurière et al., 2022). Different from previous works, we formulate the problem in the continuous state and action spaces as follows. The state-space of the model is described by a 2-dimensional space in $[0, 11]^2$ which is split into 4 equal-sized rooms separated by unit-sized walls with one corridor connecting neighboring rooms (See Fig. 1a). Agents are free to move around within the area and are stopped by the walls if they try to move through them. Each episode t consists of 21 steps starting from $h = 0$ and in each time-step h the representative agent chooses its actions from the action space $\mathsf{A} = [0, 1]^2$. The dynamics of the system in Eq. (1) are of the following form

$$f(s_{t,h}, a_{t,h}, \mu_{t,h}) = s_{t,h} + a_{t,h}, \tag{7}$$

and the additive noise is Gaussian with zero mean, variance $\boldsymbol{\sigma}^2$ and independent dimensions, i.e., $\omega_{t,h}$ $N(0, \boldsymbol{\sigma}^2 I_2)$ for all t and h where I_2 is the 2×2 unit matrix. The individual reward function is defined as $r(s, a, \mu) = -\log(\mu(s))$ meaning that the representative agent aims to avoid crowded areas. On the whole population's level, the average one-step reward is given by $\mathsf{E}_{s \quad \mu}[-\log(\mu(s)] = -\int_{s \quad \mathsf{S}} \mu(s) \log(\mu(s))$, so maximizing the average reward is equivalent of maximizing the population's entropy. The population is initially condensed in the bottom-left corner (as shown in Fig. 1a) in a unit square, therefore, a policy is optimal if it disperses the population quickly to reach a uniform distribution.

We parameterize our policy with a Neural Network and use a Deep Ensemble model (Lakshminarayanan et al., 2017) for estimating the system dynamics. To represent the mean-field distribution, we discretize the state-space to unit squares and assign each cell the probability of the representative agent is inside of it. We provide further details on the implementation in Appendix D. Additionally, we ran experiments with Gaussian Process models for the swarm motion problem to showcase the flexibility of design choices in M^3–UCRL. Details and results are described in Appendix E.2.

Laurière et al. (2022) shows that exploration is not trivial in the discrete state and action space equivalent of this environment by considering the uniform random policy which fails to reach the uniform distribution within the episode. This result holds in the continuous environment as well as shown by Fig. 1b.

Results. Due to the lack of an analytical solution to the problem, we compare the rewards achieved by M^3–UCRL in the policy roll-outs (i.e., the interactions with the environment) to the rewards achieved by the policy optimised via the back-propagation through time (BPTT) algorithm with *known* dynamics. We denote this policy by $\boldsymbol{\pi}^{BPTT}$ and find it to be adequate since when it is used, the population's entropy quickly reaches the maximum achievable corresponding to a uniform distribution. We provide further analysis of this policy in Appendix D.4.

Fig. 2a shows the rewards achieved by M^3–UCRL over 30 episodes. The learning can be characterized by two phases; in the first 10 episodes M^3–UCRL learns how to navigate the whole state-space, after which it reduces exploration and focuses more on fine-grained improvements. Learning to navigate the environment is challenging due to the precise movements needed to enter the corridors, especially, from an angle. Small changes in the action can decide whether the representative agent moves through the corridor to another room or is stopped by a wall. As shown by the swift improvement in M^3–UCRL's performance, it successfully accumulates information in the first phase of the learning process in order to have sufficiently precise policies that can navigate the agent population through the corridors. In the second phase, it gradually refines its policy further to distribute the agents among the rooms quicker and achieve a much closer to uniform distribution as shown on Fig. 2b. The optimised policy after 30 episodes achieves a reward of 77.86 compared to 78.26 that of $\boldsymbol{\pi}^{BPTT}$, i.e., less than 1% difference. [2]

While M^3–UCRL converges within 30 episodes, a comparable tabular Q-learning algorithm for the Mean-Field Control problem fails to achieve similar performance in more than 10 million training episodes in the discrete problem defined in Geist et al. (2021) and Laurière et al. (2022). This is most likely due to the large state space, the limitation of deterministic actions, and challenging exploration in the environment. We provide further details of the comparison and why Q-learning requires long training for convergence in Appendix E.1.

6 Conclusion

In this work, we propose the first model-based algorithm for the Mean-Field Control problem. The M^3–UCRL algorithm runs in episodes, builds optimistic performance estimates, and uses them to efficiently explore the space during policy learning. We proved general regret bounds and showed how they can be specialized to Gaussian Process models. Furthermore, our experiments showcase the practicality of the algorithm and how it can be easily combined with deep Neural Networks. We believe that our approach provides an important step towards making model-based MARL algorithms more principled, scalable, and sample-efficient. We see the following topics important for the further progression of the model-based mean-field learning for both the mean-field control and mean-field game problems: 1) Planning with known dynamics to solve Eq. (5) similarly to Gu et al. (2020); Motte & Pham (2019), 2) Eliminating the need of recalibration to have Neural Networks satisfying Assumption 3, 3) Maximum mutual information bounds for kernels defined on probability spaces.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation program grant agreement No 815943 and ETH Zurich Postdoctoral Fellowship 19-2 FEL-47. This publication was made possible by an ETH AI Center doctoral fellowship to Barna Pasztor.

References

Yasin Abbasi-Yadkori and Csaba Szepesvaŕi. Regret bounds for the adaptive control of linear quadratic systems. *Journal of Machine Learning Research*, 19:1–26, 2011.

Mridul Agarwal, Vaneet Aggarwal, Arnob Ghosh, and Nilay Tiwari. Reinforcement learning for mean field game. *arXiv preprint arXiv:1905.13357*, 2019.

Noha Almulla, Rita Ferreira, and Diogo Gomes. Two numerical approaches to stationary mean-field games. *Dynamic Games and Applications*, 7(4):657–682, 2017.

[2] Animations of episodes with the BPTT policy (known_dynamics_animation.mp4) and the M^3–UCRL policy (m3_ucrl_animation_episode_t.mp4) are included in the supplementary material. Episode 1 to 10 illustrate the quick learning during the exploration phase while episode 26 is the best-performing one also used for Fig. 2b.

Andrea Angiuli, Jean-Pierre Fouque, and Mathieu Lauriere. Reinforcement Learning for Mean Field Games, with Applications to Economics. *arXiv preprint arXiv:2106.13755*, 2021.

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*, pp. 1–55, 2022.

Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 2009.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In *Advances in Neural Information Processing Systems*, 2009.

Nicole Bäuerle. Mean Field Markov Decision Processes. *arXiv preprint arXiv:2106.08755*, 2021.

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

Ronen I Brafman and Moshe Tennenholtz. A near-optimal polynomial time algorithm for learning in certain classes of stochastic games. *Artificial Intelligence*, 121(1-2):31–47, 2000.

Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, 2002.

Lucian Buşoniu, Robert Babuška, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, 38(2): 156–172, 2008.

Pierre Cardaliaguet and Saeed Hadikhanloo. Learning in mean field games: the fictitious play. *ESAIM: Control, Optimisation and Calculus of Variations*, 23(2):569–591, 2017.

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

René Carmona, François Delarue, et al. *Probabilistic Theory of Mean Field Games with Applications I-II*. Springer, 2018.

René Carmona, Mathieu Laurière, and Zongjun Tan. Linear-quadratic mean-field reinforcement learning: convergence of policy gradient methods. *arXiv preprint arXiv:1910.04295*, 2019a.

René Carmona, Mathieu Laurière, and Zongjun Tan. Model-free mean-field reinforcement learning: mean-field mdp and mean-field q-learning. *arXiv preprint arXiv:1910.12802*, 2019b.

René Carmona, Kenza Hamidouche, Mathieu LauriÉre, and Zongjun Tan. Linear-quadratic zero-sum mean-field type games: Optimality conditions and policy optimization. *Journal of Dynamics and Games. 2021, Volume 8, Pages 403-443*, 8(4):403, 2021. ISSN 21646074.

Cristiano Castelfranchi. The theory of social functions: challenges for computational social science and multi-agent learning. *Cognitive Systems Research*, pp. 5–38, 2001.

Minshuo Chen, Yan Li, Ethan Wang, Zhuoran Yang, Zhaoran Wang, and Tuo Zhao. Pessimism Meets Invariance: Provably Efficient Offline Mean-Field Multi-Agent RL. *Advances in Neural Information Processing Systems*, 34:17913–17926, 12 2021.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, 2017.

Sayak Ray Chowdhury and Aditya Gopalan. Online learning in kernelized Markov decision processes. In *Proceedings of Machine Learning Research*, volume 89, pp. 3197–3205, 2019.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems*, pp. 4754–4765, 2018.

Sebastian Curi, Felix Berkenkamp, and Andreas Krause. Efficient model-based reinforcement learning through optimistic policy search and planning. In *Advances in Neural Information Processing Systems*, 2020.

Omar Darwiche Domingues, Pierre Ménard, Matteo Pirotta, Emilie Kaufmann, and Michal Valko. Regret bounds for kernel-based reinforcement learning. *arXiv preprint arXiv:2004.05599*, 2020.

Audrey Durand, Odalric-Ambrym Maillard, and Joelle Pineau. Streaming kernel regression with provably adaptive mean, variance, and regularization. *The Journal of Machine Learning Research*, 2018.

Yonathan Efroni, Nadav Merlis, Mohammad Ghavamzadeh, and Shie Mannor. Tight regret bounds for model-based reinforcement learning with greedy policies. In *Advances in Neural Information Processing Systems*, pp. 12224–12234, 2019.

S. El-Tantawy, B. Abdulhai, and H. Abdelgawad. Multiagent reinforcement learning for integrated network of adaptive traffic signal controllers (marlin-atsc): Methodology and large-scale application on downtown toronto. *IEEE Transactions on Intelligent Transportation Systems*, 14(3):1140–1150, 2013.

Romuald Elie, Julien Pérolat, Mathieu Laurière, Matthieu Geist, and Olivier Pietquin. On the convergence of model free learning in mean field games. *AAAI Conference on Artificial Intelligence*, 34(05):7143–7150, 2020.

Zuyue Fu, Zhuoran Yang, Yongxin Chen, and Zhaoran Wang. Actor-critic provably finds nash equilibria of linear-quadratic mean-field games. In *International Conference on Learning Representations*. OpenReview.net, 2020.

Nicolas Gast, Bruno Gaujal, and Jean-Yves Le Boudec. Mean field for Markov decision processes: from discrete to continuous optimization. *IEEE Transactions on Automatic Control*, pp. 2266–2280, 2012.

Matthieu Geist, Julien Pérolat, Mathieu Laurière, Romuald Elie, Sarah Perrin, Olivier Bachem, Rémi Munos, and Olivier Pietquin. Concave Utility Reinforcement Learning: the Mean-Field Game Viewpoint. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, 1:489–497, 6 2021. ISSN 15582914.

Diogo A Gomes et al. 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. Dynamic programming principles for mean-field controls with learning. *arXiv preprint arXiv:1911.07314*, 2020.

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.

Olivier Guéant, Jean-Michel Lasry, and Pierre-Louis Lions. Mean field games and applications. In *Paris-Princeton lectures on mathematical finance 2010*, pp. 205–266. Springer, 2011.

Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. Learning mean-field games. In *Advances in Neural Information Processing Systems*, volume 32, pp. 4966–4976, 2019.

Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. A general framework for learning mean-field games. *arXiv preprint arXiv:2003.06069*, 2020.

Pablo Hernandez-Leal, Michael Kaisers, Tim Baarslag, and Enrique Munoz de Cote. A survey of learning in multiagent environments: Dealing with non-stationarity. *arXiv preprint arXiv:1707.09183*, 2017.

Pablo Hernandez-Leal, Bilal Kartal, and Matthew E. Taylor. A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, 33(6):750–797, 2019.

Minyi Huang, Roland P Malhamé, Peter E Caines, et al. Large population stochastic dynamic games: closed-loop mckean-vlasov systems and the nash certainty equivalence principle. *Communications in Information & Systems*, 6(3):221–252, 2006.

Minyi Huang, Peter E Caines, and Roland P Malhamé. Large-population cost-coupled lqg problems with nonuniform agents: individual-mass behavior and decentralized ϵ-equilibria. *IEEE Transactions on Automatic Control*, 52(9):1560–1571, 2007.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 2010.

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

Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. *arXiv preprint arXiv:2006.12466*, 2020.

Leonid Kantorovich and Gennady S. Rubinstein. On a space of totally additive functions. *Vestnik Leningrad. Univ*, 13:52–59, 1958.

Andreas Krause and Cheng Soon Ong. Contextual Gaussian process bandit optimization. *Advances in Neural Information Processing Systems*, 2011.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning*, pp. 2796–2804, 2018.

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

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

Jean-Michel Lasry and Pierre-Louis Lions. Jeux à champ moyen. i–le cas stationnaire. *Comptes Rendus Mathématique*, 343(9):619–625, 2006a.

Jean-Michel Lasry and Pierre-Louis Lions. Jeux à champ moyen. ii–horizon fini et contrôle optimal. *Comptes Rendus Mathématique*, 343(10):679–684, 2006b.

Mathieu Laurière, Sarah Perrin, Matthieu Geist, and Olivier Pietquin. Learning Mean Field Games: A Survey. *arXiv preprint arXiv:2205.12944v2*, 2022.

J. W. Lee, J. Park, J. O, J. Lee, and E. Hong. A multiagent approach to q-learning for daily stock trading. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, pp. 864–877, 2007.

Jae Won Lee, Byoung-Tak Zhang, et al. Stock trading system using reinforcement learning with cooperative agents. In *International Conference on Machine Learning*, pp. 451–458, 2002.

Charles-Albert Lehalle and Charafeddine Mouzouni. A mean field game of portfolio trading and its consequences on perceived correlations. *arXiv preprint arXiv:1902.09606*, 2019.

Joel Z. Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *Conference on Autonomous Agents and MultiAgent Systems*, pp. 464–473, 2017.

Yan Li, Lingxiao Wang, Jiachen Yang, Ethan Wang, Zhaoran Wang, Tuo Zhao, and Hongyuan Zha. Permutation Invariant Policy Optimization for Mean-Field Multi-Agent Reinforcement Learning: A Principled Approach. *arXiv preprint arXiv:2105.08268*, 2021.

Teodor Mihai Moldovan, Sergey Levine, Michael I Jordan, and Pieter Abbeel. Optimism-driven exploration for nonlinear systems. In *IEEE International Conference on Robotics and Automation*, 2015.

Médéric Motte and Huyên Pham. Mean-field Markov decision processes with common noise and open-loop controls. *arXiv preprint arXiv:1912.07883*, 2019.

Thanh Thi Nguyen, Ngoc Duy Nguyen, and Saeid Nahavandi. Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications. *IEEE transactions on cybernetics*, 50(9): 3826–3839, 2020.

Afshin OroojlooyJadid and Davood Hajinezhad. A review of cooperative multi-agent deep reinforcement learning. *arXiv preprint arXiv:1908.03963*, 2019.

Jakub Pachocki, Greg Brockman, Jonathan Raiman, Susan Zhang, Henrique Pondé, Jie Tang, Filip Wolski, Christy Dennison, Rafal Jozefowicz, Przemyslaw Debiak, et al. Openai five. URL https://blog. openai. com/openai-five, 2018.

Victor M. Panaretos and Yoav Zemel. Statistical Aspects of Wasserstein Distances. *Annual Review of Statistics and Its Application*, 6(1):405–431, 2019. ISSN 2326-8298.

Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, 2003.

Jonathan Scarlett, Ilija Bogunovic, and Volkan Cevher. Lower bounds on regret for noisy Gaussian process bandit optimization. In Satyen Kale and Ohad Shamir (eds.), *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 1723–1742. PMLR, 07–10 Jul 2017.

Pier Giuseppe Sessa, Maryam Kamgarpour, and Andreas Krause. Efficient Model-based Multi-agent Reinforcement Learning via Optimistic Equilibrium Computation, 2022. ISSN 2640-3498.

Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning*, pp. 1015–1022, 2010.

Ingo Steinwart and Christmann Andreas. Support vector machines. In *Climate Change 2013 - The Physical Science Basis*. Springer, 2008. ISBN 978-0-387-77241-7.

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

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

Lingxiao Wang, Zhuoran Yang, and Zhaoran Wang. Breaking the curse of many agents: Provable mean embedding q-iteration for mean-field reinforcement learning. In *International Conference on Machine Learning*, pp. 10092–10103, 2020.

Weichen Wang, Jiequn Han, Zhuoran Yang, and Zhaoran Wang. Global Convergence of Policy Gradient for Linear-Quadratic Mean-Field Control/Game in Continuous Time, 2021. ISSN 2640-3498.

Pierre Weiss. L'hypothèse du champ moléculaire et la propriété ferromagnétique. *J. Phys. Theor. Appl.*, 6 (1):661–690, 1907.

Jiachen Yang, Xiaojing Ye, Rakshit Trivedi, Huan Xu, and Hongyuan Zha. Learning deep mean field games for modeling large population behavior. *International Conference on Learning Representations*, pp. 1–15, 2017.

Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean field multi-agent reinforcement learning. In *International Conference on Machine Learning*, pp. 5571–5580, 2018.

Huibing Yin, Prashant G Mehta, Sean P Meyn, and Uday V Shanbhag. Learning in mean-field oscillator games. In *49th IEEE Conference on Decision and Control (CDC)*, pp. 3125–3132, 2010.

Huibing Yin, Prashant G Mehta, Sean P Meyn, and Uday V Shanbhag. Learning in mean-field games. *IEEE Transactions on Automatic Control*, 59(3):629–644, 2013.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *arXiv preprint arXiv:1911.10635*, 2019.

A Preliminaries on the Wasserstein Distance

In this section, we provide a brief overview on the Wasserstein distance and its properties used in the proof of our general regret bound in Theorem 1.

Definition 1 (Wasserstein Distance)**.** *Let μ and ν be two probability distributions on R^d, then for $k \in \{1, 2, ...\}$ the Wasserstein distance of order k is defined by*

$$W_k(\mu, \nu) = \left[\inf_{\gamma} \int_{\Omega \times \Omega} \|x - y\|^k \, d\gamma(x, y)\right]^{1/k}$$

where the infimum is taken over all joint distributions γ with marginals μ, ν. For $k = 1$, a standard dual formulation derived by Kantorovich & Rubinstein (1958) is given by:

$$W_1(\mu, \nu) = \sup_{f \in \mathsf{F}} \left\{ \int f(x) d(\mu - \nu)(x) \right\} \tag{8}$$

where $\mathsf{F} = \{f : \mathsf{R}^d \to \mathsf{R}$ such that $|f(x) - f(y)| \leq \|x - y\| \quad \forall x, y \in \mathsf{R}^d\}$. Both formulations above generalises to laws defined on more general space, e.g., complete and separable metric spaces and infinite-dimensional function spaces such as $L^2[0, 1]$.

The Wasserstein distance is usually interpreted as the minimum transportation distance between the distributions μ and ν. For random variables $X \sim \mu$ and $Y \sim \nu$, we use the notations $W_k(X, Y)$ and $W_k(\mu, \nu)$ interchangeably. Assuming that $W_k(X, Y)$ is finite, i.e., $\mathsf{E}[\|X\|^k] + \mathsf{E}[\|Y\|^k] < \infty$, the distance obeys the following properties Panaretos & Zemel (2019):

- For $m \leq n$, by Jensen's Inequality, we have:

$$W_m(X, Y) \leq W_n(X, Y) \tag{9}$$

- For $a \in \mathsf{R}$, it holds

$$W_k(aX, aY) = |a| W_k(X, Y) \tag{10}$$

- For $x \in \mathsf{R}^p$, it holds

$$W_k(X + x, Y + x) = W_k(X, Y) \tag{11}$$

- For $x \in \mathsf{R}^p$, we have that

$$W_2^2(X + x, Y) = \|x + \mathsf{E}[X] - \mathsf{E}[Y]\|^2 + W_2^2(X, Y) \tag{12}$$

Corollary 1. *Let X, Y be independent identically distributed random variables on R^p with $\mathsf{E}[\|X\|^k] < \infty$ and $x, y \in \mathsf{R}^p$. Then,*

$$W_1(X + x, Y + y) \leq \|x - y\|_2$$

Proof.

$$\begin{aligned} W_2^2(X + x, Y + y) &= W_2^2(X + x - y, Y) && \text{By } Eq. \text{ (11)} \\ &= \|x - y + \mathsf{E}[X] - \mathsf{E}[Y]\|_2^2 + W_2^2(X, Y) && \text{By } Eq. \text{ (12)} \\ &= \|x - y\|_2^2 && X \text{ and } Y \text{ are i.i.d.} \end{aligned}$$

Therefore,

$$\begin{aligned} W_1(X + x, Y + y) &\leq W_2(X + x, Y + y) && \text{By } Eq. \text{ (9)} \\ &\leq \|x - y\|_2 \end{aligned}$$

□

B Regret Bound Proof

In the following subsections, we present the theoretical analysis that leads to the results of Theorem 1.

B.1 Measuring performance under arbitrary transition function

Before we start proving the main theorem, we introduce the following notation of $\tilde{\mathsf{J}}(\boldsymbol{\pi}, \tilde{f})$ to measure the performance of a policy $\boldsymbol{\pi}$ under the transition function $\tilde{f}$:

$$
\begin{aligned}
\tilde{\mathsf{J}}(\boldsymbol{\pi}, \tilde{f}) &:= \mathsf{E}\left[\sum_{h=0}^{H-1} r(\tilde{s}_h, \tilde{a}_h, \tilde{\mu}_h)\right] && \text{(13a)}\\
\text{s.t.} \quad \tilde{a}_h &= \pi_h(\tilde{s}_h, \tilde{\mu}_h) && \text{(13b)}\\
\tilde{s}_{h+1} &= \tilde{f}(\tilde{s}_h, \tilde{a}_h, \tilde{\mu}_h) + \tilde{\omega}_h && \text{(13c)}\\
\tilde{\mu}_{h+1} &= \Phi(\tilde{\mu}_h, \pi_h, \tilde{f}), && \text{(13d)}
\end{aligned}
$$

where the expectation in Eq. (13a) is taken with respect to initial state distribution μ_0 and noise in transitions. We note that we can rewrite the optimization problems in Eq. (4) and Eq. (5) such that $\boldsymbol{\pi} \quad \arg\max_{\boldsymbol{\pi} \quad \Pi} \mathsf{J}(\boldsymbol{\pi}) = \arg\max_{\boldsymbol{\pi} \quad \Pi} \tilde{\mathsf{J}}(\boldsymbol{\pi}, f)$ and $\boldsymbol{\pi}_t = \arg\max_{\boldsymbol{\pi} \quad \Pi} \max_{\tilde{f} \quad \mathsf{F}_{t-1}} \tilde{\mathsf{J}}(\boldsymbol{\pi}, \tilde{f})$, respectively.

B.2 Measure Formulation

First, we provide the proof of Lemma 1. Then, we reformulate the optimization problems in Eq. (4) and Eq. (5a) as a Markov Decision Process on a probability measure space with transition dynamics defined in Lemma 1. The following proof follows Lemma 2.2 in Gu et al. (2020). We are stating it here for the sake of completeness.

Proof of Lemma 1. Fix $\boldsymbol{\pi}$ and let φ be a bounded measurable function on S. Recall that s_h denotes random variable of the representative agent's state at time t and, by the definition of μ_h, $s_h \quad \mu_h$.

First note that

$$
\mathsf{E}_{s_{h+1}}[\varphi(s_{h+1})] = \int_{s \quad \mathsf{S}} \mu_{h+1}(ds \ )\varphi(s \ ).
$$

On the other hand, by the law of iterated conditional expectation,

$$
\begin{aligned}
\mathsf{E}_{s_{h+1}}[\varphi(s_{h+1})] &= \mathsf{E}_{s_0,\ldots,s_h}\Big[\mathsf{E}_{s_{h+1}}[\varphi(s_{h+1}) | s_0, \ldots, s_h]\Big] \\
&= \mathsf{E}_{s_0,\ldots,s_h}\left[\int_{s \quad \mathsf{S}} \varphi(s \ )\mathsf{P}[s_{h+1} \quad ds \ | s_0, \ldots, s_h]\right] \\
&= \mathsf{E}_{s_0,\ldots,s_h}\left[\int_{s \quad \mathsf{S}} \varphi(s \ )\mathsf{P}[f(s_h, \pi_h(s_h, \mu_h), \mu_h) + \omega_h \quad ds \ ]\right] \\
&= \int_{s \quad \mathsf{S}} \varphi(s \ )\mathsf{E}_{s_0,\ldots,s_h}\Big[\mathsf{P}[f(s_h, \pi_h(s_h, \mu_h), \mu_h) + \omega_h \quad ds \ ]\Big] \\
&= \int_{s \quad \mathsf{S}} \varphi(s \ ) \int_{s \quad \mathsf{S}} \mu_h(ds)\mathsf{P}[f(s, \pi_h(s, \mu_h), \mu_h) + \omega_h \quad ds \ ].
\end{aligned}
$$

In the third equality, we used the dynamics of the system Eq. (1) which states that s_{h+1} is determined by s_h, a_h, μ_h as $s_{h+1} \quad f(s_h, a_h, \mu_h) + \omega_h$. The last equality follows from $\mu_h(ds) = \mathsf{P}[s_h \quad ds]$. Therefore,

$$
\int_{s \quad \mathsf{S}} \mu_{h+1}(ds \ )\varphi(s \ ) = \int_{s \quad \mathsf{S}} \varphi(s \ ) \int_{s \quad \mathsf{S}} \mu_h(ds)\mathsf{P}[f(s, \pi_h(s, \mu_h), \mu_h) + \omega_h \quad ds \ ],
$$

which implies the desired results

$$\mu_{h+1}(ds\ ) = \int_{s \quad \mathsf{S}} \mu_h(ds)\mathsf{P}[f(s, \pi_h(s, \mu_h), \mu_h) + \omega_h \quad ds\ ].$$

□

Remark 2. *The relationship described in Lemma 1 is the discrete-time equivalent of the Fokker-Planck (Kolmogorov forward) equation for the McKean-Vlasov Process. While we provided a proof for the dynamics in Eq.* (1)*, it holds for any other functions $\tilde{f}_t \quad \mathsf{F}_t$ for $t = 1, 2, \ldots$.*

Now, we consider the objective under the true dynamics, Eq. (4). The following lemma is based on Lemma 2.1. in Gu et al. (2020).

Lemma 2. *For any policy $\boldsymbol{\pi} = (\pi_0, \ldots, \pi_{H-1})$,*

$$\mathsf{J}(\boldsymbol{\pi}) = \sum_{h=0}^{H-1} \hat{r}(\mu_h, \pi_h),$$

where $\hat{r}$ is the integrated reward function defined as

$$\hat{r}(\mu, \pi) = \int_{\mathsf{S}} \mu(ds) r(s, \pi(s, \mu), \mu), \tag{14}$$

and μ_h is the mean-field distribution induced by the policy $\boldsymbol{\pi}$ following the dynamics defined in Eq. (2)*.*

Proof.

$$\begin{aligned}
\mathsf{J}(\boldsymbol{\pi}) &= \mathsf{E}_{s_0 \quad \mu_0, \omega_0, \ldots, \omega_{H-1}} \left[\sum_{h=0}^{H-1} r(s_h, \pi_h(s_h, \mu_h), \mu_h) \right] \\
&= \sum_{h=0}^{H-1} \left[\mathsf{E}_{s_0 \quad \mu_0, \omega_0, \ldots, \omega_{h-1}} r(s_h, \pi_h(s_h, \mu_h), \mu_h) \right] \\
&= \sum_{h=0}^{H-1} \int_{\mathsf{S}} \mathsf{P}[s_h \quad ds] r(s, \pi_h(s, \mu_h), \mu_h) \\
&= \sum_{h=0}^{H-1} \int_{\mathsf{S}} \mu_h(ds) r(s, \pi_h(s, \mu_h), \mu_h) \\
&= \sum_{h=0}^{H-1} \hat{r}(\mu_h, \pi_h).
\end{aligned}$$

The expectation in the first line is taken over the random variables $s_0, \omega_0, \ldots, \omega_{H-1}$ that induce the state trajectory $s_0, s_1, \ldots, s_H$. In particular, for a fixed policy the continuous random variable s_h is a function of $s_0, \omega_0, \ldots, \omega_{h-1}$. In the second line, we use the linearity of expectation and s_h's independence of $\omega_h, \ldots, \omega_{H-1}$. The third line follows from the definition of expectation.

Note that the fixed and known initial distribution μ_0, the true transition function f, and the fixed policy $\boldsymbol{\pi}$ define the mean-field distributions μ_h, for all $0 \leq h \leq H$ via Eq. (2). The fourth equality follows from Lemma 1, i.e., $\mu_h(ds) = \mathsf{P}[s_h \quad ds]$. □

Corollary 2. *The arguments in Lemma 2 apply in the case when the transition function of the system is unknown but approximated via $\tilde{f}$. We can rewrite $\tilde{\mathsf{J}}(\boldsymbol{\pi}, \tilde{f})$ defined in Eq.* (13a) *as*

$$\tilde{\mathsf{J}}(\boldsymbol{\pi}, \tilde{f}) = \sum_{h=0}^{H-1} \hat{r}(\tilde{\mu}_{t,h}, \pi_h).$$

where $\tilde{\mu}_{t,h}$ is the mean-field distribution induced by the policy $\boldsymbol{\pi}$ under the estimated transition function $\tilde{f}$.

Corollary 3. *By using the objective formulation from Lemma 2, we can restate the optimization problem Eq.* (4) *as*

$$\max_{\boldsymbol{\pi}} \mathsf{J}(\boldsymbol{\pi}) = \sum_{h=0}^{H-1} \hat{r}(\mu_h, \pi_h) \tag{15}$$

$$s.t.\ \ \mu_{h+1} = \Phi(\mu_h, \pi_h, f), \tag{16}$$

where Φ *is defined in Eq.* (2)*. The same holds for the optimization under any other transition function* $\tilde{f}$ *in Eq.* (5)*.*

Remark 3. *Corollary 3 turns the optimization problem Eq.* (4) *into a Markov Decision Process where the state space is the space of probability measures* $\mathcal{P}(\mathcal{S})$ *and the action space is the space of functions* $\Pi = \{\pi : \mathcal{S} \times \mathcal{P}(\mathcal{S}) \to \mathcal{A}\}$*. This reformulation comes with a certain trade-off:* $\mathcal{P}(\mathcal{S})$ *and* Π *are more complex than* $\mathcal{S} \subseteq \mathbb{R}^q$ *and* $\mathcal{A} \subseteq \mathbb{R}^p$ *but, in contrast to Eq.* (4)*, the transition function in Eq.* (16) *is deterministic.*

B.3 Proof of Theorem 1

As described in Section 3, at the beginning of episode t, the representative agent optimizes over both the set of admissible policies, Π, and plausible transition functions, $\mathcal{F}_{t-1}$, satisfying Assumptions 1 to 3. In the following, we denote the transition function corresponding to the optimal policy by $\tilde{f}_{t-1} \in \mathcal{F}_{t-1}$, i.e.,

$$\boldsymbol{\pi}_t, \tilde{f}_{t-1} = \arg\max_{\boldsymbol{\pi}, \tilde{f}} \tilde{\mathsf{J}}(\boldsymbol{\pi}, \tilde{f}).$$

Lemma 3. *Conditioning on the event in Assumption 3 holding true in episode* t*, the following holds for episodic regret:*

$$r_t = \mathsf{J}(\boldsymbol{\pi}^*) - \mathsf{J}(\boldsymbol{\pi}_t) \leq \tilde{\mathsf{J}}(\boldsymbol{\pi}_t, \tilde{f}_{t-1}) - \mathsf{J}(\boldsymbol{\pi}_t). \tag{17}$$

Proof. Recall that $\boldsymbol{\pi}^*$ denotes the socially optimal policy from Eq. (4) and let $\tilde{f}_{t-1} = \arg\max_{\tilde{f} \in \mathcal{F}_{t-1}} \tilde{\mathsf{J}}(\boldsymbol{\pi}^*, \tilde{f})$. The well-calibrated property from Assumption 3 implies that the true transition function f is in the set of functions $\mathcal{F}_{t-1}$ meaning that the true dynamics of the system f has been considered in Eq. (5) at the beginning of episode t. Then, since we provide additional flexibility to the optimization, we have

$$\tilde{\mathsf{J}}(\boldsymbol{\pi}^*, \tilde{f}_{t-1}) \geq \mathsf{J}(\boldsymbol{\pi}^*)$$

Using this inequality we can bound the episodic regret as follows:

$$\begin{aligned} r_t = \mathsf{J}(\boldsymbol{\pi}^*) - \mathsf{J}(\boldsymbol{\pi}_t) &\leq \tilde{\mathsf{J}}(\boldsymbol{\pi}^*, \tilde{f}_{t-1}) - \mathsf{J}(\boldsymbol{\pi}_t) \\ &\leq \tilde{\mathsf{J}}(\boldsymbol{\pi}_t, \tilde{f}_{t-1}) - \mathsf{J}(\boldsymbol{\pi}_t), \end{aligned}$$

where the last inequality comes from the fact that M^3–UCRL selects the policy $\boldsymbol{\pi}_t$ according to $(\boldsymbol{\pi}_t, \tilde{f}_{t-1}) = \arg\max_{\boldsymbol{\pi}, \tilde{f}} \tilde{\mathsf{J}}(\boldsymbol{\pi}, \tilde{f})$. □

Lemma 4. *Under Assumptions 2 to 3 and conditioning on that the event defined in Assumption 3 holds true, it follows for all* $t \geq 1$*:*

$$r_t \leq \tilde{\mathsf{J}}(\boldsymbol{\pi}_t, \tilde{f}_{t-1}) - \mathsf{J}(\boldsymbol{\pi}_t) \leq 2L_r(1 + L_\pi) \sum_{h=1}^{H-1} W_1(\tilde{\mu}_{t,h}, \mu_{t,h}), \tag{18}$$

where $\mu_{t,h}$ *is the mean-field distribution observed in the true unknown environment with dynamics* f *as in Eq.* (3) *when agents follow* $\boldsymbol{\pi}_t$ *and* $\tilde{\mu}_{t,h}$ *is the mean-field distribution under the approximated dynamics* $\tilde{f}_{t-1}$ *as in Eq.* (13a)*.*

Proof. By Lemma 3, we have that

$$r_t \leq \tilde{\mathsf{J}}(\boldsymbol{\pi}_t, \tilde{f}_{t-1}) - \mathsf{J}(\boldsymbol{\pi}_t).$$

By using Lemma 2, Corollary 2 and the triangle inequality we get

$$
\begin{aligned}
|\tilde{\mathsf{J}}(\boldsymbol{\pi}_t, \tilde{f}_{t-1}) - \mathsf{J}(\boldsymbol{\pi}_t)| &= \left| \sum_{h=0}^{H-1} \hat{r}(\tilde{\mu}_{t,h}, \pi_{t,h}) - \sum_{h=0}^{H-1} \hat{r}(\mu_{t,h}, \pi_{t,h}) \right| \\
&\leq \sum_{h=0}^{H-1} |\hat{r}(\tilde{\mu}_{t,h}, \pi_{t,h}) - \hat{r}(\mu_{t,h}, \pi_{t,h})|.
\end{aligned}
$$

Consider a fixed $h \quad \{0, ..., H-1\}$, then

$$
\begin{aligned}
&|\hat{r}(\tilde{\mu}_{t,h}, \pi_{t,h}) - \hat{r}(\mu_{t,h}, \pi_{t,h})| \\
&= \left| \int_{\mathcal{S}} r(s, \pi_{t,h}(s, \tilde{\mu}_{t,h}), \tilde{\mu}_{t,h}) \tilde{\mu}_{t,h}(ds) - \int_{\mathcal{S}} r(s, \pi_{t,h}(s, \mu_{t,h}), \mu_{t,h}) \mu_{t,h}(ds) \right| && \text{From } Eq. \text{ (14)} \\
&= \left| \int_{\mathcal{S}} r(s, \pi_{t,h}(s, \tilde{\mu}_{t,h}), \tilde{\mu}_{t,h}) \tilde{\mu}_{t,h}(ds) - \int_{\mathcal{S}} r(s, \pi_{t,h}(s, \mu_{t,h}), \mu_{t,h}) \tilde{\mu}_{t,h}(ds) \right. \\
&\qquad \left. + \int_{\mathcal{S}} r(s, \pi_{t,h}(s, \mu_{t,h}), \mu_{t,h}) \tilde{\mu}_{t,h}(ds) - \int_{\mathcal{S}} r(s, \pi_{t,h}(s, \mu_{t,h}), \mu_{t,h}) \mu_{t,h}(ds) \right| \\
&\leq \int_{\mathcal{S}} |r(s, \pi_{t,h}(s, \tilde{\mu}_{t,h}), \tilde{\mu}_{t,h}) - r(s, \pi_{t,h}(s, \mu_{t,h}), \mu_{t,h}))| \tilde{\mu}_{t,h}(ds) && (19) \\
&\qquad + \left| \int_{\mathcal{S}} r(s, \pi_{t,h}(s, \mu_{t,h}), \mu_{t,h}) (\tilde{\mu}_{t,h}(ds) - \mu_{t,h}(ds)) \right|, && (20)
\end{aligned}
$$

where the last inequality follows from the triangle inequality and the inequality for the absolute value of definite integrals.

We start by considering the integrand of the term in Eq. (19)

$$
\begin{aligned}
&|r(s, \pi_{t,h}(s, \tilde{\mu}_{t,h}), \tilde{\mu}_{t,h}) - r(s, \pi_{t,h}(s, \mu_{t,h}), \mu_{t,h})| \\
&\leq L_r(\ s - s \ _2 + \ \pi_{t,h}(s, \tilde{\mu}_{t,h}) - \pi_{t,h}(s, \mu_{t,h}) \ _2 + W_1(\tilde{\mu}_{t,h}, \mu_{t,h})) \\
&\leq L_r(L_\pi(\ s - s \ _2 + W_1(\tilde{\mu}_{t,h}, \mu_{t,h})) + W_1(\tilde{\mu}_{t,h}, \mu_{t,h})) \\
&\leq L_r(1 + L_\pi) W_1(\tilde{\mu}_{t,h}, \mu_{t,h}), && (21)
\end{aligned}
$$

where the inequalities follow from Assumption 2.

Next, we consider the function $g(s) = \frac{1}{L_r(1+L_\pi)} r(s, \pi_{t,h}(s, \mu_{t,h}), \mu_{t,h})$

$$
\begin{aligned}
|g(x) - g(y)| &= \frac{1}{L_r(1 + L_\pi)} |r(x, \pi_{t,h}(x, \mu_{t,h}), \mu_{t,h}) - r(y, \pi_{t,h}(y, \mu_{t,h}), \mu_{t,h})| \\
&\leq \frac{L_r}{L_r(1 + L_\pi)} (\ x - y \ _2 + \ \pi_{t,h}(x, \mu_{t,h}) - \pi_{t,h}(y, \mu_{t,h}) \ _2 + \underbrace{W_1(\mu_{t,h}, \mu_{t,h})}_{=0}) \\
&\leq \frac{L_r}{L_r(1 + L_\pi)} (\ x - y \ _2 + L_\pi(\ x - y \ _2 + \underbrace{W_1(\mu_{t,h}, \mu_{t,h})}_{=0})) \\
&= \ x - y \ _2,
\end{aligned}
$$

where Assumption 2 has been applied in the second and third lines. Therefore, g is a Lipschitz-1 function.

We bound the second term in Eq. (20) by using the previously defined function g

$$\begin{aligned}
&\int_{\mathsf{S}} r(s, \pi_{t,h}(s, \mu_{t,h}), \mu_{t,h})(\tilde{\mu}_{t,h}(ds) - \mu_{t,h}(ds)) \\
&= L_r(1 + L_\pi) \int_{\mathsf{S}} \underbrace{\frac{1}{L_r(1 + L_\pi)} r(s, \pi_{t,h}(s, \mu_{t,h}), \mu_{t,h})}_{=g(s)} (\tilde{\mu}_{t,h}(ds) - \mu_{t,h}(ds)) \\
&\leq L_r(1 + L_\pi) W_1(\tilde{\mu}_{t,h}, \mu_{t,h}). \qquad (22)
\end{aligned}$$

The last inequality comes from the Kantorovich-Rubinstein dual formulation of the Wasserstein-1 distance in Eq. (8). Using the inequalities in Eq. (21) and Eq. (22) for Eq. (19) and Eq. (20), respectively, we get that

$$|\hat{r}(\tilde{\mu}_{t,h}, \pi_{t,h}) - \hat{r}(\mu_{t,h}, \pi_{t,h})| \leq 2L_r(1 + L_\pi) W_1(\tilde{\mu}_{t,h}, \mu_{t,h}).$$

Summing it up from $h = 1$ to $h = H - 1$, we obtain the desired result:

$$|\tilde{\mathsf{J}}(\boldsymbol{\pi}_t, \tilde{f}_{t-1}) - \mathsf{J}(\boldsymbol{\pi}_t)| \leq 2L_r(1 + L_\pi) \sum_{h=1}^{H-1} W_1(\tilde{\mu}_{t,h}, \mu_{t,h}).$$

We note that $h = 0$ is removed from the sum since the initial distribution is fixed, $\tilde{\mu}_{t,0} = \mu_{t,0} = \mu_0$, therefore $W_1(\tilde{\mu}_{t,0}, \mu_{t,0}) = 0$. □

Lemma 5. *Under Assumptions 1 to 4, and assuming that event in Assumption 3 holds true, and for $h \quad \{1, ..., H\}$ and $t \geq 1$ and fixed policy $\boldsymbol{\pi}_t$, it holds:*

$$W_1(\tilde{\mu}_{t,h}, \mu_{t,h}) \leq 2\beta_{t-1} \overline{L}_{t-1}^{h-1} \sum_{i=0}^{h-1} \int_{\mathsf{S}} \quad \boldsymbol{\sigma}_{t-1}(s, \pi_{t,hh}(s, \mu_{t,i}), \mu_{t,i}) \quad {}_2\, \mu_{t,i}(ds), \qquad (23)$$

where $\overline{L}_{t-1} = 1 + 2(1 + L_\pi)\,[L_f + 2\beta_{t-1} L_{\boldsymbol{\sigma}}]$.

Proof. To simplify the notation when it comes to both f and $\tilde{f}_{t-1}$, we omit the action variable as it is determined via $\boldsymbol{\pi}, \boldsymbol{\mu}$ and the current state, i.e., $f(s, \mu) := f(s, \pi_t(s, \mu), \mu)$ and similarly for $\tilde{f}_{t-1}$. Furthermore, we fix $t \geq 1$.

First, we establish a relationship between $W_1(\tilde{\mu}_{t,h+1}, \mu_{t,h+1})$ and $W_1(\tilde{\mu}_{t,h}, \mu_{t,h})$, i.e., the change in the Wasserstein distance between the state distributions under the true dynamics and the approximated dynamics $\tilde{f}_{t-1}$. Then, we will use this relationship and the fact that the initial distribution is fixed, i.e., $\mu_{t,0} = \tilde{\mu}_{t,0}$, to bound the distance at time h.

First, we rewrite $W_1(\tilde{\mu}_{t,h+1}, \mu_{t,h+1})$ in terms of $\pi_{t,h}, \tilde{\mu}_{t,h}$ and $\mu_{t,h}$, use the triangle inequality for the Wasserstein-1 distance, and use the Kantorovich and Rubinstein formulation from Eq. (8).

$$\begin{aligned}
&W_1(\tilde{\mu}_{t,h+1}, \mu_{t,h+1}) \\
&= W_1(\Phi(\pi_{t,h}, \tilde{\mu}_{t,h}, \tilde{f}_{t-1}), \Phi(\pi_{t,h}, \mu_{t,h}, f)) \\
&\leq W_1(\Phi(\pi_{t,h}, \tilde{\mu}_{t,h}, \tilde{f}_{t-1}), \Phi(\pi_{t,h}, \mu_{t,h}, \tilde{f}_{t-1})) + W_1(\Phi(\pi_{t,h}, \mu_{t,h}, \tilde{f}_{t-1}), \Phi(\pi_{t,h}, \mu_{t,h}, f)) \\
&= \sup_{v: Lip(v) \leq 1} \int_{s \quad \mathsf{S}} v(s\ ) \left(\int_{s \quad \mathsf{S}} \mathsf{P}[\tilde{f}_{t-1}(s, \tilde{\mu}_{t,h}) + \tilde{\omega}_{t,h} \quad ds\ ](\tilde{\mu}_{t,h}(ds) - \mu_{t,h}(ds)) \right) \qquad (24) \\
&\quad + \sup_{v: Lip(v) \leq 1} \int_{s \quad \mathsf{S}} v(s\ ) \left(\int_{s \quad \mathsf{S}} \big(\mathsf{P}[\tilde{f}_{t-1}(s, \tilde{\mu}_{t,h}) + \tilde{\omega}_{t,h} \quad ds\ ] - \mathsf{P}[f(s, \mu_{t,h}) + \omega_{t,h} \quad ds\ ]\big) \mu_{t,h}(ds) \right). \qquad (25)
\end{aligned}$$

Next, we consider terms from Eq. (24) and Eq. (25) separately, and obtain upper-bounds for both of them in terms of $W_1(\tilde{\mu}_{t,h}, \mu_{t,h})$ and $\boldsymbol{\sigma}_{t-1}(s, \pi_{t,h}(s, \tilde{\mu}_{t,h}), \tilde{\mu}_{t,h})$.

Step 1: Bounding Eq. (25) First, we fix v such that $Lip(v) \leq 1$. Then, consider Eq. (25)

$$
\begin{aligned}
&\int_{s\quad S} v(s\ )\left(\int_{s\quad S}\left(\mathsf{P}[\tilde{f}_{t-1}(s,\tilde{\mu}_{t,h})+\tilde{\omega}_{t,h}\quad ds\ ]-\mathsf{P}[f(s,\mu_{t,h})+\omega_{t,h}\quad ds\ ]\right)\mu_{t,h}(ds)\right) \\
&=\int_{s\quad S}\int_{s\quad S} v(s\ )\Big[\mathsf{P}[\tilde{f}_{t-1}(s,\tilde{\mu}_{t,h})+\tilde{\omega}_{t,h}\quad ds\ ]-\mathsf{P}[f(s,\mu_{t,h})+\omega_{t,h}\quad ds\ ]\Big]\mu_{t,h}(ds) && \text{Fubini's thm.} \\
&\leq \int_{s\quad S} W_1(\tilde{f}_{t-1}(s,\tilde{\mu}_{t,h})+\tilde{\omega}_{t,h}, f(s,\mu_{t,h})+\omega_{t,h})\mu_{t,h}(ds) && \textit{Eq. } (8) \\
&\leq \int_{s\quad S} \left\|\tilde{f}_{t-1}(s,\tilde{\mu}_{t,h})-f(s,\mu_{t,h})\right\|_2 \mu_{t,h}(ds). && \textit{Corollary } 1 \qquad (26)
\end{aligned}
$$

Note that the state space is compact hence the first moment of $\tilde{\mu}_{t,h+1}$ and $\mu_{t,h+1}$ exist and $W_1(\tilde{\mu}_{t,h+1},\mu_{t,h+1}) < \infty$, therefore, the integration above is finite and Fubini's theorem is applicable. Next, we consider the following two claims.

Claim 5.1. *Under Assumption 3 and assuming that the event defined in it holds true, we have that*

$$\left\|\tilde{f}_{t-1}(s,\tilde{\mu}_{t,h})-f(s,\tilde{\mu}_{t,h})\right\|_2 \leq 2\beta_{t-1}\quad \boldsymbol{\sigma}_{t-1}(s,\pi_{t,h}(s,\tilde{\mu}_{t,h}),\tilde{\mu}_{t,h})\quad {}_2\,.$$

Proof.

$$
\begin{aligned}
&\left\|\tilde{f}_{t-1}(s,\tilde{\mu}_{t,h})-f(s,\tilde{\mu}_{t,h})\right\|_2 \\
&=\left\|\tilde{f}_{t-1}(s,\tilde{\mu}_{t,h})-\boldsymbol{m}_{t-1}(s,\tilde{\mu}_{t,h})+\boldsymbol{m}_{t-1}(s,\tilde{\mu}_{t,h})-f(s,\tilde{\mu}_{t,h})\right\|_2 \\
&\leq \left\|\tilde{f}_{t-1}(s,\tilde{\mu}_{t,h})-\boldsymbol{m}_{t-1}(s,\tilde{\mu}_{t,h})\right\|_2 + \quad \boldsymbol{m}_{t-1}(s,\tilde{\mu}_{t,h})-f(s,\tilde{\mu}_{t,h})\quad {}_2 \\
&=\sqrt{\sum_{i=1}^{p}\left[\tilde{f}_{t-1}(s,\tilde{\mu}_{t,h})-\boldsymbol{m}_{t-1}(s,\tilde{\mu}_{t,h})\right]_i^2}+\sqrt{\sum_{i=1}^{p}\left[\boldsymbol{m}_{t-1}(s,\tilde{\mu}_{t,h})-f(s,\tilde{\mu}_{t,h})\right]_i^2} \\
&\leq 2\sqrt{\sum_{i=1}^{p}\beta_{t-1}^2\left[\boldsymbol{\sigma}_{t-1}(s,\pi_{t,h}(s,\tilde{\mu}_{t,h}),\tilde{\mu}_{t,h})\right]_i^2} \\
&=2\beta_{t-1}\quad \boldsymbol{\sigma}_{t-1}(s,\pi_{t,h}(s,\tilde{\mu}_{t,h}),\tilde{\mu}_{t,h})\quad {}_2
\end{aligned}
$$

where $[\cdot]_i$ denotes the i-th coordinate. The first inequality follows from the triangle inequality, after which, we used the event defined in Assumption 3 and the fact that $\tilde{f}_{t-1}$ is chosen to satisfy the same condition. □

Claim 5.2. *Under Assumption 1 and Assumption 2, we have that*

$$f(s,\tilde{\mu}_{t,h})-f(s,\mu_{t,h})\quad {}_2 \leq L_f(1+L_\pi)W_1(\tilde{\mu}_{t,h}\mu_{t,h})).$$

Proof. From Assumption 1 and Assumption 2, it follows that

$$
\begin{aligned}
f(s,\tilde{\mu}_{t,h})-f(s,\mu_{t,h})\quad {}_2 &\leq L_f(\quad \pi_{t,h}(s,\tilde{\mu}_{t,h})-\pi_{t,h}(s,\mu_{t,h})\quad {}_2+W_1(\tilde{\mu}_{t,h}\mu_{t,h})) \\
&\leq L_f(1+L_\pi)W_1(\tilde{\mu}_{t,h}\mu_{t,h}).
\end{aligned}
$$

□

Now, we can combine Claim 5.1 and Claim 5.2 to get

$$
\begin{aligned}
&\left\|\tilde{f}_{t-1}(s,\tilde{\mu}_{t,h})-f(s,\mu_{t,h})\right\|_2 && (27) \\
&\quad\leq \left\|\tilde{f}_{t-1}(s,\tilde{\mu}_{t,h})-f(s,\tilde{\mu}_{t,h})\right\|_2 + \quad f(s,\tilde{\mu}_{t,h})-f(s,\mu_{t,h})\quad {}_2 \\
&\quad\leq 2\beta_{t-1}\quad \boldsymbol{\sigma}_{t-1}(s,\pi_{t,h}(s,\tilde{\mu}_{t,h}),\tilde{\mu}_{t,h})\quad {}_2+L_f(1+L_\pi)W_1(\tilde{\mu}_{t,h},\mu_{t,h}). && (28)
\end{aligned}
$$

Substituting back Eq. (28) into Eq. (26) at the beginning of Step 1, we obtain:

$$\begin{aligned}
&\int_{s' \in \mathcal{S}} v(s') \int_{s \in \mathcal{S}} \Big[\mathsf{P}[\tilde{f}_{t-1}(s, \tilde{\mu}_{t,h}) + \tilde{\omega}_{t,h} \in ds'] - \mathsf{P}[f(s, \mu_{t,h}) + \omega_{t,h} \in ds']\Big]\mu_{t,h}(ds) \\
&\leq \int_{s \in \mathcal{S}} \big\|\tilde{f}_{t-1}(s, \tilde{\mu}_{t,h}) - f(s, \mu_{t,h})\big\|_2 \, \mu_{t,h}(ds) \\
&\leq L_f(1 + L_\pi)W_1(\tilde{\mu}_{t,h}, \mu_{t,h}) + 2\beta_{t-1} \int_{s \in \mathcal{S}} \|\boldsymbol{\sigma}_{t-1}(s, \pi_{t,h}(s, \tilde{\mu}_{t,h}), \tilde{\mu}_{t,h})\|_2 \, \mu_{t,h}(ds). \qquad (29)
\end{aligned}$$

Therefore, we obtain an upper bound for Eq. (25).

Step 2: Bounding Eq. (24)

Next, we consider the other part, Eq. (24). Again let v be a fixed function such that $Lip(v) \leq 1$. We have:

$$\begin{aligned}
&\int_{s' \in \mathcal{S}} \int_{s \in \mathcal{S}} v(s')\mathsf{P}[\tilde{f}_{t-1}(s, \tilde{\mu}_{t,h}) + \tilde{\omega}_{t,h} \in ds'](\tilde{\mu}_{t,h}(ds) - \mu_{t,h}(ds)) \\
&= \int_{s' \in \mathcal{S}} \int_{s \in \mathcal{S}} v(s')\Big(\mathsf{P}[\tilde{f}_{t-1}(s, \tilde{\mu}_{t,h}) + \tilde{\omega}_{t,h} \in ds'] - \mathsf{P}[f(s, \tilde{\mu}_{t,h}) + \omega_{t,h} \in ds']\Big)(\tilde{\mu}_{t,h}(ds) - \mu_{t,h}(ds)) \qquad (30) \\
&\quad + \int_{s' \in \mathcal{S}} \int_{s \in \mathcal{S}} v(s')\mathsf{P}[f(s, \tilde{\mu}_{t,h}) + \omega_{t,h} \in ds'](\tilde{\mu}_{t,h}(ds) - \mu_{t,h}(ds)). \qquad (31)
\end{aligned}$$

We use Fubini's theorem again to change the order of integration, and we consider the inner integration from Eq. (30):

$$\begin{aligned}
&\int_{s' \in \mathcal{S}} v(s')\Big(\mathsf{P}[\tilde{f}_{t-1}(s, \tilde{\mu}_{t,h}) + \tilde{\omega}_{t,h} \in ds'] - \mathsf{P}[f(s, \tilde{\mu}_{t,h}) + \omega_{t,h} \in ds']\Big) \\
&\leq W_1(\tilde{f}_{t-1}(s, \tilde{\mu}_{t,h}) + \tilde{\omega}_{t,h}, f(s, \tilde{\mu}_{t,h}) + \omega_{t,h}) && \textit{Eq. } (8) \\
&\leq \big\|\tilde{f}_{t-1}(s, \tilde{\mu}_{t,h}) - f(s, \tilde{\mu}_{t,h})\big\|_2 && \textit{Corollary } 1 \\
&\leq 2\beta_{t-1}\|\boldsymbol{\sigma}_{t-1}(s, \pi_{t,h}(s, \tilde{\mu}_{t,h}), \tilde{\mu}_{t,h})\|_2 . && \textit{Claim } 5.1
\end{aligned}$$

Let $g(s) = \frac{1}{L_{\boldsymbol{\sigma}}(1+L_\pi)}\|\boldsymbol{\sigma}_{t-1}(s, \pi_{t,h}(s, \tilde{\mu}_{t,h}), \tilde{\mu}_{t,h})\|_2$, then note that

$$\begin{aligned}
|g(x) - g(y)| &= \frac{1}{L_{\boldsymbol{\sigma}}(1 + L_\pi)}\Big|\|\boldsymbol{\sigma}_{t-1}(x, \pi_{t,h}(x, \tilde{\mu}_{t,h}), \tilde{\mu}_{t,h})\|_2 - \|\boldsymbol{\sigma}_{t-1}(y, \pi_{t,h}(y, \tilde{\mu}_{t,h}), \tilde{\mu}_{t,h})\|_2\Big| \\
&\leq \frac{1}{L_{\boldsymbol{\sigma}}(1 + L_\pi)}\|\boldsymbol{\sigma}_{t-1}(x, \pi_{t,h}(x, \tilde{\mu}_{t,h}), \tilde{\mu}_{t,h}) - \boldsymbol{\sigma}_{t-1}(y, \pi_{t,h}(y, \tilde{\mu}_{t,h}), \tilde{\mu}_{t,h})\|_2 \\
&\leq \frac{1}{1 + L_\pi}(\|x - y\|_2 + \|\pi(x, \tilde{\mu}_{t,h}) - \pi(y, \tilde{\mu}_{t,h})\|_2) && \textit{Assumption } 4 \\
&\leq \|x - y\|_2 , && \textit{Assumption } 2
\end{aligned}$$

where the first inequality follows from the reverse triangle inequality. Therefore, we can bound Eq. (30) as

$$\begin{aligned}
&\int_{s \in \mathcal{S}} \int_{s' \in \mathcal{S}} v(s')\left(\mathsf{P}[\tilde{f}_{t-1}(s, \tilde{\mu}_{t,h}) + \tilde{\omega}_{t,h} \in ds'] - \mathsf{P}[f(s, \mu_{t,h}) + \omega_{t,h} \in ds']\right)(\tilde{\mu}_{t,h}(ds) - \mu_{t,h}(ds)) \\
&\leq 2\beta_{t-1} \int_{s \in \mathcal{S}} \|\boldsymbol{\sigma}_{t-1}(s, \pi_{t,h}(s, \tilde{\mu}_{t,h}), \tilde{\mu}_{t,h})\|_2 \, (\tilde{\mu}_{t,h}(ds) - \mu_{t,h}(ds)) \\
&= 2\beta_{t-1}L_{\boldsymbol{\sigma}}(1 + L_\pi) \int_{s \in \mathcal{S}} g(s)(\tilde{\mu}_{t,h}(ds) - \mu_{t,h}(ds)) \\
&\leq 2\beta_{t-1}L_{\boldsymbol{\sigma}}(1 + L_\pi)W_1(\tilde{\mu}_{t,h}, \mu_{t,h}). \qquad (32)
\end{aligned}$$

The last inequality comes from the dual formulation of the Wasserstein-1 distance Eq. (8) and the previously shown fact that g is 1-Lipschitz.

Next, we consider the remaining term in Eq. (31), namely,

$$\int_{s\in\mathcal{S}}\int_{s'\in\mathcal{S}} v(s')\mathsf{P}[f(s,\tilde{\mu}_{t,h})+\omega_{t,h}\in ds'](\tilde{\mu}_{t,h}(ds)-\mu_{t,h}(ds)).$$

Let $q(s)=\frac{1}{L_f(1+L_\pi)}\int_{s'\in\mathcal{S}} v(s')\mathsf{P}[f(s,\tilde{\mu}_{t,h})+\omega_{t,h}\in ds']$, then

$$\begin{aligned}
|q(x)-q(y)| &= \frac{1}{L_f(1+L_\pi)}\bigg|\int_{s'\in\mathcal{S}} v(s')\Big(\mathsf{P}[f(x,\tilde{\mu}_{t,h})+\omega_{t,h}\in ds'] \\
&\qquad\qquad - \mathsf{P}[f(y,\tilde{\mu}_{t,h})+\omega_{t,h}\in ds']\Big)\bigg| \\
&\leq \frac{1}{L_f(1+L_\pi)}\Big|W_1(f(x,\tilde{\mu}_{t,h})+\omega_{t,h}, f(y,\tilde{\mu}_{t,h})+\omega_{t,h})\Big| && \textit{Eq. } (8)\\
&\leq \frac{1}{L_f(1+L_\pi)}\|f(x,\tilde{\mu}_{t,h})-f(y,\tilde{\mu}_{t,h})\|_2 && \textit{Corollary } 1\\
&\leq \frac{1}{L_f(1+L_\pi)}\Big(\|x-y\|_2+\|\pi_{t,h}(x,\tilde{\mu}_{t,h})-\pi_{t,h}(y,\tilde{\mu}_{t,h})\|_2\Big) && \textit{Assumption } 1\\
&\leq \|x-y\|_2. && \textit{Assumption } 2
\end{aligned}$$

It follows that q is 1-Lipschitz, and therefore

$$\begin{aligned}
&\int_{s\in\mathcal{S}}\int_{s'\in\mathcal{S}} v(s')\mathsf{P}[f(s,\mu_{t,h})+\omega_{t,h}\in ds'](\tilde{\mu}_{t,h}(ds)-\mu_{t,h}(ds)) \\
&= L_f(1+L_\pi)\int_{s\in\mathcal{S}} q(s)(\tilde{\mu}_{t,h}(ds)-\mu_{t,h}(ds)) \\
&\leq L_f(1+L_\pi)W_1(\tilde{\mu}_{t,h},\mu_{t,h}) \qquad (33)
\end{aligned}$$

The first equality follows directly from the definition of the function q, while the last inequality comes from the Wasserstein Dual formulation of W_1 in Eq. (8).

Combining Step 1 and Step 2:

By combining Eq. (29), Eq. (32) and Eq. (33), we get

$$\begin{aligned}
W_1(\tilde{\mu}_{t,h+1},\mu_{t,h+1}) &\leq [2L_f(1+L_\pi)+2\beta_{t-1}L_{\boldsymbol{\sigma}}(1+L_\pi)]W_1(\tilde{\mu}_{t,h}\mu_{t,h})) \\
&\quad + 2\beta_{t-1}\int_{s\in\mathcal{S}}\|\boldsymbol{\sigma}_{t-1}(s,\pi_{t,h}(s,\tilde{\mu}_{t,h}),\tilde{\mu}_{t,h})\|_2\,\mu_{t,h}(ds).
\end{aligned}$$

Also, note that

$$\begin{aligned}
&\|\boldsymbol{\sigma}_{t-1}(s,\pi_{t,h}(s,\tilde{\mu}_{t,h}),\tilde{\mu}_{t,h})\|_2 \\
&\quad= \|\boldsymbol{\sigma}_{t-1}(s,\pi_{t,h}(s,\tilde{\mu}_{t,h}),\tilde{\mu}_{t,h})-\boldsymbol{\sigma}_{t-1}(s,\pi_{t,h}(s,\mu_{t,h}),\mu_{t,h})+\boldsymbol{\sigma}_{t-1}(s,\pi_{t,h}(s,\mu_{t,h}),\mu_{t,h})\|_2 \\
&\quad\leq L_{\boldsymbol{\sigma}}(\|\pi_{t,h}(s,\tilde{\mu}_{t,h})-\pi_{t,h}(s,\mu_{t,h})\|_2+W_1(\tilde{\mu}_{t,h},\mu_{t,h}))+\|\boldsymbol{\sigma}_{t-1}(s,\pi_{t,h}(s,\mu_{t,h}),\mu_{t,h})\|_2 \\
&\quad\leq L_{\boldsymbol{\sigma}}(1+L_\pi)W_1(\tilde{\mu}_{t,h},\mu_{t,h})+\|\boldsymbol{\sigma}_{t-1}(s,\pi_{t,h}(s,\mu_{t,h}),\mu_{t,h})\|_2,
\end{aligned}$$

where the first inequality comes from Assumption 4 and the second from Assumption 2. Therefore, we have the following upper-bound on $W_1(\tilde{\mu}_{t,h+1},\mu_{t,h+1})$:

$$\begin{aligned}
W_1(\tilde{\mu}_{t,h+1},\mu_{t,h+1}) &\leq 2(1+L_\pi)\,[L_f+2\beta_{t-1}L_{\boldsymbol{\sigma}}]\,W_1(\tilde{\mu}_{t,h}\mu_{t,h}) \\
&\quad + 2\beta_{t-1}\int_{s\in\mathcal{S}}\|\boldsymbol{\sigma}_{t-1}(s,\pi_{t,h}(s,\mu_{t,h}),\mu_{t,h})\|_2\,\mu_{t,h}(ds).
\end{aligned}$$

The above inequality establishes the relationship between $W_1(\tilde{\mu}_{t,h+1},\mu_{t,h+1})$ and $W_1(\tilde{\mu}_{t,h},\mu_{t,h})$. We use the fact that $W_1(\tilde{\mu}_{t,0},\mu_{t,0})=0$ and apply this relationship repeatedly $h-1$ times to derive the following upper bound for $W_1(\tilde{\mu}_{t,h}, \mu_{t,h})$:

$$\begin{aligned}
&W_1(\tilde{\mu}_{t,h}, \mu_{t,h}) \\
&\leq 2\beta_{t-1} \sum_{i=0}^{h-1} \left[2(1+L_\pi)\left[L_f + 2\beta_{t-1}L_{\boldsymbol{\sigma}}\right]\right]^{h-1-i} \int_{s\in\mathsf{S}} \|\boldsymbol{\sigma}_{t-1}(s, \pi_{t,i}(s, \mu_{t,i}), \mu_{t,i})\|_2\, \mu_{t,i}(ds) \\
&\leq 2\beta_{t-1} \sum_{i=0}^{h-1} \Big[\underbrace{1 + 2(1+L_\pi)\left[L_f + 2\beta_{t-1}L_{\boldsymbol{\sigma}}\right]}_{=:\overline{L}_{t-1}}\Big]^{h-1-i} \int_{s\in\mathsf{S}} \|\boldsymbol{\sigma}_{t-1}(s, \pi_{t,i}(s, \mu_{t,i}), \mu_{t,i})\|_2\, \mu_{t,i}(ds) \\
&\leq 2\beta_{t-1}\overline{L}_{t-1}^{h-1} \sum_{i=0}^{h-1} \int_{s\in\mathsf{S}} \|\boldsymbol{\sigma}_{t-1}(s, \pi_{t,i}(s, \mu_{t,i}), \mu_{t,i})\|_2\, \mu_{t,i}(ds).
\end{aligned}$$

□

Remark 4. *Consider the setting of Lemma 5. Note that $\{\mu_{t,i}\}_{i=0}^{h-1}$ is the trajectory of state distributions i.e. $\mathsf{P}[s_{t,i} \in A] = \mu_{t,i}(A)$ for all $A \subseteq \mathsf{S}$ as shown in Lemma 1. Therefore,*

$$\begin{aligned}
\sum_{i=0}^{h-1} \int_{s\in\mathsf{S}} \|\boldsymbol{\sigma}_{t-1}(s, \pi_{t,i}(s, \mu_{t,i}), \mu_{t,i})\|_2\, \mu_{t,i}(ds) &= \sum_{i=0}^{h-1} \mathsf{E}\left[\|\boldsymbol{\sigma}_{t-1}(z_{t,i})\|_2\right] \\
&= \mathsf{E}\left[\sum_{i=0}^{h-1} \|\boldsymbol{\sigma}_{t-1}(z_{t,i})\|_2\right],
\end{aligned}$$

where $z_{t,i} = (s_{t,i}, \pi_{t,i}(s_{t,i}, \mu_{t,i}), \mu_{t,i})$ and the expectation is taken over the initial distribution and the transition stochasticity. The first equality comes from the aforementioned fact, $\mathsf{P}[s_{t,i} \in A] = \mu_{t,i}(A)$, while the second equality comes from the linearity of expectation.

Therefore,

$$W_1(\tilde{\mu}_{t,h}, \mu_{t,h}) \leq 2\beta_{t-1}\overline{L}_{t-1}^{h-1}\mathsf{E}\left[\sum_{i=0}^{h-1} \|\boldsymbol{\sigma}_{t-1}(z_{t,i})\|_2\right]. \tag{34}$$

Lemma 6. *Under the setting of Lemma 5 and assuming that the event in Assumption 3 holds true,*

$$r_t \leq 4\beta_{t-1}L_r(1+L_\pi)\overline{L}_{t-1}^{H-1}H\mathsf{E}\left[\sum_{h=0}^{H-1} \|\boldsymbol{\sigma}_{t-1}(z_{t,h})\|_2\right].$$

Proof. First, note that under the setting of Lemma 6, Lemmas 3 to 5 also hold.

$$\begin{aligned}
r_t &\leq |\tilde{\mathsf{J}}(\boldsymbol{\pi}_t, \tilde{f}_{t-1}) - \mathsf{J}(\boldsymbol{\pi}_t)| && \textit{Eq. } (17) \\
&\leq 2L_r(1+L_\pi)\sum_{h=1}^{H-1} W_1(\tilde{\mu}_{t,h}, \mu_{t,h}) && \textit{Eq. } (18) \\
&\leq 4\beta_{t-1}L_r(1+L_\pi)\sum_{h=1}^{H-1} \overline{L}_{t-1}^{h-1}\mathsf{E}\left[\sum_{i=0}^{h-1} \|\boldsymbol{\sigma}_{t-1}(z_{t,i})\|_2\right] && \textit{Eq. } (34) \\
&\leq 4\beta_{t-1}L_r(1+L_\pi)\overline{L}_{t-1}^{H-1}H\mathsf{E}\left[\sum_{i=0}^{H-1} \|\boldsymbol{\sigma}_{t-1}(z_{t,i})\|_2\right].
\end{aligned}$$

In the last inequality, we used the fact that $\overline{L}_{t-1} \geq 1$ hence $\overline{L}_{t-1}^{H-1} \geq \overline{L}_{t-1}^{h-1}$. Furthermore, $\|\boldsymbol{\sigma}_{t-1}(z_{t,i})\|_2 \geq 0$ for any input combination, hence $\sum_{i=0}^{H-1} \|\boldsymbol{\sigma}_{t-1}(z_{t,i})\|_2 \geq \sum_{i=0}^{h-1} \|\boldsymbol{\sigma}_{t-1}(z_{t,i})\|_2$. □

Now, we are ready to prove Theorem 1.

Proof of Theorem 1.

$$\begin{aligned}
R_T^2 &= \left(\sum_{t=1}^{T} r_t\right)^2 \\
&\leq T\sum_{t=1}^{T} r_t^2 && \text{Cauchy-Schwartz ineq.} \\
&\leq T\sum_{t=1}^{T} 16p\beta_{T-1}^2 L_r^2 (1+L_\pi)^2 \bar{L}_{T-1}^{2H-2} H^2 \mathbb{E}\left[\sum_{i=0}^{H-1} \|\boldsymbol{\sigma}_{t-1}(z_{t,i})\|_2\right]^2 && \textit{Lemma } 6 \\
&\leq 16pT\beta_{T-1}^2 L_r^2 (1+L_\pi)^2 \bar{L}_{T-1}^{2H-2} H^2 \sum_{t=1}^{T} \mathbb{E}\left[\sum_{i=0}^{H-1} \|\boldsymbol{\sigma}_{t-1}(z_{t,i})\|_2\right]^2 \\
&\leq 16pT\beta_{T-1}^2 L_r^2 (1+L_\pi)^2 \bar{L}_{T-1}^{2H-2} H^3 \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=0}^{H-1} \|\boldsymbol{\sigma}_{t-1}(z_{t,i})\|_2^2\right] && \text{Jensen's ineq.} \\
&\leq 16pT\beta_{T-1}^2 L_r^2 (1+L_\pi)^2 \bar{L}_{T-1}^{2H-2} H^3 I_T.
\end{aligned}$$

The expectations in the expressions above are taken over the initial state and transition noises. For the second inequality, we used that β_{t-1} is non-decreasing in t by Assumption 3 while the last inequality follows from $\sum_{t=1}^{T}\sum_{i=0}^{H-1} \|\boldsymbol{\sigma}_{t-1}(z_{t,i})\|_2^2 \leq I_T$ for all $z_{t,i} \in \mathcal{Z}$ for $t = 1, \dots, T$ and $i = 0, \dots, H-1$. Taking square root of both sides yields the desired result. □

C Gaussian Processes

In this section, we specialize our main theorem, Theorem 1, to Gaussian Process models.

C.1 Gaussian Process

The main step of relating Theorem 1 to the Gaussian Process models is to use a Gaussian Process to approximate the unknown dynamics function f. In particular, we select a GP with prior $GP(0, k)$ where k is a positive semi-definite kernel on $\mathcal{Z} \times [p]$ where $[p] = \{1, \dots, p\}$ and $i \in [p]$ specifies the index of the output dimension. The posterior of this Gaussian Process is calculated under the assumption that the observed noise, $\xi_{i,h} = s_{i,h+1} - f(z_{i,h})$, is drawn independently from $\mathcal{N}(0, \lambda \boldsymbol{I}_p)$ for all i and h. Here λ is a free-parameter that does not necessarily depend on the distribution of $\omega_{t,h}$ for any t and h. Then, at the beginning of episode t, the posterior mean and variance functions conditioned on $\mathcal{D}_1 \cup \cdots \cup \mathcal{D}_{t-1}$ obtain a closed-form solution (Rasmussen, 2003):

$$\boldsymbol{m}_{t-1}(z, j) = \boldsymbol{k}_{t-1}(z, j)^\top (\boldsymbol{K}_{t-1} + \lambda \boldsymbol{I}_{t-1})^{-1} \boldsymbol{y}_{t-1}, \tag{35a}$$

$$\boldsymbol{\sigma}_{t-1}^2(z, j) = k((z, j), (z, j)) - \boldsymbol{k}_{t-1}(z, j)^\top (\boldsymbol{K}_{t-1} + \lambda \boldsymbol{I}_{t-1})^{-1} \boldsymbol{k}_{t-1}(z, j), \tag{35b}$$

where $\boldsymbol{I}_{t-1} \in \mathbb{R}^{(t-1)Hp \times (t-1)Hp}$ is the identity matrix and

$$\boldsymbol{y}_{t-1} = \left[[s_{i,h}]_l\right]_{i,h,l=1}^{t-1,H,p} \in \mathbb{R}^{(t-1)Hp},$$

$$\boldsymbol{k}_{t-1}(z) = \left[k((z, j), (z_{i,h}, l))\right]_{i,h,l=1}^{t-1,H,p} \in \mathbb{R}^{(t-1)Hp},$$

$$\boldsymbol{K}_{t-1} = \left[k((z_{i,h}, l), (z_{i',h'}, l'))\right]_{i,h,l,i',h',l'=1}^{t-1,H,p,t-1,H,p} \in \mathbb{R}^{(t-1)Hp \times (t-1)Hp},$$

where $[s_{i,h}]_l$ denotes the l-th element of the vector $s_{i,h}$. We use the following notation for the mean and variance vectors $\boldsymbol{m}_{t-1}(z) := [\boldsymbol{m}_{t-1}(z, 1), \dots, \boldsymbol{m}_{t-1}(z, p)]$ and $\boldsymbol{\sigma}_{t-1}^2(z) := [\boldsymbol{\sigma}_{t-1}^2(z, 1), \dots, \boldsymbol{\sigma}_{t-1}^2(z, p)]$ for all z in $\mathcal{Z}$. In particular, we choose $\lambda = pH$.

C.2 Assumptions and Relevant Results

We make the following assumptions on the semi-definite kernel k and the true dynamics f of the system.

Assumption 5. *The unknown function f belongs to the Reproducing Kernel Hilbert Space (RKHS) of a positive semi-definite kernel $k : (\mathsf{Z} \times [p]) \times (\mathsf{Z} \times [p]) \to \mathsf{R}$, and has bounded RKHS norm, i.e., $\ f\ _k = \sqrt{\ f, f\ _k} \leq B_f$ where $\ \cdot, \cdot\ _k$ is the inner product of the RKHS and B_f is a fixed known positive constant. Furthermore, we assume that $\omega_{t,h}$ is σ-sub-Gaussian for all t and h.*

This assumption is standard when Gaussian Process models are used (e.g., Srinivas et al. (2010); Chowdhury & Gopalan (2019); Curi et al. (2020)).

Assumption 6. *The positive semi-definite kernel k is symmetric, continuously differentiable with bounded derivative, and $k(z, z) \leq 1$, $\ z\ \ \mathsf{Z}$. We define the kernel metric as $d_k(z, z\ ) = \sqrt{k(z, z) + k(z\ , z\ ) - 2k(z, z\ )}$ and assume that it is Lipschitz-continuous, i.e., $d_k(z, z\ ) \leq L_d d(z, z\ )$ where $d(z, z\ ) :=\ \ s - s\ \ _2 +\ \ a - a\ \ _2 + W_1(\mu, \mu\ )$ and L_d is some positive constant.*

We note that the previous assumption is very mild since it holds for the most commonly used kernel functions. Assumption 5 and Assumption 6 also imply Assumption 1 which states that the unknown function f is Lipschitz-continuous, i.e., $\ f(z) - f(z\ )\ _2 \leq L_f d(z, z\ )$ where L_f is a positive constant (see Corollary 4.36 in Steinwart & Andreas (2008)).

First, we show that under Assumption 5 the Gaussian Process conditioned on $\mathsf{D}_1\ \ \cdots\ \ \mathsf{D}_t$ is calibrated for all $t \geq 1$. Our argument follows from the following lemma:

Lemma 7 (Concentration of an RKHS member, Lemma 5 in Chowdhury & Gopalan (2019))**.** *Let $k : \mathsf{X} \times \mathsf{X} \to \mathsf{R}$ be a symmetric, positive semi-definite kernel and $f : \mathsf{X} \to \mathsf{R}$ be a member of the RKHS $\mathsf{H}_k(\mathsf{X})$ of real-valued functions on X with kernel k. Let $\{x_t\}_{t \geq 1}$ and $\{\epsilon_t\}_{t \geq 1}$ is conditionally R-sub-Gaussian for a positive constant R, i.e.,*

$$\ t \geq 0,\ \ \lambda\ \ \mathsf{R}, \mathsf{E}\left[e^{\lambda \epsilon_t} | \mathsf{F}_{t-1}\right] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right),$$

where F_{t-1} is the σ-algebra generated by $\{x_s, \epsilon_s\}_{s=1}^{t-1}$ and x_t. Let $\{y_t\}_{t \geq 1}$ be a sequence of noisy observations at the query points $\{x_t\}_{t \geq 1}$, where $y_t = f(x_t) + \epsilon_t$. For $\lambda > 0$ and $x\ \ \mathsf{X}$, let

$$\mu_{t-1}(x) := k_{t-1}(x)^{|}(K_{t-1} + \lambda I)^{-1} Y_{t-1},$$
$$\sigma_{t-1}^2(x) := k(x, x) - k_{t-1}(x)^{|}(K_{t-1} + \lambda I)^{-1} k_{t-1}(x),$$

where $Y_{t-1} := [y_1, \ldots, y_{t-1}]^{|}$ denotes the vector of observations at $\{x_1, \ldots, x_{t-1}\}$. Then, for any $0 < \delta \leq 1$, with probability at least $1 - \delta$, uniformly over $t \geq 1, x\ \ \mathsf{X}$,

$$|f(x) - \mu_{t-1}(x)| \leq \left(\ f\ _k + \frac{R}{\sqrt{\lambda}} \sqrt{2\left(\log(1/\delta) + \frac{1}{2} \sum_{s=1}^{t-1} \log(1 + \lambda^{-1} \sigma_{t-1}^2(x_s))\right)}\right) \sigma_{t-1}(x).$$

We note that similar results have been shown independently in Durand et al. (2018).

Before showing that the above specified Gaussian Process is calibrated, we define the *maximum mutual information*, γ_t, to measure the maximum information gain the representative agent could possibly obtain on f by observing any set $A\ \ \mathsf{Z} \times [p]$ of size t.

Definition 2 (Maximum Mutual Information)**.** *Let $f : \mathsf{X} \to \mathsf{R}$ be a real-valued function defined on a domain X, and t a positive integer. For each subset $A\ \ \mathsf{X}$, let $\boldsymbol{y}_A$ denote a noisy version of $\boldsymbol{f}_A$ with $\mathsf{P}[\boldsymbol{y}_A | \boldsymbol{f}_A]$. The Maximum Mutual Information after t noisy observations is defined as*

$$\gamma_t(f, \mathsf{X}) := \max_{A\ \ \mathsf{X} : |A| = t} I(\boldsymbol{f}_A; \boldsymbol{y}_A),$$

where $I(\boldsymbol{f}_A; \boldsymbol{y}_A)$ denotes the mutual information between $\boldsymbol{f}_A$ and $\boldsymbol{y}_A$.

Using the Maximum Mutual Information, the following corollary shows that the Gaussian Process satisfies Assumption 3 under Assumption 5 and Assumption 6.

Corollary 4. *Under Assumption 5 and Assumption 6, with probability* $1 - \delta$ *for all* $t \geq 0$ *and* $z \in \mathcal{Z}$

$$\|f(z) - \boldsymbol{m}_t(z)\|_2 \leq \beta_t \|\boldsymbol{\sigma}_t(z)\|_2, \tag{37}$$

where

$$\beta_t = B_f + \frac{\sigma}{\sqrt{\lambda}}\sqrt{2(\log(1/\delta) + \gamma_{ptH}(k, \mathcal{Z} \times [p]))}.$$

Proof. This corollary follows from Lemma 10 in Chowdhury & Gopalan (2019). □

Additionally, the GP also satisfies Assumption 4 under Assumption 5 and Assumption 6. Lemma 8 follows from Lemma 12 from Curi et al. (2020)

Lemma 8. *For all* z *and* z' *in* $\mathcal{Z}$ *and all* $t \geq 0$*, we have*

$$\|\boldsymbol{\sigma}_t(z) - \boldsymbol{\sigma}_t(z')\|_2 \leq d_k(z, z')$$

where d_k *is the kernel metric defined in Assumption 6.*

We note that Lemma 8 and Assumption 6 immediately imply that $\boldsymbol{\sigma}_t$ satisfies Assumption 4 with Lipschitz constant L_d for all $t \geq 0$.

C.3 Regret Bound

Now, we are ready to show the regret bound under the choice of a Gaussian Process model.

First, we recall a partial result in Eq. (39) from Lemma 11 in Chowdhury & Gopalan (2019) stating that under Assumption 5 and Assumption 6

$$\sum_{t=1}^{T} \sum_{z \in \tilde{D}_1 \cup \cdots \cup \tilde{D}_{t-1}} \|\boldsymbol{\sigma}_{t-1}(z)\|_2^2 \leq 2epH\gamma_{pHT}(k, \mathcal{Z} \times [p]) \tag{38}$$

where $\tilde{D}_t$ is an arbitrary observation sets during episode t for $t = 1, 2, 3, \ldots$.

Theorem 2. *Let* $\overline{L}_{T-1} = 1 + 2(1 + L_\pi)\left[L_f + 2\beta_{T-1}L_{\boldsymbol{\sigma}}\right]$*, and let* $\mu_{t,h} \in \mathcal{P}(\mathcal{S})$ *for all* $t, h > 0$*. Then, under Assumption 2, Assumption 5, and Assumption 6 and using a Gaussian Process as described in Appendix C.1, for all* $T \geq 1$ *and fixed constant* $H > 0$*, with probability at least* $1 - \delta$*, the regret of* M^3*–UCRL is at most:*

$$R_T \leq \mathcal{O}\left(\beta_{T-1}L_r(1 + L_\pi)\overline{L}_{T-1}^{H-1}H^2\sqrt{T\gamma_{pHT}(k, \mathcal{Z} \times [p])}\right). \tag{39}$$

Proof. As noted before, Assumption 5 and Assumption 6 imply Assumption 1, Assumption 3, and Assumption 4, hence, by Theorem 1, with probability at least $1 - \delta$

$$\begin{aligned} R_T &= \mathcal{O}\left(\beta_T L_r(1 + L_\pi)\overline{L}_{T-1}^{H-1}\sqrt{H^3 T I_T}\right) \\ &= \mathcal{O}\left(\beta_T L_r(1 + L_\pi)\overline{L}_{T-1}^{H-1}\sqrt{H^3 T \max_{\tilde{D}_1, \ldots, \tilde{D}_T} \sum_{t=1}^{T} \sum_{z \in \tilde{D}_1 \cup \cdots \cup \tilde{D}_t} \|\boldsymbol{\sigma}_{t-1}(z)\|_2^2}\right), \end{aligned}$$

where $\overline{L}_{t-1} = 1 + 2(1 + L_\pi)\left[L_f + 2\beta_{t-1}L_{\boldsymbol{\sigma}}\right]$. Then we can substitute in Eq. (38) to reach the desired result. □

Theorem 2 shows that the sublinearity of our regret depends on the Maximum Mutual Information rates, in particular, for sublinear MMI rate the regret is sublinear. Srinivas et al. (2010) consider the most commonly used kernels, and obtain the following sublinear rates for $D \quad \mathbb{R}^d$: Linear kernels $\gamma_t(f, D) = \mathcal{O}(d \log t)$, Squared Exponential kernels $\gamma_t(f, D) = \mathcal{O}((\log t)^{d+1})$, and Matérn kernels $\gamma_t(f, D) = \mathcal{O}(t^{d(d+1)/(2\alpha+d(d+1))}(\log t))$ with $\alpha > 1$. Furthermore, Krause & Ong (2011) show the sublinear property of specific composite kernels. We note that this line of work provides large freedom for kernel design on Euclidean spaces, however, the state space in our problem includes the mean-field distribution space $\mathcal{P}(\mathcal{S})$, therefore, the results are not readily applicable. Proving similar bounds on probability measure spaces is a promising direction for future works.

D Experiment Implementation

D.1 Hallucinated Control Implementation

In Section 3, we introduced M^3–UCRL, an algorithm that optimizes over the set of plausible dynamics $\mathcal{F}_{t-1}$ and admissible policies Π (see Eq. (5)). However, as noted in the paper, optimizing over $\mathcal{F}_{t-1}$ is intractable in most of the cases. In this sections, we describe a practical and equivalent reformulation of Eq. (5) that helps parametrizing the problem and enables using gradient based optimization to find $\boldsymbol{\pi}_t$ at every episode t. First, we introduce an auxiliary function $\eta : \mathcal{Z} \to [-1, 1]^p$ (as in Moldovan et al. (2015); Curi et al. (2020)) where p is the dimensionality of the state space $\mathcal{S}$ and define the following hypothetical dynamics model:

$$\tilde{f}_{t-1}(z) = \boldsymbol{m}_{t-1}(z) + \beta_{t-1}\boldsymbol{\Sigma}_{t-1}(z)\eta(z). \tag{40}$$

Conditioned on the event from Assumption 3 holding true, $\tilde{f}_{t-1}$ for any η is calibrated, i.e., $\tilde{f}_{t-1} \quad \mathcal{F}_{t-1}$. Furthermore, by the confidence interval in Assumption 3, every function in $\mathcal{F}_{t-1}$ can be written in the form of Eq. (40), i.e.,

$$\tilde{f}_{t-1} \quad \mathcal{F}_{t-1}, \qquad \eta : \mathcal{Z} \to [-1, 1]^p \quad \text{such that} \quad \tilde{f}_{t-1}(z) = \boldsymbol{m}_{t-1}(z) + \beta_{t-1}\boldsymbol{\Sigma}_{t-1}(z)\eta(z) \qquad z \quad \mathcal{Z}.$$

Therefore, we can reformulate Eq. (5) as an optimization over the set of admissible policies Π and auxiliary function $\eta : \mathcal{Z} \to [-1, 1]^p$ as

$$\boldsymbol{\pi}_t = \arg\max_{\boldsymbol{\pi} \quad \Pi} \max_{\eta:\mathcal{Z} \to [-1,1]^p} \mathbb{E}\left[\sum_{h=0}^{H-1} r(\tilde{s}_{t,h}, \tilde{a}_{t,h}, \tilde{\mu}_{t,h})\right] \tag{41a}$$

$$\text{s.t.} \quad \tilde{a}_{t,h} = \pi_{t,h}(\tilde{s}_{t,h}, \tilde{\mu}_{t,h}), \quad \tilde{z}_{t,h} = (\tilde{z}_{t,h}), \tag{41b}$$

$$\tilde{f}_{t-1}(\tilde{z}_{t,h}) = \boldsymbol{m}_{t-1}(\tilde{z}_{t,h}) + \beta_{t-1}\boldsymbol{\Sigma}_{t-1}(\tilde{z}_{t,h})\eta(\tilde{z}_{t,h}), \tag{41c}$$

$$\tilde{s}_{t,h+1} = \tilde{f}_{t-1}(\tilde{z}_{t,h}) + \tilde{\omega}_{t,h}, \tag{41d}$$

$$\tilde{\mu}_{t,h+1} = \Phi(\tilde{\mu}_{t,h}, \pi_{t,h}, \tilde{f}_{t-1}). \tag{41e}$$

We note that for a fixed policy π, the auxiliary function $\eta(\cdot)$ can be rewritten as $\eta(s, a, \mu) = \eta(s, \pi(s, \mu), \mu) = \eta(s, \mu)$. In essence, this turns the function $\eta(\cdot)$ into an additional policy that exerts "hallucinated" control over the set of plausible models $\mathcal{F}$ (Curi et al., 2020).[3]

Taking expectation in Eq. (41a) still hinders the practicality of the algorithm, however, we note that Corollary 3 applies to Eq. (41) and the following optimization problem is equivalent,

$$\boldsymbol{\pi}_t = \arg\max_{\boldsymbol{\pi} \quad \Pi} \max_{\eta:\mathcal{Z} \to [-1,1]^p} \sum_{h=0}^{H-1} \hat{r}(\tilde{\mu}_h, \pi_{t,h}) \tag{42a}$$

$$\text{s.t.} \quad \tilde{f}_{t-1}(\tilde{z}_{t,h}) = \boldsymbol{m}_{t-1}(\tilde{z}_{t,h}) + \beta_{t-1}\boldsymbol{\Sigma}_{t-1}(\tilde{z}_{t,h})\eta(\tilde{z}_{t,h}), \tag{42b}$$

$$\tilde{\mu}_{t,h+1} = \Phi(\tilde{\mu}_{t,h}, \pi_{t,h}, \tilde{f}_{t-1}) \tag{42c}$$

[3] We remark that the introduction of $\eta(\cdot)$ policy is a simple trick used before in Moldovan et al. (2015) and Curi et al. (2020) that is suitable for practical implementation of model-based reinforcement learning algortihms.

where Φ is defined in Eq. (2) and $\hat{r}$ is defined in Eq. (14) as

$$\hat{r}(\mu, \pi) = \int_{\mathcal{S}} \mu(ds) r(s, \pi(s, \mu), \mu).$$

Reformulating the optimization problem as in Eq. (42) and choosing parametrizable functions, e.g., Neural Networks, for $\boldsymbol{\pi}$, η, and the statistical estimators $\boldsymbol{m}_{t-1}$ and $\boldsymbol{\Sigma}_{t-1}$ allow using gradient ascent to find an optima for Eq. (42a). Even though running gradient ascent does not guarantee the discovery of global optimum, we found in our experiments (see in Section 5) that it still provides a close approximation. We provide further details on the implementation of the transition function Φ in Appendix D.2

D.2 Transition Function Implementation

The main challenges of implementing the mean-field distribution function Φ from Eq. (2) are representing the mean-field distributions and taking the integral over the state space $\mathcal{S}$. For our experiments, we discretize the state-space $\mathcal{S}$ with 104 unit-sized cells. We denote these cells by C_{ij} for $i, j \quad \{0, \dots, 10\}$ where C_{ij} denotes the cell $[i, i+1) \times [j, j+1)$. Note that cells corresponding to walls (see Fig. 1a), e.g., $C_{5,0}$, $C_{5,1}$, or C_{55}, are not part of the state-space. For each episode t and time-step h, the i, j entry of $\mu_{t,h}$ represents the probability that the representative agent lies in that cell, i.e., $[\mu_{t,h}]_{i,j} = \mathsf{P}[s_{t,h} \quad [i, i+1) \times [j, j+1)]$. Furthermore, we denote the middle of each cell $C_{i,j}$ by $c_{i,j}$, i.e., $c_{i,j} = (i + 0.5, j + 0.5)$.

The transition function Φ is then implemented as follows, for every $i, j \quad \{0, \dots, 10\}$ (for which C_{ij} does not correspond to a wall), every episode $t = 1, 2, \dots$, and every time-step $h = 0, \dots, 19$

$$[\mu_{t,h+1}]_{i,j} = \sum_{k,l} \mathsf{P}[f(c_{k,l}, \pi_{t,h}(c_{k,l}, \mu_{t,h}), \mu_{t,h}) + \omega_{t,h} \quad C_{i,j}][\mu_{t,h}]_{k,l}$$

We assume that the noise term $\omega_{t,h}$ is Gaussian with 0 mean and $\boldsymbol{\sigma}^2$ variance and the two dimensions are independent, therefore,

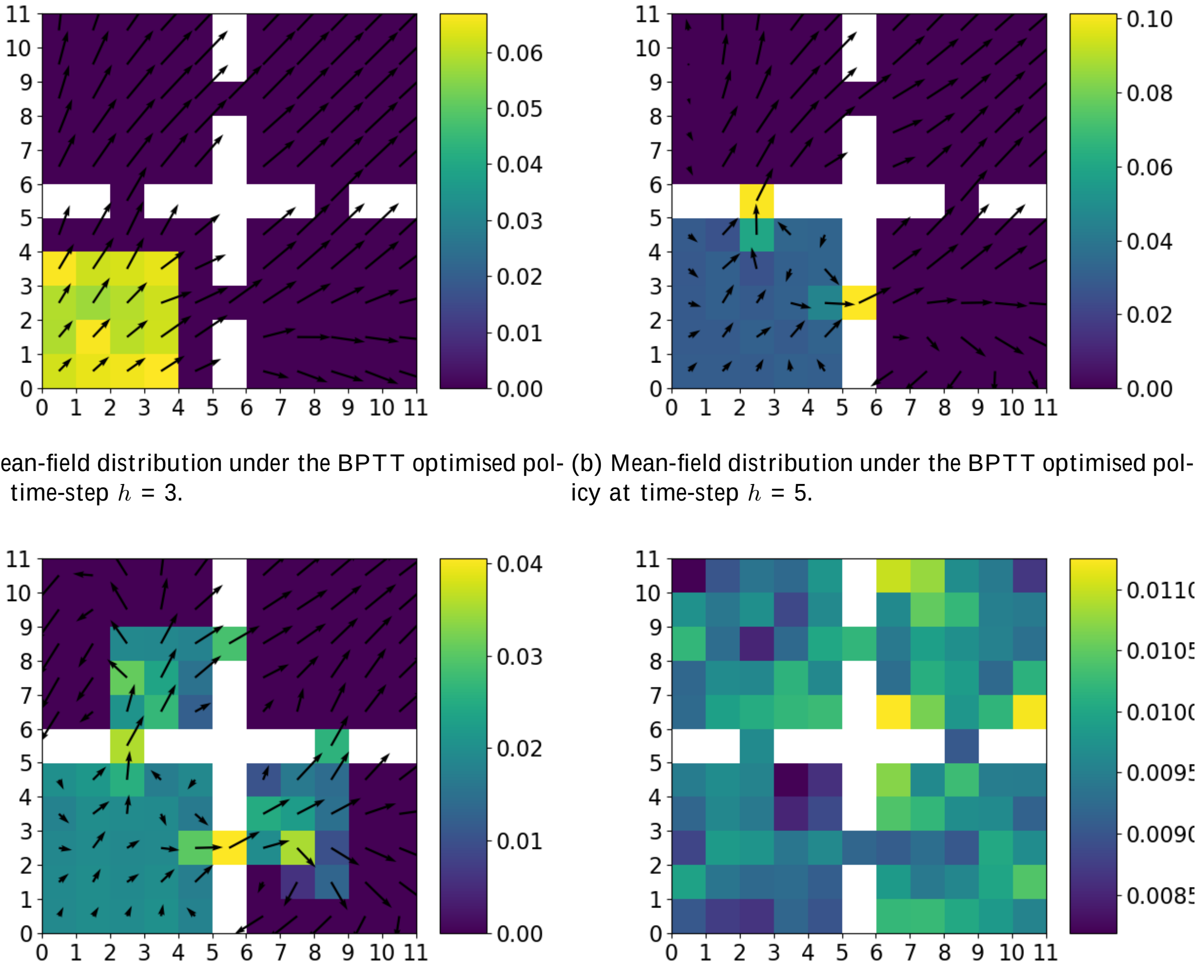
$$\begin{aligned}
&\mathsf{P}[f(c_{k,l}, \pi_{t,h}(c_{k,l}, \mu_{t,h}), \mu_{t,h}) + \omega_{t,h} \quad C_{i,j}] \\
&= \mathsf{P}[f(c_{k,l}, \pi_{t,h}(c_{k,l}, \mu_{t,h}), \mu_{t,h})_1 + \omega_{t,h,1} \quad [i, i+1)] \times \mathsf{P}[f(c_{k,l}, \pi_{t,h}(c_{k,l}, \mu_{t,h}), \mu_{t,h})_2 + \omega_{t,h,2} \quad [j, j+1)] \\
&= (\phi(i - f(c_{k,l}, \pi_{t,h}(c_{k,l}, \mu_{t,h}), \mu_{t,h})_1) - \phi(i + 1 - f(c_{k,l}, \pi_{t,h}(c_{k,l}, \mu_{t,h}), \mu_{t,h})_1)) \\
&\qquad \times (\phi(j - f(c_{k,l}, \pi_{t,h}(c_{k,l}, \mu_{t,h}), \mu_{t,h})_2) - \phi(j + 1 - f(c_{k,l}, \pi_{t,h}(c_{k,l}, \mu_{t,h}), \mu_{t,h})_2))
\end{aligned}$$

where ϕ is the cumulative distribution function of $N(0, \boldsymbol{\sigma}^2)$.

In the implementation, we also redistribute some of the probability mass during transition based on the walls location. If there is a wall between $C_{i,j}$ and $f(c_{k,l}, \pi_{t,h}(c_{k,l}, \mu_{t,h}), \mu_{t,h}) + \omega_{t,h}$, we set $\mathsf{P}[f(c_{k,l}, \pi_{t,h}(c_{k,l}, \mu_{t,h}), \mu_{t,h}) + \omega_{t,h} \quad C_{i,j}] = 0$. Similarly, for $C_{i,j}$ adjacent to a wall or border we increase $\mathsf{P}[f(c_{k,l}, \pi_{t,h}(c_{k,l}, \mu_{t,h}), \mu_{t,h}) + \omega_{t,h} \quad C_{i,j}]$ by the amount that is set to zero due to the adjacent wall.

D.3 Transition Model estimation

To estimate the unknown transition function f in Section 5, we use deep ensemble models (Lakshminarayanan et al., 2017). In particular, we use an ensemble of 10 feed-forward Neural Networks (NNs) with one hidden layer of size 32 and leaky ReLu action functions. We use two output layers joined to the hidden middle layer. The first one uses linear activation and returns the mean of the function while the second returns the estimated variance using softplus activation. We follow the optimisation procedure described in (Lakshminarayanan et al., 2017) that minimizes the negative log-likelihood for each NN under the assumption of heteroscedastic Gaussian noise. We included the adversarial training procedure as well for robustness and smoothing. In prediction time, the ensemble estimates the mean and variance of the unknown function via the empirical mean and variance of the Neural Networks' mean outputs.



(a) Mean-field distribution under the BPTT optimised policy at time-step h = 3.

(b) Mean-field distribution under the BPTT optimised policy at time-step h = 5.

(c) Mean-field distribution under the BPTT optimised policy at time-step h = 8.

(d) Mean-field distribution under the BPTT optimised policy at time-step h = 20.

Figure 3: Mean-field distribution following the policy optimised via BPTT at four time-steps ($h = 3, 5, 8, 20$) in a single policy roll-out. Arrows depict the actions taken by the agents in the corresponding squares. Note that the coloring range changes with the time-steps.

D.4 Policy optimisation with known environment

We consider the policy optimised via back-propagation through time, $\boldsymbol{\pi}^{BPTT}$, as a reasonable point-of-reference for our convergence analysis as it successfully disperses the population in the state space quickly without any notable improvements in its strategy.

Fig. 3 shows the mean-field distribution over one episode following $\boldsymbol{\pi}^{BPTT}$ [4]. As shown on Fig. 3a, in the first few steps $\boldsymbol{\pi}^{BPTT}$ spreads the distribution evenly in the bottom-left corner, however, in the next two steps as shown on Fig. 3b it concentrates a larger mass into the corridors to the neighbouring rooms. It is an optimal choice of action because the corridors act as bottlenecks hence the more it can push through at once the easier it will be to achieve a uniform distribution in the latter time-steps. In the next 3 steps until $h = 8$, $\boldsymbol{\pi}^{BPTT}$ further propagates a larger portion of the population mass to the empty top-right room. In the meanwhile, it starts to distribute the population in the top-left and bottom-right rooms as well. In the following time-steps, $\boldsymbol{\pi}^{BPTT}$ focuses on achieving an even distribution within the rooms without pushing

[4] The supplementary material includes animation of the whole episode named as `known_dynamics_animation.mp4`

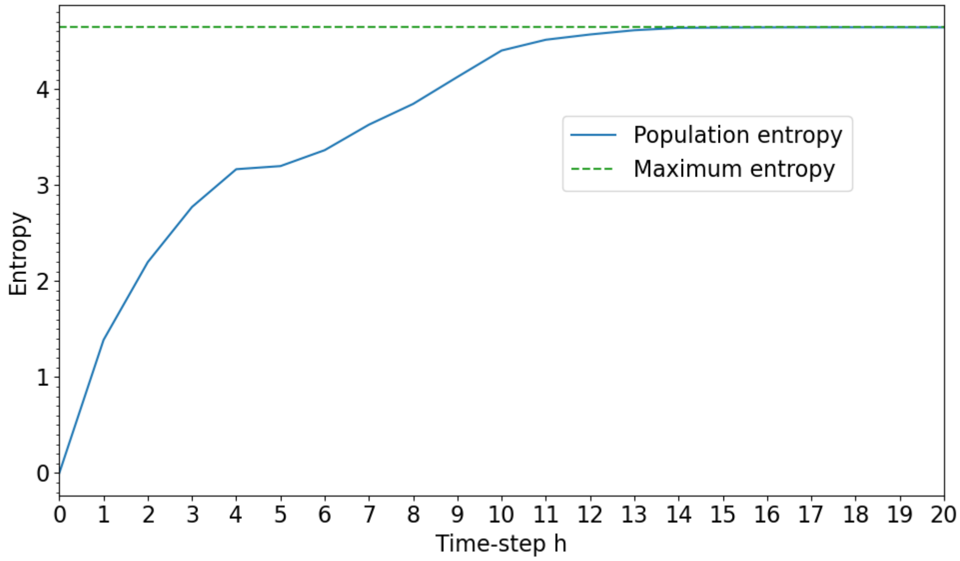


Figure 4: Reward and population action size per time-step in one episode. The policy followed by the agents were optimised via BPTT.

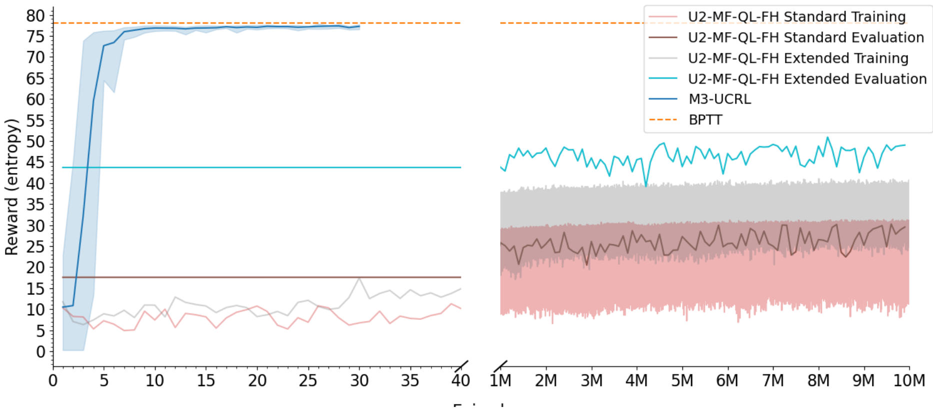


Figure 5: Rewards achieved by *U2-MF-QL-FH* over 10 million training episodes. *Training Rewards* show the estimated reward by *U2-MF-QL-FH* while *Evaluation Rewards* show the performance of the greedy policy in a roll-out. The policy is evaluated after every million training episodes. The hyperparameters of the training were $\omega^Q = 0.7, \omega^\nu = 0.05$ and $\epsilon = 0.15$.

significant population mass through the corridors. Finally, Fig. 3d shows the mean-field distribution at the end of the episode, $h = 20$, when the population is distributed uniformly and maximized its entropy. We provide a further animation with every time-steps as part of the supplementary material.

Fig. 4 shows the mean-field distribution entropy over the time-steps of the same episode as in Fig. 3. $\boldsymbol{\pi}^{BPTT}$ successfully increases the entropy until time-step 14 where it reaches the maximum achievable by the uniform distribution and then maintains this level. The small stagnation in entropy from $h = 4$ to $h = 5$ is due to the bottleneck of the corridors as shown on Fig. 3b.

E Additional Experiments

E.1 Comparison to model-free reinforcement learning

M^3–UCRL is a unique algorithm in the Mean-Field Control literature because it considers continuous state and action spaces and finite time-horizon. As described in Section 2, the goal in this setting is to find an optimal policy that considers the evolution of the mean-field distribution over one episode for a fixed initial distribution. In comparison, the majority of the works consider finite spaces and infinite time-horizon in which the goal is to find the optimal stationary pair of policy and mean-field distribution. We note that this is a significantly different solution concept to ours, hence, proposed algorithms are not suitable to solve the exploration via entropy maximization problem described in Section 5. To the best of our knowledge, the only algorithm that considers a finite time-horizon problem is the *U2-MF-QL-FH* proposed by Angiuli et al. (2021). This algorithm is a Q-Learning variant that optimises the Q values via a two timescale approach. Even though, the problem formulation for *U2-MF-QL-FH* is similar to ours described in Section 2, there are notable differences. First, *U2-MF-QL-FH* is proposed for finite state and action spaces, uses a tabular representation for the Q values and select deterministic actions. Second, it is model-free and assumes access to simulator of one-step transitions. On the other hand, M^3–UCRL does not rely on a tabular representation, learns in continuous spaces, and have limited interactions with the environment to reduce costs associated with these interactions.

Due to the formulation of *U2-MF-QL-FH*, we ran experiments in the finite state and action space variant of the Exploration via entropy maximization problem as described in Geist et al. (2021); Laurière et al. (2022). As we demonstrate below, the deterministic policy used by *U2-MF-QL-FH* is a significant limitation, therefore, we also alter its action space to be a set of 9 distributions over the 5 potential actions from which *U2-MF-QL-FH* chooses [5]. We call the former *Standard* formulation and the latter *Extended*.

As shown on Fig. 5, *U2-MF-QL-FH* converges slowly with high variance in rewards for both formulations, however, the *Extended* version outperforms the *Standard* version for both training and evaluation performance as expected. One of the main factors for slow learning is that the Q values are hard to estimate due to the large number of state-action-time pairs and infrequent visits of many states. In particular, the Q value table has $|\mathcal{S}| \times |\mathcal{A}| \times H = 104 \times 5 \times 20 = 10,400$ entries ($18,720$ for the *Extended* version), and states far from the bottom left corner are rarely visited because they can be reached only in the later steps of an episode. While extending the action-space to non-deterministic actions clearly improves performance, it increases significantly the tabular representation leading to even longer training required. Even after 10 million episodes, we observed large changes in the Q value table indicating that the algorithm would need further iterations, but we had to terminate it due to computational limits. These results further highlights the contribution of M^3–UCRL to the literature.

E.2 Swarm Motion Experiments

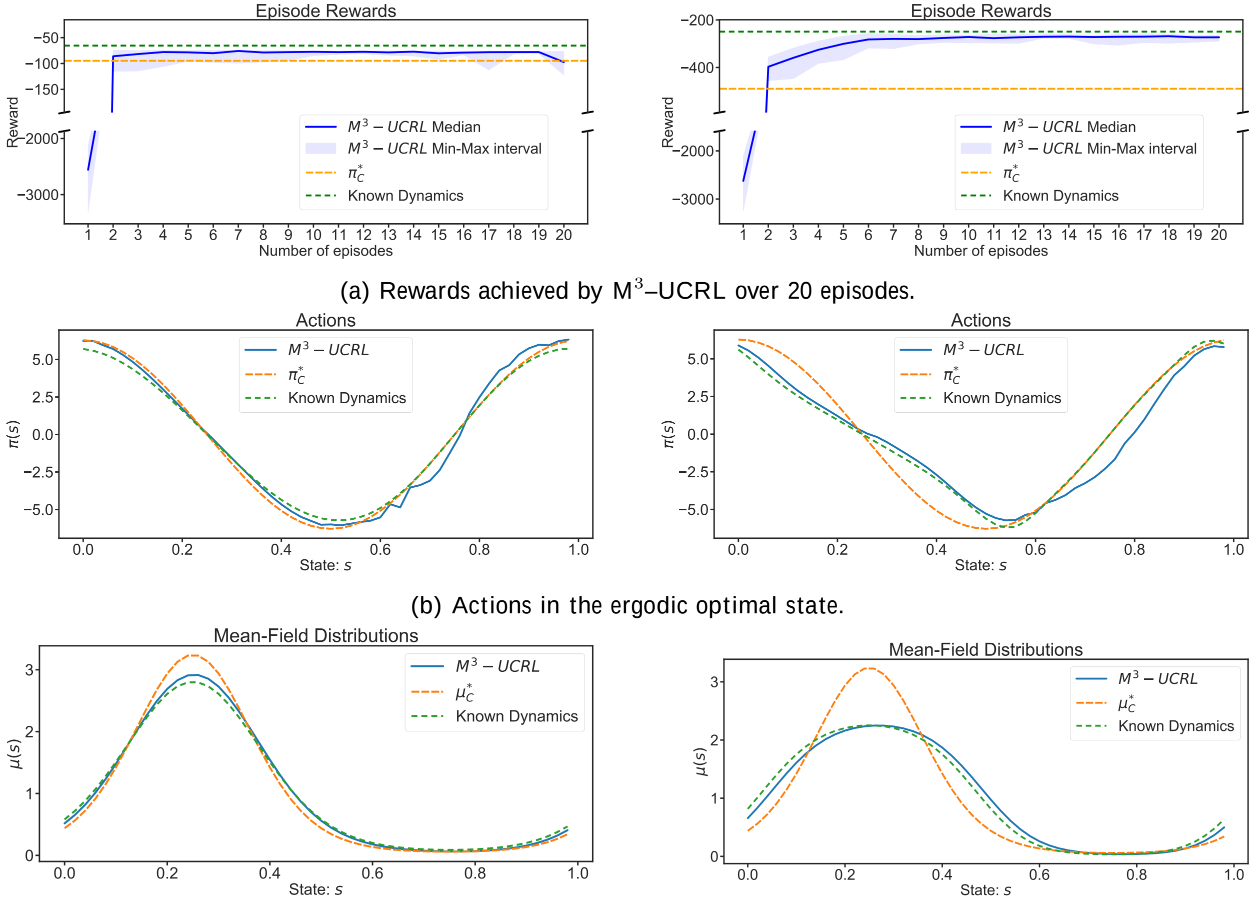
To complement experiments in Section 5, we also implemented M^3–UCRL with Gaussian Process models for the *swarm motion* problem. This setting considers an asymptotically large ($N \to \infty$) population of agents moving around in a space, maximizing a location-dependent reward function, and avoiding congested areas (Almulla et al., 2017).

We choose the state space $\mathcal{S}$ as the $[0, 1]$ interval with periodic boundary conditions, i.e., the unit torus, and the action space $\mathcal{A}$ as the compact interval $[-7, 7]$. Each episode partitions the unit time length into $H = 200$ equal steps of length $\Delta h = 1/200$. The unknown dynamics of the system is given by

$$f(s, a) = s + a\Delta h, \tag{43}$$

and $\omega_{t,h}$ $N(0, \Delta h)$ for all t and h. We note that, for now, the dynamics do not depend on the mean-field distribution, but later on in Eq. (45), we also consider such a case. The reward function is of the form

[5] The set of distribution consists of the following: 5 deterministic actions (4 directions plus staying idle), 4 50% − 50% split between adjacent moves (up-right, right-down, down-left, left-up) and 1 distribution that assigns 25% to all 4 directions. Note that the action space of the *Standard* formulation is a subset of this action space, therefore, it is expected that the *Extended* version achieves better performance.



(a) Rewards achieved by M^3–UCRL over 20 episodes.

(b) Actions in the ergodic optimal state.

(c) Mean-field distribution in the ergodic optimal state.

Figure 6: Results for dynamics model: **(Left column)** $f(s, a) = s + a\Delta h$, and **(Right column)** $f(s, a, \mu) = s + a(4 - 4\mu(s))\Delta h$. (π_C, μ_C) represent an analytic solution to the continuous time swarm motion problem, and thus, it is not discrete-time optimal. The main discrete-time benchmark is Known Dynamics that unlike our M^3–UCRL knows the true dynamics. In both cases (left and right), M^3–UCRL converges in a small number of episodes and discovers a near-optimal policy.

$r(s, a, \mu) = \varphi(s) - \frac{1}{2}|a| - \ln(\mu(s))$, where the second term penalizes large actions, the third term introduces the aversion to crowded regions and the first term defines the reward received at the position s, namely, $\varphi(s) = -2\pi^2[-\sin(2\pi s) + |\cos(2\pi s)|^2] + 2\sin(2\pi s)$.

The advantage of this task is that one can analytically obtain the following optimal solution in the *continuous-time* setting when $\Delta h \to 0$ and the time horizon is infinite, i.e., $H \to \infty$ Almulla et al. (2017):

$$\begin{aligned} \pi_C(s, \mu) &= 2\pi \cos(2\pi s), \\ \mu_C(s) &= \frac{\exp(2\sin(2\pi s))}{\int \exp(2\sin(2\pi s\ ))ds}, \end{aligned} \tag{44}$$

where π_C and μ_C form an ergodic solution, i.e., $\mu_C = \Phi(\mu_C, \pi_C, f)$. We use the continuous time optimal solutions from Eq. (44) (where the subscript C stands for continuous-time solutions) as a benchmark for comparison, however, we note that our time discretization introduces certain deviations. Consequently, π_C might no longer be optimal, and so we also compare our solution to the one obtained under the same setup of discrete time steps and optimization procedure when the true dynamics are *known* (Known Dynamics). This corresponds to our main benchmark.

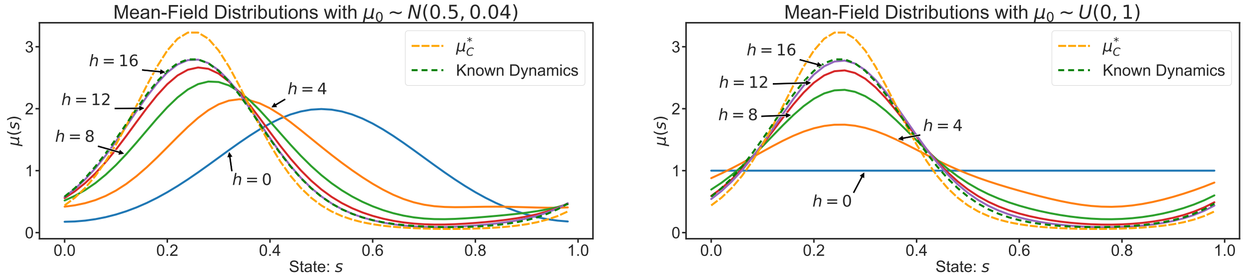


Figure 7: Mean-Field distributions over one episode with different initial distributions and dynamics given by Eq. (43) after training has converged. (**Left side**) μ_0 $N(0.5, 0.04)$ (**Right side**) μ_0 $U(0, 1)$. Already at $h = 16$, the distribution induced by M^3–UCRL policy nearly-matches the one when dynamics are known.

Finally, we also consider the modified swarm motion problem with dynamics given by

$$f(s, a, \mu) = s + a(4 - 4\mu(s))\Delta h. \tag{45}$$

The additional term that depends on the mean-field distribution models the effect of congestion on the transition. Specifically, in comparison to Eq. (43), the more congested an area is, the larger the action an agent must exert to achieve the same movement.

While swarm motion resembles the exploration via entropy maximization problem in Section 5, there are meaningful differences besides the lower dimensional state and action spaces. Namely, certain areas in the state space rewarded differently and large actions are explicitly penalized, therefore, agents not only have to explore the state space but also learn the trade-off between areas and actions. Eq. (45) further extends the swarm motion problem with a dependency on the mean-field distribution that increases the complexity of the agent's learning problem.

Fig. 6 depicts the results when the true dynamics are given by Eq. (43) (Left column) and Eq. (45) (Right column). In Fig. 6a (Left), we show the rewards achieved over 20 episodes. In Fig. 6b (Left) and Fig. 6c (Left), we show the ergodic solution of M^3–UCRL and compare it with (π_C, μ_C) and the solution found when dynamics are known. As expected, the introduced time-discretization results in some deviations from Eq. (44) which allow the policy found when the dynamics are known to exploit them and achieve higher reward. On the other hand, M^3–UCRL converges in a small number of episodes and drives the environment into an ergodic state with policy and mean-field distribution almost identical to the solution under known dynamics (see Fig. 6b (Left) and Fig. 6c (Left)) . We note that, while minimal fluctuations in the episodic rewards are present, the M^3–UCRL is robust in finding close to optimal solutions.

In Fig. 6 (Right column), we show the results obtained for the dynamics from Eq. (45). Even though π and μ in Eq. (44) are no longer applicable for this problem, we include them in the figures to show the effect of the introduced congestion term. Similarly to our previous experiments, M^3–UCRL successfully converges to the solution obtained under known dynamics. Fig. 6b (Right) shows that the two policies are not completely aligned after the same number of episodes as before, which is due to the harder estimation problem and larger confidence estimates.

In the previous experiments, we selected μ as the initial distribution to focus on finding an optimal policy function π. To evaluate the robustness of M^3–UCRL and its ability to drive an arbitrary initial distribution towards the optimal solution, we perform additional experiments with uniform and normal initial distributions. Fig. 7 shows the subsequent mean-field distributions in one episode after the algorithm reached a stable policy. We observe that M^3–UCRL is robust to the initial distribution and swiftly drives the mean-field distribution towards a steady state that maximizes the rewards over time.