Steering No-Regret Agents in MFGs under Model Uncertainty

Leo Widmer **Jiawei Huang** **Niao He**
lewidmer@student.ethz.ch, {jiawei.huang, niao.he}@inf.ethz.ch
Department of Computer Science
ETH Zürich

Abstract

Incentive design is a popular framework for guiding agents' learning dynamics towards desired outcomes by providing additional payments beyond intrinsic rewards. However, most existing works focus on a finite, small set of agents or assume complete knowledge of the game, limiting their applicability to real-world scenarios involving large populations and model uncertainty. To address this gap, we study the design of steering rewards in Mean-Field Games (MFGs) with density-independent transitions, where both the transition dynamics and intrinsic reward functions are unknown. This setting presents non-trivial challenges, as the mediator must incentivize the agents to explore for its model learning under uncertainty, while simultaneously steer them to converge to desired behaviors without incurring excessive incentive payments. Assuming agents exhibit no(-adaptive) regret behaviors, we contribute novel optimistic exploration algorithms. Theoretically, we establish sub-linear regret guarantees for the cumulative gaps between the agents' behaviors and the desired ones. In terms of the steering cost, we demonstrate that our total incentive payments incur only sub-linear excess, competing with a baseline steering strategy that stabilizes the target policy as an equilibrium. Our work presents an effective framework for steering agents behaviors in large-population systems under uncertainty.

1 INTRODUCTION

Mean-Field Games (MFGs) (Huang et al., 2006; Lasry and Lions, 2007) are a widely-used and powerful framework to model the competition and cooperation of large population systems involving symmetric and interchangeable agents. MFGs effectively capture the dynamics of many real-world scenarios, such as macroeconomic models (Steinbacher et al., 2021), road traffic systems (Chen and Cheng, 2010), autonomous vehicle systems (Dinneweth et al., 2022) and auctions (Iyer et al., 2014), and it has been successfully applied in those domains (Gomes et al., 2014; Cabannes et al., 2021; Achdou and Lasry, 2019; Guo et al., 2021). Similar to the finite-agent systems (Roughgarden and Tardos, 2007), MFGs with self-interested agents may lead to undesirable collective behaviors. Typically, the agents' learning dynamics may converge to equilibria where all the participants are worse off compared to other possible outcomes (Guo et al., 2023a).

To address this dilemma, the field of incentive design explores methods to guide agents towards more favorable behaviors by modifying the reward structure. A widely-studied formulation, known as the *steering problem* (Zhang et al., 2024; Canyakmaz et al., 2024; Huang et al., 2024b), assumes the presence of a mediator (incentive designer) outside the game, who can influence the agents' learning dynamics by providing additional steering rewards. However, previous research on steering mainly focuses on either Extensive-Form Games (Zhang et al., 2024) or Markov Games (Canyakmaz et al., 2024; Huang et al., 2024b) with a limited number of agents. The methods developed for those small-scale settings become intractable when applied to large-population scenarios, as the number of agents increases, which is known as the curse of multi-agency.

To address this gap, in this work, we study incentive design in Mean-Field Games (MFGs). More concretely, we focus on the finite-horizon MFGs with density-independent transitions, a standard model in literature (Huang et al., 2006; Lasry and Lions, 2007; Perolat

Proceedings of the 28$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

et al., 2021). In the steering problem setup, we play the role of the mediator, with access to a utility function dependent on the collective behavior of the agents through the population density. During the interactions with the mediator, the agents are continuously learning and adapting. We assume the agents are self-interested no-adaptive-regret learners (Hazan and Seshadhri, 2007); similar no-regret assumptions have been widely adopted in previous literature (Camara et al., 2020; Ge et al., 2024). Following previous works (Zhang et al., 2024; Huang et al., 2024b), our primary goal is to design steering rewards that guide the agents towards desired policies (i.e., minimize the *steering gap*), such that the resulting behaviors maximize the utility function. Meanwhile, the incentives paid by the mediator to the agents, referred to as the *steering cost*, should remain low.

In practice, the mediator usually lacks knowledge of the transition dynamics and the intrinsic reward functions of the MFGs. Therefore, in this work, we focus on the design of steering strategies without prior knowledge of the game model. The model uncertainty makes the steering problem much more challenging, and requires the mediator to strategically balance the exploration and exploitation. Typically, without knowledge of the MFG model, the mediator does not know which agents' behaviors (population densities) are feasible, let alone how to maximize the utility function. Therefore, the mediator needs not only to steer the agents to explore the MFG for its own learning, but also ensure the agents converge to desired outcomes, while keeping the accumulative incentive payments affordable. In summary, the key question we would like to address is:

How can we design effective steering strategies for no-regret agents in MFGs under model uncertainty?

Main contributions We address the above open question by proposing novel exploration algorithms with provable guarantees. We highlight our main contributions in the following. A summary of the main theorems in this paper can be found in Appx. B.

- Firstly, in Sec. 3, we contribute the first formulation for steering in mean-field games, with details about the problem setting and learning objectives.
- Secondly, as preparation, in Sec. 4, we investigate how to steer the agents to a given density or policy. Notably, in Sec. 4.2, we propose a novel steering strategy, which can guide the no-adaptive-regret agents towards any target policy without prior knowledge of the model. This method serves as the key ingredient of our steering algorithms in the following sections.
- Thirdly, in Sec. 5 and Sec. 6, we investigate strategic exploration methods for steering agents in MFGs under uncertainty. In Sec. 5, we start with the setting where the intrinsic reward is zero, and propose an optimism-based exploration algorithm, which guarantees that both the cumulative steering gap and cost only have sub-linear growth. Furthermore, in Sec. 6, we extend our methods to the setting with non-zero and unknown intrinsic reward by integrating a pessimism-based reward estimation strategy. We establish sub-linear regret in steering gap, and show that the total steering cost is only sub-linearly worse, compared to a baseline strategy that stabilizes the target policy as an equilibrium by offsetting differences in intrinsic rewards.

1.1 Closely Related Work

Due to the limit of space, we only discuss closely related works here and defer the others to Appendix C.

Incentive Design in Multi-Agent Systems The problem of incentive design broadly refers to the design of mechanisms for shaping the behavior of autonomous agents (Ehtamo et al., 2002; Ratliff et al., 2019). A recently popular framework for incentive design is known as the steering problem (Zhang et al., 2024; Canyakmaz et al., 2024; Huang et al., 2024b), which considers a repeated interaction between a mediator and learning agents. All of them focuses on small-scale problems (e.g., Markov Games or Extensive-Form Games) and their proposed methods become intractable when extending to large-population setting, including MFGs. Besides, Canyakmaz et al. (2024); Huang et al. (2024b) consider the agents' learning dynamics to be memoryless, which is different from our no-regret assumptions.

Another related direction is contract design[1] (DellaVigna and Malmendier, 2004), which studies the interactions between a principal and agents when the two parties transact in the presence of private information. The fundamental question is how the principal should design the incentives for the agents to maximize its own utility after deducting the payments to agents. However, most of literature study the single-agent setting (Zhu et al., 2022; Ho et al., 2014; Scheid et al., 2024), or focus on the computational aspects without addressing exploration under uncertainty (Dütting et al., 2023; Castiglioni et al., 2023). (Carmona and Wang, 2021; Elie et al., 2019) study contract design in the MFGs setting, but none of them consider model uncertainty. Moreover, a common assumption in those works is that the agents always do the best response (or take equilibrium policies) to the principal's intervention, which is much stronger than our no-adaptive-regret assumption.

[1] To save space, we defer to Appx. C more elaboration of the comparisons between our steering framework and contract design setting.

Besides, Sanjari et al. (2024) consider incentive design in a large-population setting, but they study the Stackelberg games with one leader and a large number of followers, which differs quite substantially to ours. Fu and Horst (2018) consider mean-field leader-follower games, however they assume knowledge of the dynamics, while we consider the steering problem without this knowledge. Moreover, they study the dynamics where the agents cooperate together and optimally respond to the leader's control signal. In contrast, we consider decentralized and self-interested agents with no-regret behaviors in maximizing individual interests.

2 PRELIMINARIES

Mean-Field Games We consider the MFG setting with a finite yet extremely large number of agents, each of which acts independently. In line with Subramanian et al. (2022), we refer to this setting as "Decentralized-MFGs", although their model allows diversity in action space and reward functions for agents.

Definition 2.1. A *Finite-Horizon Decentralized MFG* is defined by a tuple $M = (N, \mathcal{S}, \mathcal{A}, H, \mathbb{P}_M, r_M, \mu_1)$, given the number of agents N; state and action spaces $\mathcal{S}, \mathcal{A}$ with sizes S and A; horizon length H; initial state distribution $\mu_1 \in \Delta_{\mathcal{S}}$. $\mathbb{P}_M := \{\mathbb{P}_{M,h}\}_{h=1}^H$ with $\mathbb{P}_{M,h} : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$ and $r_M := \{r_{M,h}\}_{h=1}^H$ with $r_{M,h} : \mathcal{S} \times \mathcal{A} \times \Delta_{\mathcal{S}\times\mathcal{A}} \to [0, r_{\max}]$ denote the transition and reward function, respectively.

In this paper, we focus on density-independent transition function, a common assumption in previous literature (Huang et al., 2006; Lasry and Lions, 2007; Perolat et al., 2021). For the reward function, we consider the general setup, where the rewards depend on the state-action density (Guo et al., 2021).

We only focus on non-stationary Markovian policies, denoted by $\Pi := \{\pi := \{\pi_h\}_{h\in[H]} | \pi_h : \mathcal{S} \to \Delta(\mathcal{A})\}$. Given a model M, considering an agent taking policy $\pi \in \Pi$, we use $\mu_M^\pi := \{\mu_{M,h}^\pi\}_{h=1}^H$ to denote its state-action density for each step $h \in [H]$. Starting with $\mu_{M,1}^\pi(s,a) = \mu_1(s)\pi_1(a|s)$, for $1 \le h \le H$, we have:

$$\mu_{M,h+1}^\pi(s,a) = \pi_{h+1}(a|s) \sum_{s',a'} \mathbb{P}_{M,h}(s|s',a')\mu_{M,h}^\pi(s',a').$$

When N agents take policies $\pi^1, ...\pi^N \in \Pi$, respectively, the trajectory of agent $n \in [N]$ is specified by:

$$s_1^n \sim \mu_1,\ \forall h \ge 1,\ a_h^n \sim \pi_h^n(\cdot|s_h^n),\ s_{h+1}^n \sim \mathbb{P}_{M,h}(\cdot|s_h^n, a_h^n), \\ r_h^n \leftarrow r_{M,h}(s_h^n, a_h^n, \bar{\mu}_{M,h}). \quad (1)$$

where we use $\bar{\mu}_{M,h} = \frac{1}{N}\sum_{n=1}^N \mu_{M,h}^{\pi^n}$ to denote the *population density* at step h. We also assume r_M is Lipschitz in the density, which is standard in previous works (Guo et al., 2021; Yardim et al., 2022).

Other Notational Convention For convenience, we implicitly treat μ_M^π as a vector in $\mathbb{R}^{HSA}$ concatenated by $\{\mu_{M,h}^\pi\}_{h\in[H]}$. We denote $\Psi_M := \{\mu_M^\pi : \pi \in \Pi\} \subseteq \Delta_{\mathcal{S}\times\mathcal{A}}^H$ to be the set of all feasible state-action densities given M. Note that Ψ_M is a convex set (see Lem. E.1), which implies $\bar{\mu}_M \in \Psi_M$. If it is not necessary to distinguish what model M we use, we omit it in the sub-scriptions, for example, μ^π/Ψ instead of μ_M^π/Ψ_M. We also omit h in s_h, a_h if it is clear from the context. With slight abuse of notation, given a population density $\bar{\mu} := \{\bar{\mu}_h\}_{h=1}^H \in \Delta_{\mathcal{S}\times\mathcal{A}}^H$ and a reward function r, we use $r(\bar{\mu}) \in \mathbb{R}^{HSA}$ to denote the reward vector where $(r(\bar{\mu}))_{h,s,a} = r_h(s,a,\bar{\mu}_h)$. In this way, given an arbitrary agent $n \in [N]$ taking policy π, its expected total return conditioning on population density $\bar{\mu}$ can be written as: $\mathbb{E}_{\pi^n}[\sum_{h=1}^H r_h(s_h^n, a_h^n, \bar{\mu}_h)] = \langle r(\bar{\mu}), \mu^{\pi^n}\rangle$.

Given that this paper considers learning under uncertainty, we use M^* to denote the true hidden mean-field model with transition $\mathbb{P}^*$ and intrinsic reward r^*, in order to distinguish it from the estimated ones.

Besides, given a population density $\bar{\mu}$ in a model M, we will use $\bar{\pi}$ to denote the policy, which induces the population density (i.e., $\mu^{\bar{\pi}} = \bar{\mu}$), defined by: $\bar{\pi}_h(\cdot|s) := \bar{\mu}_h(s,\cdot)/\bar{\mu}_h(s)$ (or $\bar{\pi}_h(\cdot|s) = 1/A$ if $\bar{\mu}_h(\cdot) = 0$).

Reward Function Approximation and Eluder Dimension In this paper, we consider the setting where the true intrinsic reward, denoted by r^*, is unknown. Note that the reward function depends on not only the state and action but also the density, which belongs to a high-dimensional continuous space. Therefore, we consider function approximation for reward estimation with the standard realizability assumption.

Assumption A. A reward function class $\mathcal{R}$ is available, s.t. (i) $\forall r \in \mathcal{R}$, $\forall h$, $r_h(\cdot,\cdot,\cdot) \in [0, r_{\max}]$; (ii) $r^* \in \mathcal{R}$.

In the function approximation setting, the fundamental sample efficiency is closely related to the complexity of the function class. We follow previous works (Russo and Van Roy, 2013; Huang et al., 2024a) and utilize the Eluder Dimension as the complexity measure of the function class. Intuitively, the Eluder Dimension is defined to be the length of the longest "independent" sequence, such that each element in the sequence "reveals" some new information about the function class comparing with previous ones.

Definition 2.2 (ε-independent sequence)**.** Given a domain $\mathcal{X}$ and a class of functions $\mathcal{F}$ defined on $\mathcal{X}$, we say $x \in \mathcal{X}$ is ε-independent on $\{x_1, ..., x_J\} \subseteq \mathcal{X}$ if there exists $f, \tilde{f} \in \mathcal{F}$, such that $\sum_{j=1}^J (f(x_j) - \tilde{f}(x_j))^2 \le \varepsilon^2$, but $|f(x) - \tilde{f}(x)| > \varepsilon$.

Definition 2.3 (Eluder Dimension)**.** Given a mean-field reward function class $\mathcal{R}$ and domain $\mathcal{X} := [H] \times \mathcal{S} \times \mathcal{A} \times \Delta_{\mathcal{S}\times\mathcal{A}}$, the Eluder Dimension of $\mathcal{R}$, denoted by $\dim_E(\mathcal{R}, \varepsilon)$, is defined to be the length of the longest sequence $\{x^j\}_{j=1}^J$, such that, for any $i \in [J]$, x^i is ε-independent w.r.t. $\{x^j\}_{j=1}^{i-1}$.

3 THE STEERING PROBLEM FORMULATION FOR MFGS

In this section, we introduce our steering setup. In Sec. 3.1, we first provide our formulation for steering protocol. Then, in Sec. 3.2 we discuss our assumptions on agent's behavior. After that, we introduce the learning objectives and other setups in Sec. 3.3 and 3.4.

3.1 Agent-Mediator Interaction Protocol

We consider a repeated game setup, and summarize the interaction procedure between agents and the mediator in Procedure 1. In each iteration $t \in [T]$, the mediator first selects a steering reward function[2] R^t, which is a mapping from the density space to the *non-negative*[3] reward vector space, upper bounded by $R_{\max}$. Besides, each agent computes a policy and plays the game. The agents' policies result in a population density $\bar{\mu}^t := \frac{1}{N}\sum_{n=1}^N \mu^{\pi^{n,t}}$, by which the mediator realizes the steering reward $R^t(\bar{\mu}^t)$. Then, each agent $n \in [N]$ receives payments from the mediator equal to the expected return induced by the steering reward and the agent's policy, i.e., $\langle R^t(\bar{\mu}^t), \mu^{\pi^{n,t}}\rangle$. We highlight here that in our setup, at each iteration t, the mediator designs the steering reward function R^t without the knowledge of the agents' policies $\pi^{n,t}$, and we do not restrict whether the agents can observe R^t or not before they make decisions. Furthermore, the agents can either independently compute their policies or collaborate. In the next section, we will characterize our assumptions on the agents' behaviors with more details.

Procedure 1 Agent-Mediator Interaction Protocol

```
1: for $t = 1, ..., T$ do
2:     Mediator chooses $R^t : \Delta_{\mathcal{S}\times\mathcal{A}}^H \to [0, R_{\max}]^{HSA}$.
3:     Each agent $n \in [N]$ computes policy $\pi^{n,t} \in \Pi$, resulting in the population density $\bar{\mu}^t$, and gets payment $\langle R^t(\bar{\mu}^t), \mu^{\pi^{n,t}}\rangle$ from the mediator.
4:     Mediator observes $\bar{\mu}^t$ and a trajectory $\{(s_h^{n,t}, a_h^{n,t}, r_h^{n,t} + \xi_h^t)\}_{h\in[H]}$ generated by $\pi^{n,t}$ following Eq. (1), where $n \sim$ Uniform$\{1, 2, ..., N\}$.
5: end for
```

At the end of each iteration, the mediator can observe a trajectory sampled from a random agent with noisy reward samples. We assume noises ξ_h^t are i.i.d. σ-sub-Gaussian random variables with zero mean. We also assume the mediator has access to the population density, which is necessary to estimate the unknown intrinsic reward function from samples.

3.2 Behavioral Assumptions on Agents

We first introduce our no-adaptive regret assumption and its implication, and then make some justification.

Assumption B (No-Adaptive Regret Behavior)**.** In Procedure 1, the adaptive regret for each agent $\forall n \in [N]$, which is defined below, can be upper bounded by some term $\text{AdaReg}(T) = (r_{\max} + R_{\max}) \cdot o(T)$:

$$\max_{\substack{1\le a<b\le T\\ \mu\in\Psi_{M^*}}} \sum_{t=a}^{b} \langle r^*(\bar{\mu}^t) + R^t(\bar{\mu}^t), \mu - \mu^{\pi^{n,t}}\rangle, \qquad (2)$$

where $(r_{\max} + R_{\max})$ is the normalization term. In Appx. D.4 we show that for all the steering rewards that we deploy in this paper we have $R_{\max} = \mathcal{O}(1 + r_{\max})$.

Justification for Assump. B We remark that it is common to consider agents exhibiting no-regret behaviors in previous literature (Deng et al., 2019; Zhang et al., 2024; Brown et al., 2024). Most of these literature assume no-external regret (directly assigning $a = 1$ and $b = T$ in Eq. (2)), which is weaker than our no-adaptive-regret assumption. However, similar stronger assumptions, such as no-dynamic-regret learners, have also been considered in some studies (Ge et al., 2024). Moreover, our no-adaptive-regret assumption is standard when interpreted through the online linear optimization perspective (Hazan and Seshadhri, 2007; Hazan, 2023), where in each iteration, each agent picks a density from the convex set Ψ_{M^*} and receives potentially adversarial feedback $R^t(\bar{\mu}^t)$. Then, Assump. B aligns with the standard no-adaptive regret guarantees in online linear optimization setting, and there are very simple algorithms (e.g., Online Gradient Descent) achieving AdaReg $= \tilde{O}(\sqrt{T})$. We defer more detailed discussion to Appx. D.2.

Under Assump. B, we have the following property, which suggests the collective population will also exhibit no-regret behaviors. This is a useful property we will leverage in algorithm design.

Proposition 3.1 (No-Adaptive-Regret Population Behavior)**.** *Under Assump. B, we have:*

$$\max_{1\le a<b\le T, \mu\in\Psi_{M^*}} \sum_{t=a}^{b} \langle r^*(\bar{\mu}^t) + R^t(\bar{\mu}^t), \mu - \bar{\mu}^t\rangle \\ \le \text{AdaReg}(T) = (r_{\max} + R_{\max}) \cdot o(T).$$

[2] We will use capital R to denote the steering reward to distinguish with intrinsic reward r.

[3] The non-negativity of the steering reward is known as limited liability (Innes, 1990), which is standard in previous works (Zhang et al., 2024; Huang et al., 2024b)

3.3 Performance Metrics

Inspired by the previous works (Zhang et al., 2024; Huang et al., 2024b), we evaluate the steering algorithm from two aspects: the steering gap and the steering cost. We provide the concrete definition in our MFGs setup as follows.

The Steering Gap Intuitively, the steering gap measures the difference between the desired outcomes and the agents' behavior under the mediator's guidance. In this paper, we assume the mediator is given a utility function $U : \Delta^H_{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}$ assigning each population density a utility value. The only assumption we make for it is about the Lipschitz continuity:

Assumption C (Lipschitz Utility Function)**.** $\forall \mu, \mu' \in \Delta^H_{\mathcal{S}\times\mathcal{A}},\ |U(\mu) - U(\mu')| \leq L_U \|\mu - \mu'\|_1$.

The steering gap up to step T is defined by:

$$\Delta_T(\{\bar{\mu}^t\}_{t=1}^T) := \max_{\pi^*\in\Pi} \sum_{t=1}^T U(\mu^{\pi^*}) - U(\bar{\mu}^t)$$

Here $U(\bar{\mu}^t)$ represents the utility paid to the mediator at each iteration $t \in [T]$, induced by the population density $\bar{\mu}^t$. Note that we consider the best density maximizing utility function as the comparator. This can be interpreted as the best population density if all the agents are restricted to take the same policy, and finding the best shared policy is a standard objective in previous MFGs literature.

The Steering Cost The motivation for introducing a steering cost is that the agents will not accept the mediator's guidance for free. A common measure of the cost is the expected total return associated with the reward received by the agents. Formally, suppose at iteration t, the mediator computes a steering reward function R^t, and the N agents select policies $\pi^{1,t}, \pi^{2,t}, ..., \pi^{N,t} \in \Pi$, which induce a population density $\bar{\mu}^t := \frac{1}{N}\sum_{n=1}^N \mu^{\pi^{n,t}}$, then the steering cost is defined to be the average payments to the agents: $C(\bar{\mu}^t, R^t) := \langle R^t(\bar{\mu}^t), \bar{\mu}^t\rangle = \frac{1}{N}\sum_{n=1}^N \langle R^t(\bar{\mu}^t), \mu^{\pi^{n,t}}\rangle$. We will use

$$C_T(\{\bar{\mu}^t, R^t\}_{t=1}^T) := \sum_{t=1}^T C(\bar{\mu}^t, R^t)$$

to denote the accumulative steering gap. Note that the steering rewards are non-negative, the steering cost effectively reflects the strength of the steering signal.

3.4 Two Steering Scenarios and Objectives

In this paper, we consider the case when the mediator does not know the true transition and reward functions of M^*. However, to make it easy for reader to understand our algorithm design and technique contributions, we will start with a special case, where the agents do not have intrinsic rewards, i.e., $r^* = 0$.

Scenario 1: No Intrinsic Reward The goal of this setting to find an incentive design algorithm producing a sequence of R^t such that both the steering gap and the steering cost are sub-linear: $\Delta_T(\{\bar{\mu}^t\}_{t=1}^T) = o(T),\ C_T(\{\bar{\mu}^t, R^t\}_{t=1}^T) = o(T)$.

The motivation for the sub-linear guarantee here is that it implies the average utility converges to the maximum and the average steering cost vanishes. This implies that the incentive design strategies fulfilling these guarantees will eventually pay off as a long-term investment. In Sec. 5, we analyze this case and provide algorithms achieving our objective.

Scenario 2: Non-Zero Intrinsic Reward In Sec. 6, we study the complete setting where the agents' original reward is non-zero and unknown. In this case, the mediator additionally has to estimate the reward function from observed noisy samples and steer the agents based on that. Similarly, we expect sub-linear steering gap $\Delta_T(\{\bar{\mu}^t\}_{t=1}^T) = o(T)$, while for steering cost, we manually choose the "sandboxing reward" as the comparator:

$$C_T(\{\bar{\mu}^t, R^t - \underbrace{(r_{\max}\cdot\mathbf{1} - r^*)}_{\text{sandboxing reward}}\}_{t=1}^T) = o(T).$$

Here we use $\mathbf{1}$ to denote the all-ones vector. Intuitively, because the intrinsic rewards are non-zero, if the desired behavior μ^{π^*} is not an equilibrium induced by r^*, the mediator has to maintain a non-zero steering rewards to avoid the agents deviating from π^*, so we can not expect the average steering cost to vanish to 0 as in Scenario 1. Therefore, we consider the sandboxing reward as a baseline comparator, which mitigates differences in the intrinsic rewards, so that π^* would be a "stable equilibrium" even if the additional steering reward $R^t - (r_{\max}\cdot\mathbf{1} - r^*(\bar{\mu}^t))$ vanishes to zero. Though, we admit that other choices of sandboxing terms may result in lower steering cost, or one can consider optimizing utility and steering cost together. We leave those interesting directions for the future work.

4 STEERING TOWARDS A FIXED TARGET

In this section, we focus on how to design rewards to guide the agents to a target population density or policy, which serves as preparation steps for the following sections. For convenience, we assume the agents' intrinsic rewards are zero and ignore them.

4.1 Warm-Up: Steering in a Known Model

We start with the case when the MFG model is known. In this case, we can also compute Ψ_{M^*} and find the best $\mu^* = \arg\max_{\mu \in \Psi_{M^*}} U(\mu)$. If we want to steer the population to μ^*, one steering reward choice is $R(\mu) = \mathbf{1}\|\mu^* - \mu\|_\infty + \mu^* - \mu$, where the first shift term is to ensure the non-negativity. The key motivation for our choice is that $\langle R(\mu), \mu^* - \mu\rangle = \|\mu - \mu^*\|_2^2$. As a result, if we consider the accumulative performance, we have the following theorem.

Theorem 4.1. *If M^* is known and r^* is zero everywhere, under Asump. B and C, by choosing the steering reward $\forall\ t \in [T]$, $R^t(\mu) = \mu^* - \mu + \mathbf{1}\|\mu^* - \mu\|_\infty$, for any $\mu \in \Delta^H_{\mathcal{S}\times\mathcal{A}}$, we have:*

$$\Delta_T(\{\bar{\mu}^t_{M^*}\}_{t=1}^T) \leq L_U \sqrt{HSAT\,\mathrm{AdaReg}(T)} = o(T)$$
$$C_T(\{\bar{\mu}^t_{M^*}, R^t\}_{t=1}^T) \leq 2H\sqrt{T\,\mathrm{AdaReg}(T)} = o(T).$$

The bound above for the known model setting, although may not be tight, can serve as a benchmark for the more challenging unknown model settings. We will see that the bounds in Theorems 5.1 and 6.1 (for the unknown model setting) are not much worse than the bound of Theorem 4.1.

4.2 Steering towards a Target Policy in an Unknown Model

Without the knowledge of transition function $\mathbb{P}^*$, the steering becomes challenging, because we can no longer compute Ψ_{M^*} or identify whether a given density (e.g. $\arg\max_{\mu\in\Delta^H_{\mathcal{S}\times\mathcal{A}}} U(\mu)$) can actually be achieved by the agents. Therefore, we shift our focus to the policy space. Interestingly, we reveal that, it is possible to steer the agents to any target policy $\pi \in \Pi$, even without the knowledge of $\mathbb{P}^*$ or Ψ_{M^*}. Our key observation is the following lemma, which suggests an upper bound to control the difference between the population density and the density regarding the target policy.

Lemma 4.2. *Given any M and target $\pi \in \Pi$, suppose the agents induce population density $\bar{\mu}_M$ in M, then:*

$$\|\bar{\mu}_M - \mu^\pi_M\|_1 \leq H \sum_{h,s} \bar{\mu}_{M,h}(s)\|\bar{\pi}_h(\cdot|s) - \pi_h(\cdot|s)\|_1,$$
$$\text{with } \bar{\mu}_{M,h}(s) := \sum_a \bar{\mu}_{M,h}(s,a) \quad (3)$$

This motivates us to design a steering reward function that penalizes the RHS of Eq. (3), which is actually doable *without the knowledge of model.* Given a policy π, we define matrix $W^\pi \in \mathbb{R}^{SAH\times SAH}$ to be the block diagonal of W^π_{h,s_h} for all $h \in [H]$ and $s_h \in \mathcal{S}$, where

$$W^\pi_{h,s} := \begin{bmatrix} \pi_h(a_1|s) & \dots & \pi_h(a_1|s) \\ \vdots & & \vdots \\ \pi_h(a_A|s) & \dots & \pi_h(a_A|s) \end{bmatrix} \in \mathbb{R}^{A\times A}, \quad (4)$$

Now, consider the steering reward function:

$$\forall \mu \in \Delta^H_{\mathcal{S}\times\mathcal{A}},\ R_\pi(\mu) := -\mu^\top (W^\pi - I)^\top (W^\pi - I). \quad (5)$$

where $I \in \mathbb{R}^{SAH\times SAH}$ is the identity matrix. We can verify that, for any possible population density $\bar{\mu}^t$ occurs at step t, we have

$$\langle R_\pi(\bar{\mu}^t), \mu^\pi - \bar{\mu}^t\rangle = \|(W^\pi - I)\bar{\mu}^t\|_2^2$$
$$= \sum_{h,s,a} (\bar{\mu}^t_h(s))^2 |\pi_h(a|s) - \bar{\pi}^t_h(a|s)|^2. \quad (6)$$

Recall that $\bar{\pi}^t_h$ denotes the policy induced by population density (see definition in Sec. 2). Here in the first equality, we use the fact that, for any μ, $\langle R_\pi(\mu), \mu^\pi\rangle = 0$ since $(W^\pi - I)\mu^\pi = 0$. Eq. (6) above is important in that it connects the one step regret (LHS) with the gap between the population density and target density (RHS through Lemma 4.2).

Combining with Prop. 3.1, if all the agents are no-regret learners, and we steer the agents with the same steering reward R_π for T steps, we should expect $\bar{\mu}^t$ to converge to μ^π, which we summarize to the following theorem. This result provides important insights for our incentive design algorithm in Section 5.

Theorem 4.3. *Let $\pi^* = \arg\max_\pi U(\mu^\pi)$ and $R^t(\mu) = R_{\pi^*}(\mu) + \|R_{\pi^*}(\mu)\|_\infty \mathbf{1}$ for all t. Under Asump. B,*

$$\Delta_T(\{\bar{\mu}^t_{M^*}\}_{t=1}^T) \leq L_U \sqrt{H^3SAT\,\mathrm{AdaReg}(T)} = o(T)$$
$$C_T(\{\bar{\mu}^t_{M^*}, R^t\}_{t=1}^T) \leq 4H\sqrt{T\,\mathrm{AdaReg}(T)} = o(T).$$

5 STEERING WITH NO INTRINSIC REWARD

In this section, we study the **Scenario 1** introduced in Sec. 3.4, where the transition function $\mathbb{P}^*$ is unknown and the original reward r^* is zero, so the steering rewards are the only incentives for the agents. The main challenge in this setting is that, without the knowledge of $\mathbb{P}^*$, we can not determine the feasible density set Ψ_{M^*} and the maximizer of the utility function. Therefore, we have to design a steering strategy to incentivize the agents to explore for the mediator to estimate $\mathbb{P}^*$, while balancing the exploration-exploitation trade-off to ensure sub-linear steering gap and cost.

Our main contribution is an optimism-based exploration algorithm in Alg. 2, which provably addresses the above challenges and achieves our objectives. The algorithm is built based on the techniques we developed in Sec. 4.2, which allows us to steer the agents to any target policy without the knowledge of model. Next, we introduce the key components in algorithm design.

Low Policy Switching Optimistic Exploration Strategy For efficient exploration, we maintain a

Algorithm 2 Steering reward design for Scenario 1

```
 1: Initialize $\mathcal{P}^1 :=$ set of all possible transition functions, $\pi^1_*$ (arbitrarily), $k = 1, T_0 = 0$.
 2: for $t = 1, ..., T$ do
                                         ▷ Recall $R_{\pi^k_*}$ as defined in Eq. (5)
 3:     Compute steering reward function
            $$R^t_{\text{z}}(\cdot) \leftarrow R_{\pi^k_*}(\cdot) + \|R_{\pi^k_*}(\cdot)\|_\infty \mathbf{1}.$$
 4:     Agents play the $t$-th game.
 5:     Obtain trajectory $((s^t_h, a^t_h))^H_{h=1}$.
 6:     if $\exists(h, s, a),\ s.t.\ n_k(h, s, a) \geq N_k(h, s, a)$ then
 7:         Update $\mathcal{P}^{k+1}$ as in (7).
 8:         $T_k \leftarrow t$; $k \leftarrow k + 1$.
 9:         $\pi^k_*, \hat{M}^k \leftarrow \arg\max_{\pi\in\Pi, \hat{M}:\hat{\mathbb{P}}_{\hat{M}}\in\mathcal{P}^k} U(\mu^\pi_{\hat{M}})$.
10:     end if
11: end for
```

confidence set for $\mathbb{P}^*$ denoted by $\mathcal{P}$:

$$\bar{\mathbb{P}}^{k+1}_h(s'|s, a) := \sum_{t=1}^{T_k} \frac{\mathbb{I}\{s^t_h = s, a^t_h = a, s^t_{h+1} = s'\}}{\max\{1, N_{k+1}(h, s, a)\}},$$

$$\mathcal{P}^{k+1} := \Big\{ \hat{\mathbb{P}} : \forall h, s, a. \|\hat{\mathbb{P}}_h(\cdot|s, a) - \bar{\mathbb{P}}^{k+1}_h(\cdot|s, a)\|_1 \leq \varepsilon_{k+1}(h, s, a) \Big\}, \qquad (7)$$

where $\varepsilon_{k+1}(h, s, a) := \sqrt{\frac{2S\ln(THSA/\delta)}{\max\{1, N_{k+1}(h,s,a)\}}}$. We highlight that we only update $\mathcal{P}$ and switch target policy in low frequency, and here we use index $1 \leq k \leq K$ to count the policy switching episodes, to distinguish with the steering steps $t \in [T]$. We use $k(t)$ to denote the index of episode at iteration t and use T_k to denote the iteration number at the end k-th policy switching. We define $n_k(h, s, a)$ to be the number of samples equal to (s, a) at time h in episode k, and $N_k(h, s, a) = \sum_{k'<k} n_{k'}(h, s, a)$. A new episode begins as soon as we have as many samples in this episode as in all the previous ones for some h, s, a, i.e., $n_k(h, s, a) \geq N_k(h, s, a)$. The main motivation for this technique is to avoid the agents' potentially adversarial behaviors. As we will see later in the proof sketch, K will appear in the steering gap upper bound.

For exploration, we select the optimistic policy $\pi^{(\cdot)}_*$ and model $\hat{M}^{(\cdot)}$ (line 10) s.t. the induced density maximizes utility. Then, we choose steering reward R_{z} to guide the agents towards $\pi^{(\cdot)}_*$ and collect data samples to update the model confidence set. Intuitively, either $\pi^{(\cdot)}_*$ indeed maximizes the utility, implying a low steering gap; or the exploration helps to reduce the uncertainty.

Managing the steering gap and cost We have the following guarantees for Alg. 2

Theorem 5.1. *Suppose the intrinsic reward $r^* = 0$, under Assump. B and C, if we run Alg. 2 with $\delta \in (0, 1)$, then with probability at least $1 - 2\delta$, $K \leq HSA\log_2 T$, and*

$$\Delta_T(\{\bar{\mu}^t_{M^*}\}^T_{t=1}) \leq L_U\sqrt{H^3SATK\,\text{AdaReg}(T)} + 36L_U H^3 S\sqrt{\ln(THSA/\delta)AT} = o(T).$$

$$C_T(\{\bar{\mu}^t_{M^*}, R^t_{\text{z}}\}^T_{t=1}) \leq 4H\sqrt{TK\,\text{AdaReg}(T)} = o(T).$$

As a concrete example, agents following Online Gradient Descent with step size $O(1/\sqrt{t})$ (Hazan, 2023) result in $\text{AdaReg}(T) = \tilde{O}(\sqrt{T})$ (ignoring H, S and A), which implies $\tilde{O}(T^{3/4})$ steering gap. Besides, if all the agents are capable enough s.t. for any $t \in [T]$, $\pi^{1,t}, ..., \pi^{N,t}$ are equilibria w.r.t. $r^* + R^t_{\text{z}}$, AdaReg would be constant-level, resulting in a $\tilde{O}(\sqrt{T})$ bound.

Proof Sketch We first analyze the steering gap. Intuitively, Alg. 2 can be interpreted as a "K-stage" version of what we did in Sec. 4.2. In each stage, we pick a target policy, and steer the agents towards it for exploration. Following this intuition, and thanks to the Lipschitz condition (Assump. C) and the optimism in planning, we can decompose the steering gap as follow:

$$\Delta_T(\{\bar{\mu}^t_{M^*}\}^T_{t=1}) \leq L_U(2H + 1)\underbrace{\sum_{t=1}^T \|\mu^{\bar{\pi}^t}_{\hat{M}^{k(t)}} - \bar{\mu}^t_{M^*}\|_1}_{\Delta_{\text{est}}}$$

$$+ L_U H\underbrace{\sum_{t=1}^T \sum_{h,s} \bar{\mu}^t_{M^*,h}(s)\|\pi^{k(t)}_{*,h}(\cdot|s) - \bar{\pi}^t_h(\cdot|s)\|_1}_{\Delta_{\text{pop}}}. \qquad (8)$$

We refer the first term Δ_{est} as model estimation error, which measures the gap between the population density $\bar{\mu}^t$ and the density induced by the population average policy $\bar{\pi}^t$ (see definition in Sec. 2) in the estimated model $\hat{M}^k$. As we collect more and more data, $\hat{M}^k$ gets closer to M^*, and we can show Δ_{est} only grows sublinearly. The second term Δ_{pop} can be interpreted as the population convergence error, which is determined by how fast the agents converge to the target policy we steer them to. Following the similar techniques in the proof of Thm. 4.3, Δ_{pop} can be upper bounded by:

$$\sqrt{HSAT\underbrace{\sum_{t=1}^T \langle R_{\pi^{k(t)}_*}(\bar{\mu}^t_{M^*}), \mu^{\pi^{k(t)}_*}_{M^*} - \bar{\mu}^t_{M^*}\rangle}_{\texttt{AgentReg}}}. \qquad (9)$$

Here we use `AgentReg` to refer the summation term, which can be interpreted as the agents' dynamic regret if choosing $\mu^{\pi^{k(t)}_*}_{M^*}$ as the comparators. Thanks to the low policy switching, `AgentReg` can be controlled by $O(K\,\text{AdaReg}(T))$, and the only remaining step is to control K. Note that we only switch policy when the number of visitation of some state-action pair got doubled, therefore, K only grows in $O(\log(T))$.

For the steering cost, we can calculate that

$$C(\bar{\mu}^t_{M^*}, R^t_{\text{z}}) \leq 2H \|R_{\pi_*^{k(t)}}(\bar{\mu}^t_{M^*})\|_\infty,$$

and for any π, μ, $\|R_\pi(\mu)\|_\infty \leq 2\|(W^\pi - I)\mu\|_2$ which, by Eq. (6), is equal to $2\sqrt{\langle R_\pi(\mu), \mu - \mu^\pi\rangle}$. Using Jensen's inequality and Assump. B, we derive the final bound.

6 STEERING WITH NON-ZERO INTRINSIC REWARD

Next, we turn to **Scenario 2** in Sec. 3.4, the complete setting where the agents' pre-existing reward function $r^* \in [0, r_{\max}]$ is both non-zero and unknown. The non-zero intrinsic reward introduces non-trivial additional challenges. Firstly, it changes the steering landscape and introduces some prior bias for our steering reward design. Secondly, since it is unknown, we must account for its interference on the steering dynamics and undertake strategic exploration to estimate r^*. In the following, we explain how we overcome these challenges by a pessimism-based reward estimation strategy.

Confidence set for r^* We recall our setup in Sec. 3.1: the mediator can observe the population density $\bar{\mu}^t$ and noisy reward $r^t = (r^*_h(s^t_h, a^t_h, \bar{\mu}^t_{M^*,h}) + \xi_h)_{h\in[H]}$ perturbed by i.i.d. zero-mean σ-sub-Gaussian noise ξ. We will use this information to estimate the original reward. At each iteration t, we maintain a confidence set $\hat{\mathcal{R}}^t$ for r^*, defined by:

$$\hat{\mathcal{R}}^t := \left\{\hat{r} \in \mathcal{R} : \|\hat{r} - \bar{r}^t\|_{2,E_t} \leq \sqrt{\beta_t}\right\},$$
$$\bar{r}^t := \arg\min_{\hat{r}\in\mathcal{R}} \sum_{i=1}^{t-1}\sum_{h=1}^{H} \left(\hat{r}_h(s^i_h, a^i_h, \bar{\mu}^i_{M^*,h}) - r^i_h\right)^2, \quad (10)$$

where $\|g\|^2_{2,E_t} := \sum_{i=1}^{t-1}\sum_{h=1}^{H}(g_h(s^i_h, a^i_h, \bar{\mu}^i_{M^*,h}))^2$ for any function g as a short note. We use β_t to denote confidence interval length to ensure r^* is contained in the confidence set at any time with high probability. We defer a detailed choice of β_t to Lem. I.1. Informally, $\beta_t = O(\sigma^2 \log N(\mathcal{R}, \frac{1}{T}))$ grows in $\log T$, where $N(\mathcal{R}, \varepsilon)$ is the ε-covering number of $\mathcal{R}$.

Steering Reward Design with Pessimism We consider the following steering reward design

$$\forall \mu \in \Delta^H_{\mathcal{S}\times\mathcal{A}}, R^t_{\text{nz}}(\mu) := R_{\pi_*^{k(t)}}(\mu) - (\bar{r}^t(\mu) - w_{\hat{\mathcal{R}}^t}(\mu)) + (r_{\max} + \|R_{\pi_*^{k(t)}}(\mu)\|_\infty)\mathbf{1} \quad (11)$$

Here $\pi_*^{k(t)}$ is computed in the same way as Alg, 2; $\bar{r}^t \in \hat{\mathcal{R}}^t$ (defined in Eq. (10)) is the reward estimation achieving the minimal empirical loss; $w_{\hat{\mathcal{R}}^t}(\mu)$ is a vector with elements $(w_{\hat{\mathcal{R}}^t}(\mu))_{h,s,a} := \sup_{r,\tilde{r}\in\hat{\mathcal{R}}^t} |r_h(s,a,\mu) - \tilde{r}_h(s,a,\mu)|$, which quantifies the estimation uncertainty for each state-action pair; the last constant shift term ensures non-negativity.

As we can see, the main difference compared with steering reward R^t_{z} in Alg. 2 is that we include an additional reward estimation term to offset the effect by the non-zero original reward r^*. In this way, the agents will follow the guidance by $R_{\pi_*^{k(t)}}$ to explore as we want. Note that here we conduct a *pessimism-based* reward estimation such that $\bar{r}^t - w_{\hat{\mathcal{R}}^t} \leq r^*$ for some technical reason, which we will explain later.

Steering Algorithm Design The algorithm design for the non-zero intrinsic reward setting only differs from Alg. 2 in the additional update of $\hat{\mathcal{R}}^t$ as in Eq. (10) and choosing Eq. (11) as the steering reward R^t_{nz}. For completeness, we defer the detailed algorithm to Alg. 4 in Appx. I.1. We have the following guarantees for steering gap and steering cost.

Theorem 6.1. *Under Assump. A, B and C, if we run Alg. 4 with* $0 < \delta < 1$*, then with probability at least* $1 - 6\delta$*,* $K \leq HSA\log_2 T$*, and*

$$\Delta_T(\{\bar{\mu}^t_{M^*}\}_{t=1}^T) \leq L_U\sqrt{H^3SAT(K\,\text{AdaReg}(T) + D)} + 36L_U H^3 S\sqrt{AT\ln(THSA/\delta)},$$
$$C_T(\{\bar{\mu}^t_{M^*}, R^t_{nz} - (r_{\max}\cdot\mathbf{1} - r^*)\}_{t=1}^T) = 4H\sqrt{T(K\,\text{AdaReg}(T) + D)} + D,$$

where $D = \tilde{O}(\sqrt{\beta_T H \dim_E(\mathcal{R}, T^{-1})T}))$.

Comparing with Theorem 5.1, we can find both the steering gap and cost only differ in the additional term D, which results from the estimation error of r^*. The term D depends on the Eluder dimension of $\mathcal{R}$ and β_T. In Appx. H.1, we show several common function classes with $\dim_E(\mathcal{R}, T^{-1}) \in \tilde{\mathcal{O}}(1)$, and where by choosing β_T appropriately, we have $D \in \tilde{\mathcal{O}}(\sqrt{T})$. As a result, both the steering gap and cost upper bounds in Thm. 6.1 will be sub-linear in T.

Proof Sketch Similar to the proof for Thm. 5.1, we can decompose the steering gap as Eq. (8), and upper bound model estimation error term Δ_{est} in the same way. The proof diverges when we upper bound `AgentReg` in Eq. (9), because the agents' no-regret behavior holds for $r^* + R^t_{\text{nz}}$ in this setting. We can write

$$\texttt{AgentReg} = \sum_{t=1}^T \langle R^t_{\text{nz}}(\bar{\mu}^t_{M^*}) + r^*(\bar{\mu}^t_{M^*}) - r^*(\bar{\mu}^t_{M^*}) + \bar{r}^t(\bar{\mu}^t_{M^*}) - w_{\hat{\mathcal{R}}^t}(\bar{\mu}^t_{M^*}), \mu^{\pi_*^{k(t)}}_{M^*} - \bar{\mu}^t_{M^*}\rangle.$$

Using pessimism, i.e., $r^* \geq \bar{r}^t - w_{\hat{\mathcal{R}}^t}$, we can bound this by

$$\sum_{t=1}^{T} \langle R^t_{\mathrm{nz}}(\bar{\mu}^t_{M^*}) + r^*(\bar{\mu}^t_{M^*}), \mu_{M^*}^{\pi_*^{k(t)}} - \bar{\mu}^t_{M^*} \rangle \\ + \sum_{t=1}^{T} \langle r^*(\bar{\mu}^t_{M^*}) - \bar{r}^t(\bar{\mu}^t_{M^*}) + w_{\hat{\mathcal{R}}^t}(\bar{\mu}^t_{M^*}), \bar{\mu}^t_{M^*} \rangle.$$

Clearly, the first term above is just agents' dynamic regret regarding the total reward they received and can be bounded again by $K\,\mathrm{AdaReg}(T)$. The second term above can be further controlled by $\mathcal{O}(\sum_t \langle w_{\hat{\mathcal{R}}^t}(\bar{\mu}^t), \bar{\mu}^t \rangle)$, which is basically the accumulative confidence interval length for reward estimation and its growth can be controlled by Eluder dimension (Lem. H.6) and is only sub-linear in T.

For the steering cost, we can provide an upper bound involving `AgentReg` and reward estimation error that we analyzed before. To save space, we do not repeat it here and refer the reader to Appx. I for the full proof.

Remark 6.2. Our strategy to deal with the intrinsic reward r^* is to try to "cancel" it with our steering reward. This approach is justified by the fact that we keep r^* and U very general, which means that the target density to maximize U may not coincide with an equilibrium associated with the original reward r^*. Therefore, to ensure the target density is still a stationary point for no-regret learners, we treat r^* as a competing force to offset. We admit that there might be other options to counteract the impact of r^* with lower steering costs, and we leave further investigation to the future work.

Remark 6.3 (Generalization to Unknown Utility Setting)**.** Although this paper focuses on the case when U is revealed to the mediator, it is possible to generalize our results to the case where the utility function U is unknown, but it lies in a known function class $\mathcal{U}$ with bounded Eluder dimension. In Appx. J, we formalize this setting and present a solution to address this case based on a simple modification of the current methods. Our established regret bound for steering gap and steering cost grow at a rate of $\tilde{\mathcal{O}}(T^{5/6})$. Although the results are worse than the rate of $\tilde{\mathcal{O}}(T^{3/4})$ in Thm. 6.1 due to the challenges in exploring the utility function, they are still sub-linear in T.

7 CONCLUSION

We study a novel problem setting for incentive design in unknown mean-field games with no-regret agents. Our optimistic algorithm introduces newly developed steering reward designs, achieving sublinear utility regret and steering costs when the intrinsic reward is zero. Extending to the setting with a non-zero and unknown intrinsic reward function, we adapted our algorithm to handle this new challenge, maintaining sublinear utility regret and vanishing steering costs competing with a baseline strategy. Future work could explore the more challenging case where the transition function is also dependent on the population density. Another interesting direction is to identify better or even optimal steering reward design to stabilize the target policy and design an algorithm with sub-linear guarantees comparing with that benchmark.

Acknowledgements

This work is supported by Swiss National Science Foundation (SNSF) Project Funding No. 200021-207343 and SNSF Starting Grant.

References

Achdou, Y. and Lasry, J.-M. (2019). Mean Field Games for Modeling Crowd Motion. In Chetverushkin, B. N., Fitzgibbon, W., Kuznetsov, Y., Neittaanmäki, P., Periaux, J., and Pironneau, O., editors, *Contributions to Partial Differential Equations and Applications*, pages 17–42. Springer International Publishing, Cham.

Baumann, T., Graepel, T., and Shawe-Taylor, J. (2020). Adaptive mechanism design: Learning to promote cooperation. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

Brown, W., Schneider, J., and Vodrahalli, K. (2024). Is learning in games good for the learners? *Advances in Neural Information Processing Systems*, 36.

Cabannes, T., Lauriere, M., Perolat, J., Marinier, R., Girgin, S., Perrin, S., Pietquin, O., Bayen, A. M., Goubault, E., and Elie, R. (2021). Solving N-player dynamic routing games with congestion: a mean field approach. arXiv:2110.11943 [cs, eess, math].

Camara, M., Hartline, J., and Johnsen, A. (2020). Mechanisms for a No-Regret Agent: Beyond the Common Prior. arXiv:2009.05518 [cs, econ].

Canyakmaz, I., Sakos, I., Lin, W., Varvitsiotis, A., and Piliouras, G. (2024). Steering game dynamics towards desired outcomes. arXiv:2404.01066 [cs, eess].

Carmona, R. and Wang, P. (2021). Finite-state contract theory with a principal and a field of agents. *Management Science*, 67(8):4725–4741.

Castiglioni, M., Marchesi, A., and Gatti, N. (2023). Multi-agent contract design: How to commission multiple agents with individual outcomes. In *Proceedings of the 24th ACM Conference on Economics and Computation*, pages 412–448.

Chen, B. and Cheng, H. H. (2010). A review of the applications of agent technology in traffic and transportation systems. *IEEE Transactions on Intelligent Transportation Systems*, 11(2):485–497.

Curry, M., Thoma, V., Chakrabarti, D., McAleer, S., Kroer, C., Sandholm, T., He, N., and Seuken, S. (2024). Automated design of affine maximizer mechanisms in dynamic settings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(9):9626–9635.

DellaVigna, S. and Malmendier, U. (2004). Contract design and self-control: Theory and evidence. *The Quarterly Journal of Economics*, 119(2):353–402.

Deng, Y., Schneider, J., and Sivan, B. (2019). Strategizing against No-regret Learners. arXiv:1909.13861 [cs].

Dinneweth, J., Boubezoul, A., Mandiau, R., and Espié, S. (2022). Multi-agent reinforcement learning for autonomous vehicles: a survey. *Autonomous Intelligent Systems*, 2(1):27.

Dütting, P., Ezra, T., Feldman, M., and Kesselheim, T. (2023). Multi-agent contracts. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 1311–1324.

Ehtamo, H., Kitti, M., and Hämäläinen, R. P. (2002). Recent studies on incentive design problems in game theory and management science. In *Optimal Control and Differential Games: Essays in Honor of Steffen Jørgensen*, pages 121–134. Springer.

Elie, R., Mastrolia, T., and Possamaï, D. (2019). A tale of a principal and many, many agents. *Mathematics of Operations Research*, 44(2):440–467.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

Fu, G. and Horst, U. (2018). Mean-field leader-follower games with terminal state constraint.

Ge, J., Wang, Y., Li, W., and Jin, C. (2024). Towards principled superhuman ai for multiplayer symmetric games.

Gomes, D. A., Velho, R. M., and Wolfram, M.-T. (2014). Socio-economic applications of finite state mean field games. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 372(2028):20130405. arXiv:1403.4217 [math].

Guo, X., Hu, A., Xu, R., and Zhang, J. (2021). Learning Mean-Field Games. arXiv:1901.09585 [math].

Guo, X., Li, L., Nabi, S., Salhab, R., and Zhang, J. (2023a). MESOB: Balancing Equilibria & Social Optimality. arXiv:2307.07911 [cs, math].

Guo, X., Li, L., Nabi, S., Salhab, R., and Zhang, J. (2023b). Mesob: Balancing equilibria & social optimality.

Hazan, E. (2023). Introduction to Online Convex Optimization. arXiv:1909.05207 [cs, math, stat].

Hazan, E. and Seshadhri, C. (2007). Adaptive algorithms for online decision problems. *Electronic Colloquium on Computational Complexity (ECCC)*, 14.

Ho, C.-J., Slivkins, A., and Vaughan, J. W. (2014). Adaptive contract design for crowdsourcing markets: Bandit algorithms for repeated principal-agent problems. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 359–376.

Holmström, B. (1979). Moral hazard and observability. *The Bell journal of economics*, pages 74–91.

Hu, A. and Zhang, J. (2024). MF-OML: Online Mean-Field Reinforcement Learning with Occupation Measures for Large Population Games. arXiv:2405.00282 [cs, math].

Huang, J., He, N., and Krause, A. (2024a). Model-Based RL for Mean-Field Games is not Statistically Harder than Single-Agent RL. arXiv:2402.05724 [cs, stat].

Huang, J., Thoma, V., Shen, Z., Nax, H. H., and He, N. (2024b). Learning to Steer Markovian Agents under Model Uncertainty. arXiv:2407.10207 [cs, stat].

Huang, J., Yardim, B., and He, N. (2023). On the Statistical Efficiency of Mean Field Reinforcement Learning with General Function Approximation. arXiv:2305.11283 [cs, stat].

Huang, M., Malhamé, R. P., and Caines, P. E. (2006). Large population stochastic dynamic games: closed-loop mckean-vlasov systems and the nash certainty equivalence principle.

Innes, R. D. (1990). Limited liability and incentive contracting with ex-ante action choices. *Journal of economic theory*, 52(1):45–67.

Iyer, K., Johari, R., and Sundararajan, M. (2014). Mean field equilibria of dynamic auctions with learning. *Management Science*, 60(12):2949–2970.

Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600.

Lasry, J.-M. and Lions, P.-L. (2007). Mean field games. *Japanese journal of mathematics*, 2(1):229–260.

Laurière, M., Perrin, S., Pérolat, J., Girgin, S., Muller, P., Élie, R., Geist, M., and Pietquin, O. (2024). Learning in Mean Field Games: A Survey. arXiv:2205.12944 [cs, math].

Liu, B., Li, J., Yang, Z., Wai, H.-T., Hong, M., Nie, Y. M., and Wang, Z. (2022). Inducing Equilibria via Incentives: Simultaneous Design-and-Play Ensures Global Convergence. arXiv:2110.01212 [cs].

Luo, Z.-Q., Pang, J.-S., and Ralph, D. (1996). *Mathematical Programs with Equilibrium Constraints*. Cambridge University Press.

Osband, I. and Roy, B. V. (2014). Model-based reinforcement learning and the eluder dimension.

Perolat, J., Perrin, S., Elie, R., Laurière, M., Piliouras, G., Geist, M., Tuyls, K., and Pietquin, O. (2021). Scaling up Mean Field Games with Online Mirror Descent. arXiv:2103.00623 [cs].

Ratliff, L. J., Dong, R., Sekar, S., and Fiez, T. (2019). A perspective on incentive design: Challenges and opportunities. *Annual Review of Control, Robotics, and Autonomous Systems*, 2(1):305–338.

Rosenberg, A. and Mansour, Y. (2019). Online convex optimization in adversarial markov decision processes.

Roughgarden, T. and Tardos, É. (2007). Introduction to the inefficiency of equilibria. In Nisan, N., Roughgarden, T., Tardos, E., and Vazirani, V. V., editors, *Algorithmic Game Theory*, pages 443–460. Cambridge University Press, Cambridge.

Russo, D. and Van Roy, B. (2013). Eluder dimension and the sample complexity of optimistic exploration. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Sanjari, S., Bose, S., and Başar, T. (2024). Incentive Designs for Stackelberg Games with a Large Number of Followers and their Mean-Field Limits. arXiv:2207.10611 [cs].

Scheid, A., Tiapkin, D., Boursier, E., Capitaine, A., Mhamdi, E. M. E., Moulines, É., Jordan, M. I., and Durmus, A. (2024). Incentivized learning in principal-agent bandit games. *arXiv preprint arXiv:2403.03811*.

Steinbacher, M., Raddant, M., Karimi, F., Camacho Cuena, E., Alfarano, S., Iori, G., and Lux, T. (2021). Advances in the agent-based modeling of economic and social behavior. *SN Business & Economics*, 1(7):99.

Subramanian, S. G., Taylor, M. E., Crowley, M., and Poupart, P. (2022). Decentralized mean field games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9439–9447.

Wang, K., Xu, L., Perrault, A., Reiter, M. K., and Tambe, M. (2022). Coordinating followers to reach better equilibria: End-to-end gradient descent for stackelberg games. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(5):5219–5227.

Weissman, T., Ordentlich, E., Seroussi, G., Verdú, S., and Weinberger, M. J. (2003). Inequalities for the l1 deviation of the empirical distribution.

Yang, J., Wang, E., Trivedi, R., Zhao, T., and Zha, H. (2022). Adaptive incentive design with multi-agent meta-gradient reinforcement learning. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, page 1436–1445, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Yardim, B., Cayci, S., Geist, M., and He, N. (2022). Policy Mirror Ascent for Efficient and Independent Learning in Mean Field Games.

Zhang, B. H., Farina, G., Anagnostides, I., Cacciamani, F., McAleer, S. M., Haupt, A. A., Celli, A., Gatti, N., Conitzer, V., and Sandholm, T. (2024). Steering No-Regret Learners to a Desired Equilibrium. arXiv:2306.05221 [cs].

Zhu, B., Bates, S., Yang, Z., Wang, Y., Jiao, J., and Jordan, M. I. (2022). The sample complexity of online contract design. *arXiv preprint arXiv:2211.05732*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Contents

A TABLE OF FREQUENTLY USED NOTATIONS

Notation	**Description**
$[n]$	$\{1, 2, ..., n\}$ for any $n \in \mathbb{N}$
$\Delta_{\mathcal{X}}$	Set of probability distributions over a finite set $\mathcal{X}$
$\mathbb{I}\{\mathcal{E}\}$	Indicator function for the event $\mathcal{E}$
$\mathbf{1}$	All-one vector
$\mathbf{e}_i$	The i-th standard-basis vector
$M = (N, \mathcal{S}, \mathcal{A}, H, \mathbb{P}_M, r_M, \mu_1)$	The model / game
N	Number of agents
$\mathcal{S}, \mathcal{A}$	State and action space
H	Horizon length of the game
μ_1	Initial state distribution
$\{\mathbb{P}_{M,h} : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}\}_{h\in[H]}$	Transition function
$\{r_{M,h} : \mathcal{S} \times \mathcal{A} \times \Delta_{\mathcal{S}\times\mathcal{A}} \to [0, r_{\max}]\}_{h\in[H]}$	Reward function
$\{R_h : \mathcal{S} \times \mathcal{A} \times \Delta_{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}\}_{h\in[H]}$	Steering reward function (capitalized)
$r : \Delta^H_{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{HSA}$	Vectorized reward function $(r(\mu))_{h,s,a} = r_h(s, a, \mu_h)$
$\{\pi_h : \mathcal{S} \to \Delta_{\mathcal{A}}\}_{h\in[H]}$	Markov policy
Π	Set of all policies
μ^π_M	State-action density of policy π in model M
Ψ_M	Set of possible state-action densities in model M
$\text{AdaReg}(T)$	Adaptive regret bound after T games
$U : \Delta^H_{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}$	Utility function
$C(\bar\mu^t, R^t) = \langle R^t(\bar\mu^t), \bar\mu^t\rangle$	Steering cost function
R_π	Reward function which incentivizes policy π
$M^*, r^*, \mathbb{P}^*$	True model, intrinsic reward, transition function
R_{z}	Steering reward for the setting where $r^* = 0$. “z” in sub-scription as a short note of “zero”.
R_{nz}	Steering reward for the setting where $r^* \in \mathcal{R}$. “nz” in sub-scription as a short note of “zero”
$\dim_E(\mathcal{F}, \varepsilon)$	Eluder dimension of function class $\mathcal{F}$
$\bar\mu$	Population density $\bar\mu := \frac{1}{N}\sum_n \mu^{\pi^n}$
$\bar\pi$	Population average policy induced by $\bar\mu$
$\mathcal{O}, \tilde{\mathcal{O}}$	Standard big-O notations

B SUMMARY OF MAIN RESULTS

In the following, we summarize the main theorems in this paper under Assump. A, B and C. We study the steering gaps and costs of four settings. The settings are categorized depending on whether M^* (or $\pi^* := \arg\max_{\pi\in\Pi} U(\mu^\pi_{M^*})$) is known or not, and whether the intrinsic reward function r^* is zero or non-zero and unknown.

Setting	$r^* = 0$?	**Steering Gap**	**Steering Cost**	**Thm.**
Known M^*	✓	$\mathcal{O}(L_U\sqrt{HSAT\,\text{AdaReg}(T)})$	$\mathcal{O}(H\sqrt{T\,\text{AdaReg}(T)})$	4.1
Unknown M^* (known π^*)	✓	$\mathcal{O}(L_U\sqrt{H^3SAT\,\text{AdaReg}(T)})$	$\mathcal{O}(H\sqrt{T\,\text{AdaReg}(T)})$	4.3
Unknown M^*	✓	$\mathcal{O}(L_U\sqrt{H^3SATK\,\text{AdaReg}(T)}$ $+ L_UH^3S\sqrt{AT\ln(THSA/\delta)})$	$\mathcal{O}(H\sqrt{TK\,\text{AdaReg}(T)})$	5.1
Unknown M^*	✗	$\mathcal{O}(L_U\sqrt{H^3SAT(K\,\text{AdaReg}(T) + D)}$ $+ L_UH^3S\sqrt{AT\ln(THSA/\delta)})$	$\mathcal{O}(H\sqrt{T(K\,\text{AdaReg}(T) + D)})$ $+D + C_T(\{\bar\mu^t, r_{\max}\cdot\mathbf{1} - r^*\}_{t=1}^T)$	6.1

Here $K = \mathcal{O}(HSA\log T)$ and $D = \tilde{\mathcal{O}}(\sqrt{\beta_T H \dim_E(\mathcal{R}, T^{-1})T}))$, where $\dim_E$ is the eluder dimension of reward function class $\mathcal{R}$, and $\beta_T = \tilde{\mathcal{O}}(1)$.

C OTHER RELATED WORKS

More Elaboration on Comparison between the Steering Setting and Contract Design Setting The steering setup differs from previous incentive design literature in two aspects: (1) it deals with "learning agents" continuously updating their policies and (2) it cares about the steering gap towards a target policy and the accumulative steering cost. One of the most related and representative existing problem setups is contract design (a.k.a. the principal-agent problem), which is a classical problem dating back to the seminal work (Holmström, 1979) in 1979. As we discussed in Sec. 1.1, it considers a similar mediator-agents interaction procedure. In the following, we elaborate more on the comparison between those two settings to support our steering setting.

(1) Contract design assumes the agents respond optimally to the mediator/principal (e.g. maximize the total return including the incentives by mediator), which is a quite strong assumption and "simplifies" the problem by making the agents' behaviors predictable.

    In contrast, the steering framework treats the agents' behavior as a dynamic process. For example, Zhang et al. (2024) and ours consider no-regret behaviors, and (Huang et al., 2024b; Canyakmaz et al., 2024) assumes Markovian learning dynamics. Such a non-stationarity is more reasonable in practice and introduces additional challenges in achieving low the steering gap and cost.

(2) Contract design considers a more challenging objective, and targets at finding the optimal incentive design to maximize the mediator's gain deducted by the incentivizing cost. Usually, it also assumes the agents' behaviors are unobservable. Due to such challenges, most of the contract design literature focuses on single-agent setting and assumes the knowledge of the model.

    On the other hand, the steering setting considers steering the agents to some target policies maximizing some utility function, which makes the framework more general. Besides, we do not pursue the optimality in steering cost but sub-linearity would be enough. This is reasonable because in many scenarios we only have budget constraints but do not have to achieve the optimum. Such a relaxation also makes the problem more tractable.

Mean-field game The mean-field game (MFG) is an important framework to model systems with a large number of symmetric agents (Laurière et al., 2024). Most works in the context of MFGs focus on learning equilibrium policies. As the pioneers, Lasry and Lions (2007) and Huang et al. (2006) reveal that learning Nash Equilibrium (NE) is computationally efficient under monotonicity conditions if the model is known in advance. Without the knowledge of the true model, many previous works contribute sample-efficient model-free (Guo et al., 2021; Yardim et al., 2022; Perolat et al., 2021) and model-based (Huang et al., 2023, 2024a) methods to compute NE. Our mean-field game definition is similar to the general MFG setting (Guo et al., 2021), but unlike them, we assume transitions are density-independent and allow independence of agents' policies. This density-independent transition assumption has been frequently considered in previous works (Lasry and Lions, 2007; Huang et al., 2006; Hu and Zhang, 2024; Perolat et al., 2021). To our knowledge, we are the first to investigate steering agents' behaviors in the context of the mean-field game.

Mathematical Programming with Equilibrium Constraints (MPEC) and Mechanism Design MPEC considers a bilevel optimization formulation, where the upper level can be utility maximization problem and the lower level involves equilibrium constraints (Luo et al., 1996). There is a line of research works (Liu et al., 2022; Wang et al., 2022; Yang et al., 2022) consider gradient-based approaches to solve MPEC problems. They usually require strong assumptions on computing hyper-gradients, which may fail to be satisfied in most games. In contrast, we do not involve those assumptions or restrict the target policies are equilibria. We only assume the agents are no-regret learners and do not require them to solve the equilibria induced by modified reward functions.

Another related field within game theory is Mechanism Design, which focuses on designing rules or systems (mechanisms) to achieve a specific objective, especially when participants (agents) have private information and act according to their own interests. Most recent works consider mechanism design on Markov Games (Curry et al., 2024; Baumann et al., 2020).

Guo et al. (2023b) consider a bi-level optimization framework and another bi-objective variant, where the goal of the social planner is to solve an equilibrium policy maximizing some social welfare function. They do not consider the usage of steering reward to intervene agents, and focus on the optimization side without considering model uncertainty. In contrast, we study the incentive design problem, and focus on how to explore and design appropriate steering rewards to guide agents' behaviors without knowledge of the model.

D REGARDING NO-ADAPTIVE REGRET ASSUMPTION

D.1 Proof Of Proposition 3.1

Proposition 3.1 (No-Adaptive-Regret Population Behavior)**.** *Under Assump. B, we have:*

$$\max_{1\leq a<b\leq T, \mu\in\Psi_{M^*}} \sum_{t=a}^{b} \langle r^*(\bar{\mu}^t) + R^t(\bar{\mu}^t), \mu - \bar{\mu}^t\rangle$$
$$\leq \mathrm{AdaReg}(T) = (r_{\max} + R_{\max})\cdot o(T).$$

Proof. We have

$$\sup_{1\leq a<b\leq T} \max_{\mu\in\Psi_{M^*}} \sum_{t=a}^{b} \langle r^*(\bar{\mu}^t) + R^t(\bar{\mu}^t), \mu - \bar{\mu}^t\rangle = \sup_{1\leq a<b\leq T} \max_{\mu\in\Psi_{M^*}} \sum_{t=a}^{b} \langle r^*(\bar{\mu}^t) + R^t(\bar{\mu}^t), \mu - \frac{1}{N}\sum_{n=1}^{N} \mu^{\pi^{n,t}}\rangle$$
$$\leq \frac{1}{N}\sum_{n=1}^{N} \sup_{1\leq a<b\leq T} \max_{\mu\in\Psi_{M^*}} \sum_{t=a}^{b} \langle r^*(\bar{\mu}^t) + R^t(\bar{\mu}^t), \mu - \mu^{\pi^{n,t}}\rangle \leq \frac{1}{N}\sum_{n=1}^{N} \mathrm{AdaReg}(T) = \mathrm{AdaReg}(T),$$

where we used Assumption B in the third step. □

D.2 Concrete Examples Satisfying No-Adaptive Regret Assumption

In this section, we provide some concrete agents learning dynamics examples to support our arguments on the practicality of Assump. B.

Example 1: Colluded Agents with Full Observation of R^t If the agents are able to observe the mediator's steering strategy R^t and R^t is Lipschitz in density (which is indeed satisfied by our proposed algorithms), the agents can collude together and take a (approximate) Nash Equilibrium policy induced by the reward function $r^* + R^t$, which is guaranteed to be exist given the Lipschitz condition (Huang et al., 2023). By the definition of Nash, each agent will have non-positive adaptive regret, which satisfies Assump. B.

Note that in the contract design literature, it is usually assumed the agents are able to do best response (Ho et al., 2014; Zhu et al., 2022) to the principal's (mediator's) strategy if there is only one agent, or take the equilibrium policies for many agents setting (Carmona and Wang, 2021; Elie et al., 2019). Based on the discussion above, those assumptions are strictly stronger than and implies our no-adaptive-regret assumption.

Example 2: Independent Agents Conducting Online Convex Learning In this second example, we consider less powerful agents who can not observe the entire R^t or coordinate with the other agents. Note that from an agent's perspective, the interaction protocol in Procedure 1 can be interpreted as an online linear optimization task, as in Procedure 3.

Procedure 3 Agent-adversary interaction

```
for $t = 1, ..., T$ do
    Agent chooses $x_t \in \mathcal{X}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ is a convex set in Euclidean space.
    Adversary chooses a reward vector $r_t \in \mathbb{R}^d$, possibly based on the history and $x_t$.
    Agent observes $r_t$ and obtains reward $\langle r_t, x_t\rangle$.
end for
```

In our setting, in each iteration $t \in [T]$, the agents pick a density (by picking a policy) from the convex set Ψ_{M^*} and receive potentially adversarial feedback $R^t(\bar{\mu}^t)$ (or $\langle R^t(\bar{\mu}^t), \mu^{\pi^{n,t}}\rangle$ in bandit feedback setting). Then, Assump. B coincides with the standard no-adaptive regret guarantees in online convex optimization setting.

Therefore, Assump. B can be realized if each agent independently adopts any no-adaptive regret online learning algorithm (Hazan and Seshadhri, 2007; Hazan, 2023).

As a concrete algorithm choice, online gradient descent (OGD) achieves a external regret bound of $\frac{3}{2}GD\sqrt{T}$ (Hazan, 2023), where $D \leq 2H$ is the diameter of $\mathcal{X} = \Psi_{M^*}$ and G an upper bound on $\|r_t\|_2 \leq \sqrt{d}\|r_t\|_\infty$. In our case, we can bound $G \leq \sqrt{HSA}(r_{\max} + R_{\max})$. A bound for $r_{\max} + R_{\max}$ is discussed in Appendix D.4. Moreover, in the full feedback setting (the agents know the model M and are able to observe $R^t(\bar{\mu}^t)$), the no-adaptive-regret assumption is not much stronger than no-external-regret, as is demonstrated by the following proposition.

Proposition D.1 (Theorem 1.3 of Hazan and Seshadhri (2007))**.** *Let* $(r_t)_{t=1}^T$ *be reward vectors in* $[0, C]^d$*. Any algorithm following Protocol 3 with external regret* $\mathrm{Reg}(T)$ *can be utilized to build an algorithm with adaptive regret at most* $\mathrm{Reg}(T) + \mathcal{O}(C\sqrt{T \log T})$*.*

Thus, Assump. B can be satisfied with an adaptive regret bound of $\tilde{\mathcal{O}}(\sqrt{T})$ if all the agents follow OGD, modified as in Prop. D.1.

D.3 Motivating Adaptive Regret

Here, we show a small example that should motivate why we need the no-adaptive-regret assumption instead of no-external-regret. External regret is one of the most common regret types, and it is the same as adaptive regret in Assumption B, but $a = 1, b = T$ are fixed. If we want to steer the agents in different directions, the no-external-regret assumption might not be enough, as we can see in the following example.

Consider the stateless setting with $|\mathcal{A}| = 2$, where the incentive designer deploys $R(\mu) = \mathbf{e}_1$ for the first $T/2$ iterations and $R(\mu) = \mathbf{e}_2$ for the remaining $T/2$ iterations.

Suppose all the agents perform the Hedge algorithm, where

$$\bar{\mu}^t(a) = \bar{\pi}^t(a) = \frac{1}{Z_t} \exp\left(1 + \eta \sum_{s=1}^{t-1} \langle R^s(\bar{\mu}^s), \mathbf{e}_a \rangle\right),$$

and Z_t is the normalizing constant. This algorithm is known to have sublinear external regret (Freund and Schapire, 1997). The population density at iteration $t \geq T/2$ is

$$\bar{\mu}^t = \frac{1}{Z_t} \begin{pmatrix} \exp(1 + \eta T/2) \\ \exp(1 + \eta(t - T/2)) \end{pmatrix},$$

while the optimal action is $\mathbf{e}_2$. Thus, over the interval $[T/2 + 1, T]$, the agents accumulate expected regret

$$\sum_{t=T/2+1}^{T} (1 - \bar{\mu}^t(2)) = \sum_{t=T/2+1}^{T} \underbrace{\frac{\exp(1 + \eta T/2)}{\exp(1 + \eta T/2) + \exp(1 + \eta(t - T/2)}}_{\geq 1/2} \geq T/4.$$

So, although this algorithm has no external regret, we still might have to wait $\Omega(T)$ many rounds to let the agents converge to a different density. One can easily observe that with the no-adaptive-regret assumption, this is not an issue.

D.4 Boundedness Of Steering Rewards

As we see in Assumption B, the adaptive regret bound $\mathrm{AdaReg}(T)$ is dependent on $r_{\max} + R_{\max}$. In this section, we show that $R_{\max} = \mathcal{O}(1 + r_{\max})$ for both of our steering rewards R_{z} and R_{nz}.

Proposition D.2. *For any* $\pi \in \Pi$ *and* $\mu \in \Psi$*,* $\|R_\pi(\mu)\|_\infty \leq 2$*, where* R_π *is defined as in Eq.* (5)*.*

Proof. As one can observe using the definition of R_π in Eq. (5) and W^π in Eq. (4), we have for any h, s, a,

$$|(R_\pi(\mu))_{h,s,a}| = \left|(\mu^\top (W^\pi - I)^\top (W^\pi - I))_{h,s,a}\right| = \left|(W^\pi \mu - \mu)^\top (W^\pi - I)_{(h,s,a)}\right|$$

$$= \left|\sum_{a'} (\pi_h(a'|s)\mu_h(s) - \mu_h(s, a'))(\pi_h(a'|s) - \mathbb{I}\{a' = a\})\right| \leq \sum_{a'} \underbrace{|\pi_h(a'|s)\mu_h(s) - \mu_h(s, a')|}_{\leq 1} \cdot |\pi_h(a'|s) - \mathbb{I}\{a' = a\}|$$

$$\leq \sum_{a' \neq a} \pi_h(a'|s) + |\pi_h(a|s) - 1| \leq 2,$$

where $(W^\pi - I)_{(h,s,a)}$ is the (h,s,a)-th column of $W^\pi - I$. □

Proposition D.3. *We have for R_z, as defined in Alg. 2, and for R_{nz}, as defined in Eq.* (11)*, that for all iterations $t \in [T]$,*

$$\|R^t_z(\mu)\|_\infty \leq 4 \quad \textit{and} \quad \|R^t_{nz}(\mu)\|_\infty \leq 2r_{\max} + 4 \quad \forall \mu \in \Delta^H_{\mathcal{S}\times\mathcal{A}}.$$

Proof. We have $\|R^t_{\mathrm{z}}(\mu)\|_\infty = \|R_{\pi_*^{k(t)}}(\mu) + \|R_{\pi_*^{k(t)}}(\mu)\|_\infty \mathbf{1}\|_\infty \leq 2\|R_{\pi_*^{k(t)}}(\mu)\|_\infty$, which is at most 4, by Prop. D.2.

By the definition of $w_{\hat{\mathcal{R}}^t}$, we know that the elements of $w_{\hat{\mathcal{R}}^t}(\mu)$ are bounded in $[0, r_{\max}]$ for any μ. Therefore, and by Prop. D.2,

$$\begin{aligned}
\left\|R^t_{\mathrm{nz}}(\mu)\right\|_\infty &= \left\|R_{\pi_*^{k(t)}}(\mu) - (\bar{r}^t(\mu) - w_{\hat{\mathcal{R}}^t}(\mu)) + (r_{\max} + \|R_{\pi_*^{k(t)}}(\mu)\|_\infty)\mathbf{1}\right\|_\infty \\
&\leq \left\|R_{\pi_*^{k(t)}}(\mu) + \|R_{\pi_*^{k(t)}}(\mu)\|_\infty \mathbf{1}\right\|_\infty + \left\|\bar{r}^t(\mu) - w_{\hat{\mathcal{R}}^t}(\mu)\right\|_\infty + \|r_{\max}\mathbf{1}\|_\infty \\
&\leq 4 + 2r_{\max}.
\end{aligned}$$

□

E STATE-ACTION DENSITY

E.1 Ψ_M Is Convex

Lemma E.1.

$$\Psi_M = \{\mu : \mu \geq 0, \sum_{a'} \mu_{h+1}(s,a') = \sum_{s',a'} \mathbb{P}_{M,h}(s|s',a')\mu_h(s',a') \forall h, s, \sum_{a'} \mu_1(s,a') = \mu_1(s)\}$$

Proof. We abbreviate $\mathbb{P} = \mathbb{P}_M$ and $\mu = \mu_M$, since the model is fixed throughout. For $\mu \in \Psi_M$, it is easy to see that the conditions on the right-hand side are fulfilled. The other direction is more involved. Suppose $\tilde{\mu}$ fulfills $\tilde{\mu} \geq 0$ and for all s, h, $\sum_{a'} \tilde{\mu}_{h+1}(s,a') = \sum_{s',a'} \mathbb{P}_h(s|s',a')\tilde{\mu}_h(s',a')$ as well as $\sum_{a'} \tilde{\mu}_1(s,a') = \mu_1(s)$. Now, define π such that for all s, a, h,

$$\pi_h(a|s) = \begin{cases} \frac{\tilde{\mu}_h(s,a)}{\sum_{a'} \tilde{\mu}_h(s,a')}, & \text{if } \sum_{a'} \tilde{\mu}_h(s,a') \neq 0 \\ 1/A, & \text{else} \end{cases}.$$

Clearly, $\pi_h(a|s) \geq 0$ and $\sum_a \pi_h(a|s) = 1$, which means $\pi \in \Pi$. First of all,

$$\mu_1^\pi(s,a) = \pi_1(a|s)\mu_1(s) = \frac{\tilde{\mu}_1(s,a)}{\sum_{a'} \tilde{\mu}_1(s,a')}\mu_1(s) = \tilde{\mu}_1(s,a).$$

By induction, for all $h \geq 1$ we have if $\sum_{a'} \tilde{\mu}_{h+1}(s,a') \neq 0$,

$$\begin{aligned} \mu_{h+1}^\pi(s,a) &= \sum_{s',a'} \pi_{h+1}(a|s)\mathbb{P}_h(s|s',a')\mu_h^\pi(s',a') \\ &= \pi_{h+1}(a|s) \sum_{s',a'} \mathbb{P}_h(s|s',a')\tilde{\mu}_h(s',a') \\ &= \frac{\tilde{\mu}_{h+1}(s,a)}{\sum_{a'} \tilde{\mu}_{h+1}(s,a')} \sum_{a'} \tilde{\mu}_{h+1}(s,a') = \tilde{\mu}_{h+1}(s,a), \end{aligned}$$

and in case $\sum_{a'} \tilde{\mu}_{h+1}(s,a') = 0$, we know that $\tilde{\mu}_{h+1}(s,a) = 0$ and therefore

$$\begin{aligned} \mu_{h+1}^\pi(s,a) &= \sum_{s',a'} \pi_{h+1}(a|s)\mathbb{P}_h(s|s',a')\mu_h^\pi(s',a') \\ &= \pi_{h+1}(a|s) \sum_{s',a'} \mathbb{P}_h(s|s',a')\tilde{\mu}_h(s',a') \\ &= 1/A \sum_{a'} \tilde{\mu}_{h+1}(s,a') = 0 = \tilde{\mu}_{h+1}(s,a). \end{aligned}$$

We can conclude that $\tilde{\mu} = \mu^\pi$ and thus $\tilde{\mu} \in \Psi_M$. □

Lemma E.2.

$$\Psi_M = \{\mu : \mu \geq 0, B\mu = b\},$$

where

$$B = \begin{pmatrix} D & & & & \\ -\mathbb{P}_{M,1}^\top & D & & & \\ & -\mathbb{P}_{M,2}^\top & D & & \\ & & \cdots & & \\ & & & -\mathbb{P}_{M,H-1}^\top & D \end{pmatrix}, \quad b = \begin{pmatrix} \mu_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

$D := I_S \otimes \mathbf{1}_A^\top$ ($\otimes$ is the tensor product) and $\mathbb{P}_M$ is viewed as a matrix such that $(\mathbb{P}_{M,h})_{(s,a),s'} = \mathbb{P}_{M,h}(s'|s,a)$. An immediate consequence of this formulation is that Ψ_M is convex.

Proof. This result is simply a reformulation of Lemma E.1. We can rewrite the condition $\sum_{a'} \mu_{h+1}(s,a') = \sum_{s',a'} \mathbb{P}_{M,h}(s|s',a')\mu_h(s',a')$ as $D\mu_{h+1} = \mathbb{P}_{M,h}^\top \mu_h$. The condition $\sum_{a'} \mu_1(s,a') = \mu_1(s)$ can be written as $D\mu_1(\cdot,\cdot) = \mu_1$. □

E.2 Inequalities

Lemma E.3. *For any model M and any $\pi, \tilde{\pi} \in \Pi$,*

$$\|\mu_M^\pi - \mu_M^{\tilde{\pi}}\|_1 \leq H \sum_{h,s} \mu_{M,h}^\pi(s) \|\pi_h(\cdot|s) - \tilde{\pi}_h(\cdot|s)\|_1,$$

where $\mu_{M,h}^\pi(s) = \sum_a \mu_{M,h}^\pi(s,a)$.

Proof. Since the model M is fixed throughout, we abbreviate $\mu = \mu_M$ and $\mathbb{P} = \mathbb{P}_M$. First of all, $\|\mu_1^\pi - \mu_1^{\tilde{\pi}}\|_1 = \sum_{s,a} \mu_1(s)|\pi_1(a|s) - \tilde{\pi}_1(a|s)|$. Furthermore, for any h,

$$\begin{aligned}
\|\mu_{h+1}^\pi - \mu_{h+1}^{\tilde{\pi}}\|_1 &= \sum_{s,a} |\mu_{h+1}^\pi(s,a) - \mu_{h+1}^{\tilde{\pi}}(s,a)| \\
&= \sum_{s,a} \left| \pi_{h+1}(a|s) \sum_{s',a'} \mu_h^\pi(s',a') \mathbb{P}_h(s|s',a') - \tilde{\pi}_{h+1}(a|s) \sum_{s',a'} \mu_h^{\tilde{\pi}}(s',a') \mathbb{P}_h(s|s',a') \right| \\
&\leq \sum_{s,a} |\pi_{h+1}(a|s) - \tilde{\pi}_{h+1}(a|s)| \sum_{s',a'} \mu_h^\pi(s',a') \mathbb{P}_h(s|s',a') \\
&\quad + \sum_{s,a} \tilde{\pi}_{h+1}(a|s) \sum_{s',a'} \left|\mu_h^\pi(s',a') - \mu_h^{\tilde{\pi}}(s',a')\right| \mathbb{P}_h(s|s',a') \\
&= \sum_{s,a} \mu_{h+1}^\pi(s) \left|\pi_{h+1}(a|s) - \tilde{\pi}_{h+1}(a|s)\right| + \sum_{s',a'} \left|\mu_h^\pi(s',a') - \mu_h^{\tilde{\pi}}(s',a')\right| \sum_{s,a} \tilde{\pi}_{h+1}(a|s) \mathbb{P}_h(s|s',a') \\
&= \sum_s \mu_{h+1}^\pi(s) \|\pi_{h+1}(\cdot|s) - \tilde{\pi}_{h+1}(\cdot|s)\|_1 + \|\mu_h^\pi - \mu_h^{\tilde{\pi}}\|_1.
\end{aligned}$$

By induction,

$$\|\mu_h^\pi - \mu_h^{\tilde{\pi}}\|_1 \leq \sum_{h'=1}^h \sum_s \mu_{h'}^\pi(s) \|\pi_{h'}(\cdot|s) - \tilde{\pi}_{h'}(\cdot|s)\|_1.$$

Finally,

$$\begin{aligned}
\|\mu^\pi - \mu^{\tilde{\pi}}\|_1 &\leq \sum_{h=1}^H \sum_{h'=1}^h \sum_s \mu_{h'}^\pi(s) \|\pi_{h'}(\cdot|s) - \tilde{\pi}_{h'}(\cdot|s)\|_1 \\
&\leq H \sum_{h=1}^H \sum_s \mu_h^\pi(s) \|\pi_h(\cdot|s) - \tilde{\pi}_h(\cdot|s)\|_1.
\end{aligned}$$

□

Lemma 4.2. *Given any M and target $\pi \in \Pi$, suppose the agents induce population density $\bar{\mu}_M$ in M, then:*

$$\begin{aligned}
\|\bar{\mu}_M - \mu_M^\pi\|_1 \leq & H \sum_{h,s} \bar{\mu}_{M,h}(s) \|\bar{\pi}_h(\cdot|s) - \pi_h(\cdot|s)\|_1, \\
&\textit{with } \bar{\mu}_{M,h}(s) := \sum_a \bar{\mu}_{M,h}(s,a)
\end{aligned} \tag{3}$$

Proof. Note that Ψ_M is a convex set and $\bar{\mu}_M \in \Psi_M$. By definition, we have $\bar{\mu}_M = \mu_M^{\bar{\pi}}$. By applying Lem. E.3 for model policy $\bar{\pi}$ and π in M, we finish the proof. □

Lemma E.4. *Consider any $\pi \in \Pi$ and models $M, \tilde{M}$, who are the same except with different transition functions $\mathbb{P}, \tilde{\mathbb{P}}$ respectively. Then,*

$$\|\mu_{\tilde{M}}^\pi - \mu_M^\pi\|_1 \leq H \sum_{h=1}^{H-1} \sum_{s,a} \mu_{M,h}^\pi(s,a) \|\tilde{\mathbb{P}}_h(\cdot|s,a) - \mathbb{P}_h(\cdot|s,a)\|_1.$$

Proof. Recall the definition of the state-action density function:

$$\mu^\pi_{M,h+1}(s,a) = \sum_{s',a'} \pi_{h+1}(a|s)\mathbb{P}_{M,h}(s|s',a')\mu^\pi_{M,h}(s',a') \quad \text{and} \quad \mu^\pi_1(s,a) = \pi_1(a|s)\mu_1(s).$$

We abbreviate $\tilde{\mu} = \mu^\pi_{\tilde{M}}, \mu = \mu^\pi_M$. Since μ_1 is the same for M and $\tilde{M}$, $\sum_{s,a} |\tilde{\mu}_1(s,a) - \mu_1(s,a)| = 0$. Furthermore, for all h,

$$\begin{aligned}
\|\tilde{\mu}_{h+1} - \mu_{h+1}\|_1 &= \sum_{s,a} |\tilde{\mu}_{h+1}(s,a) - \mu_{h+1}(s,a)| \\
&= \sum_{s,a}\sum_{s',a'} \pi_{h+1}(a|s) \left|\tilde{\mathbb{P}}_h(s|s',a')\tilde{\mu}_h(s',a') - \mathbb{P}_h(s|s',a')\mu_h(s',a')\right| \\
&= \sum_{s,a,s'} \left|\tilde{\mathbb{P}}_h(s'|s,a)\tilde{\mu}_h(s,a) - \mathbb{P}_h(s'|s,a)\mu_h(s,a)\right| \\
&\leq \sum_{s,a,s'} \tilde{\mathbb{P}}_h(s'|s,a)\,|\tilde{\mu}_h(s,a) - \mu_h(s,a)| + \mu_h(s,a)\left|\tilde{\mathbb{P}}_h(s'|s,a) - \mathbb{P}_h(s'|s,a)\right| \\
&= \sum_{s,a} |\tilde{\mu}_h(s,a) - \mu_h(s,a)| + \sum_{s,a} \mu_h(s,a) \sum_{s'} \left|\tilde{\mathbb{P}}_h(s'|s,a) - \mathbb{P}_h(s'|s,a)\right| \\
&= \|\tilde{\mu}_h - \mu_h\|_1 + \sum_{s,a} \mu_h(s,a)\|\tilde{\mathbb{P}}_{h'}(\cdot|s,a) - \mathbb{P}_{h'}(\cdot|s,a)\|_1.
\end{aligned}$$

Using induction on h, we obtain $\|\tilde{\mu}_h - \mu_h\|_1 \leq \sum_{h'=1}^{h-1}\sum_{s,a} \mu_{h'}(s,a)\|\tilde{\mathbb{P}}_{h'}(\cdot|s,a) - \mathbb{P}_{h'}(\cdot|s,a)\|_1$. Thus,

$$\begin{aligned}
\|\tilde{\mu} - \mu\|_1 &\leq \sum_{h=1}^{H}\sum_{h'=1}^{h-1}\sum_{s,a} \mu_{h'}(s,a)\|\tilde{\mathbb{P}}_{h'}(\cdot|s,a) - \mathbb{P}_{h'}(\cdot|s,a)\|_1 \\
&\leq H\sum_{h=1}^{H-1}\sum_{s,a} \mu_h(s,a)\|\tilde{\mathbb{P}}_h(\cdot|s,a) - \mathbb{P}_h(\cdot|s,a)\|_1.
\end{aligned}$$

□

F PROOFS OF RESULTS IN SECTION 4

Theorem 4.1. *If M^* is known and r^* is zero everywhere, under Asump. B and C, by choosing the steering reward $\forall\ t \in [T]$, $R^t(\mu) = \mu^* - \mu + \mathbf{1}\|\mu^* - \mu\|_\infty$, for any $\mu \in \Delta^H_{\mathcal{S}\times\mathcal{A}}$, we have:*

$$\Delta_T(\{\bar{\mu}^t_{M^*}\}_{t=1}^T) \leq L_U\sqrt{HSAT\,\text{AdaReg}(T)} = o(T)$$
$$C_T(\{\bar{\mu}^t_{M^*}, R^t\}_{t=1}^T) \leq 2H\sqrt{T\,\text{AdaReg}(T)} = o(T).$$

Proof. We abbreviate $\bar{\mu}^t = \bar{\mu}^t_{M^*}$. By Assumption B, $\sum_{t=1}^T \|\bar{\mu}^t - \mu^*\|_2^2 = \sum_{t=1}^T \langle R^t(\bar{\mu}^t), \mu^* - \bar{\mu}^t\rangle \leq \text{AdaReg}(T)$. Thus,

$$\sum_{t=1}^T \|\bar{\mu}^t - \mu^*\|_1^2 \leq HSA\sum_{t=1}^T \|\bar{\mu}^t - \mu^*\|_2^2 \leq HSA\,\text{AdaReg}(T).$$

By Assumption C and Jensen's inequality,

$$\max_\mu \sum_{t=1}^T U(\mu) - U(\bar{\mu}^t) \leq L_U \cdot \sum_{t=1}^T \|\mu^* - \bar{\mu}^t\|_1 \leq L_U \cdot \sqrt{HSAT\,\text{AdaReg}(T)}.$$

The steering cost can be bounded similarly.

$$\begin{aligned}\sum_{t=1}^T \langle R^t(\bar{\mu}^t), \bar{\mu}^t\rangle &= \sum_{t=1}^T H\|\mu^* - \bar{\mu}^t\|_\infty + \langle \mu^* - \bar{\mu}^t, \bar{\mu}^t\rangle \leq 2H\sum_{t=1}^T \|\mu^* - \bar{\mu}^t\|_\infty \\ &\leq 2H\sum_{t=1}^T \|\mu^* - \bar{\mu}^t\|_2 \leq 2H\sqrt{T\,\text{AdaReg}(T)}.\end{aligned}$$

Given that AdaReg is sub-linear in T, we finish the proof. □

Theorem 4.3. *Let $\pi^* = \arg\max_\pi U(\mu^\pi)$ and $R^t(\mu) = R_{\pi^*}(\mu) + \|R_{\pi^*}(\mu)\|_\infty \mathbf{1}$ for all t. Under Assump. B,*

$$\Delta_T(\{\bar{\mu}^t_{M^*}\}_{t=1}^T) \leq L_U\sqrt{H^3SAT\,\text{AdaReg}(T)} = o(T)$$
$$C_T(\{\bar{\mu}^t_{M^*}, R^t\}_{t=1}^T) \leq 4H\sqrt{T\,\text{AdaReg}(T)} = o(T).$$

Proof. The bound for the steering gap can be shown by first using the L_U-Lipschitzness of U and then applying Lem. G.1 under Assump. B, where $\pi^t_* = \pi$ for all t. The calculation of the steering cost is the same as in the proof of Theorem 5.1 with $K = 1$. □

G PROOF OF THEOREM 5.1

Lemma G.1. *Let $(\pi_*^t)_{t=1}^T$ be a sequence of policies. We abbreviate $\bar{\mu}^t = \bar{\mu}^t_{M^*}, \mu^{\pi_*^t} = \mu^{\pi_*^t}_{M^*}$. Then,*

$$\frac{1}{H}\sum_{t=1}^T \|\bar{\mu}^t - \mu^{\pi_*^t}\|_1 \leq \sum_{t=1}^T\sum_{h=1}^H\sum_{s\in\mathcal{S}} \bar{\mu}_h^t(s)\|\bar{\pi}_h^t(\cdot|s) - \pi_{*,h}^t(\cdot|s)\|_1$$
$$\leq \sqrt{HSAT\sum_{t=1}^T \langle R_{\pi_*^t}(\bar{\mu}^t), \mu^{\pi_*^t} - \bar{\mu}^t\rangle},$$

where $\bar{\pi}^t$ is the (population) policy which induces $\bar{\mu}^t = \mu^{\bar{\pi}^t}$.

Proof. The first inequality follows from Lemma E.3. We can write

$$\sum_{t=1}^T \left\langle R_{\pi_*^t}(\bar{\mu}^t), \mu^{\pi_*^t} - \bar{\mu}^t\right\rangle = \sum_{t=1}^T \left\|(W^{\pi_*^t} - I)\bar{\mu}^t\right\|_2^2$$
$$= \sum_{t=1}^T\sum_{h,s,a}\left(\pi_{*,h}^t(a|s)\sum_{a'}\bar{\mu}_h^t(s,a') - \bar{\mu}_h^t(s,a)\right)^2$$
$$= \sum_{t=1}^T\sum_{h,s,a}(\bar{\mu}_h^t(s))^2\left(\pi_{*,h}^t(a|s) - \bar{\pi}_h^t(a|s)\right)^2$$
$$= \sum_{t=1}^T\sum_{h,s}(\bar{\mu}_h^t(s))^2\left\|\pi_{*,h}^t(\cdot|s) - \bar{\pi}_h^t(\cdot|s)\right\|_2^2.$$

Furthermore, by Jensen's inequality,

$$\sum_{t=1}^T\sum_{h,s}\bar{\mu}_h^t(s)\|\bar{\pi}_h^t(\cdot|s) - \pi_{*,h}^t(\cdot|s)\|_1 \leq \sqrt{A}\sum_{t=1}^T\sum_{h,s}\bar{\mu}_h^t(s)\|\bar{\pi}_h^t(\cdot|s) - \pi_{*,h}^t(\cdot|s)\|_2$$
$$\leq \sqrt{HSAT\sum_{t=1}^T\sum_{h,s}(\bar{\mu}_h^t(s))^2\|\bar{\pi}_h^t(\cdot|s) - \pi_{*,h}^t(\cdot|s)\|_2^2}$$
$$= \sqrt{HSAT\sum_{t=1}^T\langle R_{\pi_*^t}(\bar{\mu}^t), \mu^{\pi_*^t} - \bar{\mu}^t\rangle}.$$

□

Lemma G.2. *For any $0 < \delta < 1$, with probability at least $1 - \delta$,*

$$\|\mathbb{P}_{\hat{M}^k,h}(\cdot|s,a) - \mathbb{P}_h^*(\cdot|s,a)\|_1 \leq 2\varepsilon_k(h,s,a)$$

for all t, h, s, a, where $\varepsilon_k(h,s,a) := \sqrt{\frac{2S\ln(THSA/\delta)}{\max\{1,N_k(h,s,a)\}}}$.

Proof. Since $\mathbb{P}_{\hat{M}^k} \in \mathcal{P}^k$, we have $\|\mathbb{P}_{\hat{M}^k,h}(\cdot|s,a) - \bar{\mathbb{P}}_h^k(\cdot|s,a)\|_1 \leq \varepsilon_k(h,s,a)$ for all k, h, s, a. By (7) and Theorem 2.1 of Weissman et al. (2003),

$$\Pr\left[\|\bar{\mathbb{P}}_h^k(\cdot|s,a) - \mathbb{P}_h^*(\cdot|s,a)\|_1 > \varepsilon\right] \leq (2^S - 2)e^{-N_k(h,s,a)\varepsilon^2/2}.$$

Plugging in $\varepsilon_k(h,s,a)$ for ε bounds this probability with $\delta/(THSA)$. The triangle inequality and a union bound over all k, h, s, a imply the result. □

Lemma G.3. *For any $0 < \delta < 1$ and respective $\varepsilon_k(h, s, a)$,*

$$\sum_{t=1}^{T} \sum_{h=1}^{H-1} \varepsilon_{k(t)}(h, s_h^t, a_h^t) \leq 3HS\sqrt{2\ln(THSA/\delta)AT}.$$

Proof. We can define $n_k(h, s, a) := \sum_{t=T_{k-1}+1}^{T_k} \mathbb{I}\{s_h^t = s, a_h^t = a\}$. Clearly, $N_k(h, s, a) = \sum_{k'<k} n_k(h, s, a)$. The condition in line 5 of the algorithm ensures that $n_k(h, s, a) \leq N_k(h, s, a)$ for all k, h, s, a. Thus, we can use Lemma 19 in Jaksch et al. (2010) and Jensen's inequality,

$$\begin{aligned}
&\sum_{t=1}^{T} \sum_{h=1}^{H-1} \frac{1}{\sqrt{\max\{1, N_{k(t)}(h, s_h^t, a_h^t)\}}} = \sum_{k=1}^{K} \sum_{t=T_{k-1}+1}^{T_k} \sum_{h=1}^{H-1} \frac{1}{\sqrt{\max\{1, N_k(h, s_h^t, a_h^t)\}}} \\
&= \sum_{k=1}^{K} \sum_{h=1}^{H-1} \sum_{s,a} \sum_{t=T_{k-1}+1}^{T_k} \frac{\mathbb{I}\{s_h^t = s, a_h^t = a\}}{\sqrt{\max\{1, N_k(h, s, a)\}}} = \sum_{k=1}^{K} \sum_{h=1}^{H-1} \sum_{s,a} \frac{n_k(h, s, a)}{\sqrt{\max\{1, N_k(h, s, a)\}}} \\
&\leq 3 \sum_{h=1}^{H-1} \sum_{s,a} \sqrt{N_K(h, s, a) + n_K(h, s, a)} \leq 3\sqrt{HSA \sum_{h=1}^{H-1} \sum_{s,a} (N_K(h, s, a) + n_K(h, s, a))} \\
&= 3\sqrt{HSA \cdot HT} = 3H\sqrt{SAT}.
\end{aligned}$$

Now, using the definition of $\varepsilon_k(h, s, a)$,

$$\begin{aligned}
\sum_{t=1}^{T} \sum_{h=1}^{H-1} \varepsilon_{k(t)}(h, s_h^t, a_h^t) &= \sum_{t=1}^{T} \sum_{h=1}^{H-1} \sqrt{\frac{2S\ln(THSA/\delta)}{\max\{1, N_{k(t)}(h, s_h^t, a_h^t)\}}} \\
&\leq \sqrt{2S\ln(THSA/\delta)} \cdot 3H\sqrt{SAT} = 3HS\sqrt{2\ln(THSA/\delta)AT}.
\end{aligned}$$

□

Lemma G.4. *Let $(\bar{\pi}^t)_{t=1}^T$ be the policy sequence of the population and $(\hat{M}^k)_{k=1}^K$ the sequence of the corresponding model estimates. We abbreviate $\bar{\mu}^t = \bar{\mu}_{M^*}^t$, $\hat{\mu}^t = \mu_{\hat{M}^{k(t)}}^{\bar{\pi}^t}$. With probability at least $1 - 2\delta$,*

$$\sum_{t=1}^{T} \left\|\hat{\mu}^t - \bar{\mu}^t\right\|_1 \leq 12H^2S\sqrt{\ln(THSA/\delta)AT}.$$

Proof. The proof is based on Rosenberg and Mansour (2019). Let $(s_h^t, a_h^t)_{h=1}^H$ be the trajectory sampled in the t-th game. We define $\xi_k(h, s, a) := \|\mathbb{P}_{\hat{M}^k,h}(\cdot|s, a) - \mathbb{P}_h^*(\cdot|s, a)\|_1$. By Lemma E.4,

$$\begin{aligned}
\sum_{t=1}^{T} \left\|\hat{\mu}^t - \bar{\mu}^t\right\|_1 &\leq H \sum_{t=1}^{T} \sum_{h=1}^{H-1} \sum_{s,a} \bar{\mu}_h^t(s, a)\xi_{k(t)}(h, s, a) \\
&= H \sum_{t=1}^{T} \sum_{h=1}^{H-1} \xi_{k(t)}(h, s_h^t, a_h^t) \\
&\quad + H \sum_{t=1}^{T} \sum_{h=1}^{H-1} \underbrace{\left(\sum_{s,a} \bar{\mu}_h^t(s, a)\xi_{k(t)}(h, s, a) - \sum_{s,a} \mathbb{I}\{s_h^t = s, a_h^t = a\}\xi_{k(t)}(h, s, a)\right)}_{=:Y_t(h)},
\end{aligned}$$

where $(Y_t(h))_t$ is a martingale difference sequence w.r.t. the trajectories sampled and with $|Y_t(h)| \leq \max_{s,a} \xi_{k(t)}(h, s, a) \leq 2$. In the following, we bound the first and second term above with high probability.

The first term can be bounded using Lemma G.2 and G.3, such that we have, with probability at least $1 - \delta$,

$$H \sum_{t=1}^{T} \sum_{h=1}^{H-1} \xi_{k(t)}(h, s_h^t, a_h^t) \leq 2H \sum_{t=1}^{T} \sum_{h=1}^{H-1} \varepsilon_{k(t)}(h, s_h^t, a_h^t) \leq 2H \cdot 3H\sqrt{2S\ln(THSA/\delta) \cdot SAT}.$$

By the Hoeffding-Azuma inequality, we have for a fixed h that with probability at least $1-\delta/H$,

$$\sum_{t=1}^{T} Y_t(h) \leq 2\sqrt{2T\ln(H/\delta)}.$$

Thus, by the union bound over all h, the second term is at most $2H^2\sqrt{2T\ln(H/\delta)}$ with probability at least $1-\delta$.

Finally, by union bound over the events used to bound the first and second term, we have with probability at least $1-2\delta$ that

$$\begin{aligned}\sum_{t=1}^{T} \left\|\hat{\mu}^t - \bar{\mu}^t\right\|_1 &\leq 2H^2\sqrt{2T\ln(H/\delta)} + 6H^2\sqrt{2S\ln(THSA/\delta)\cdot SAT} \\ &\leq 12H^2 S\sqrt{\ln(THSA/\delta)AT}.\end{aligned}$$

□

Theorem 5.1. *Suppose the intrinsic reward $r^*=0$, under Assump. B and C, if we run Alg. 2 with $\delta\in(0,1)$, then with probability at least $1-2\delta$, $K\leq HSA\log_2 T$, and*

$$\begin{aligned}\Delta_T(\{\bar{\mu}^t_{M^*}\}_{t=1}^T) &\leq L_U\sqrt{H^3SATK\,\mathrm{AdaReg}(T)} \\ &\quad + 36L_U H^3 S\sqrt{\ln(THSA/\delta)AT} = o(T). \\ C_T(\{\bar{\mu}^t_{M^*}, R^t_z\}_{t=1}^T) &\leq 4H\sqrt{TK\,\mathrm{AdaReg}(T)} = o(T).\end{aligned}$$

Proof. We first establish the upper bound for steering gap and then investigate the steering cost.

Proof for Steering Gap We denote with $k(t)$ the episode index at the t-th game and denote $\pi^* = \arg\max_\pi U(\mu^\pi_{M^*})$. Furthermore, we abbreviate $\bar{\mu}^t = \bar{\mu}^t_{M^*}, \mu^k_* = \mu^{\pi^k_*}_{M^*}, \hat{\mu}^t = \mu^{\bar{\pi}^t}_{\hat{M}^{k(t)}}$ and $\hat{\mu}^k_* = \mu^{\pi^k_*}_{\hat{M}^k}$. Consider a fixed t and $k=k(t)$. We can decompose the steering gap term of round t as follows:

$$U(\mu^{\pi^*}) - U(\bar{\mu}^t) = \Big(U(\mu^{\pi^*}) - U(\hat{\mu}^k_*)\Big) + \big(U(\hat{\mu}^k_*) - U(\bar{\mu}^t)\big)$$

The first term can be bounded by 0 using the optimism of the algorithm. We use the L_U-Lipschitzness of U and the triangle inequality to further decompose the second term.

$$U(\hat{\mu}^k_*) - U(\bar{\mu}^t) \leq L_U\|\hat{\mu}^k_* - \bar{\mu}^t\|_1 \leq L_U\|\hat{\mu}^k_* - \hat{\mu}^t\|_1 + L_U\|\hat{\mu}^t - \bar{\mu}^t\|_1.$$

Applying Lemma E.3, we get

$$\begin{aligned}\|\hat{\mu}^k_* - \hat{\mu}^t\|_1 &\leq H\sum_{h,s}\hat{\mu}^t_h(s)\|\pi^k_{*,h}(\cdot|s) - \bar{\pi}^t_h(\cdot|s)\|_1 \\ &\leq H\sum_{h,s}\bar{\mu}^t_h(s)\cdot\|\pi^k_{*,h}(\cdot|s) - \bar{\pi}^t_h(\cdot|s)\|_1 + \underbrace{H\sum_{h,s}|\hat{\mu}^t_h(s) - \bar{\mu}^t_h(s)|\cdot\|\pi^k_{*,h}(\cdot|s) - \bar{\pi}^t_h(\cdot|s)\|_1}_{(*)},\end{aligned}$$

where the second term can be bounded with

$$(*) \leq 2\sum_{h,s}|\hat{\mu}^t_h(s) - \bar{\mu}^t_h(s)| \leq 2\sum_{h,s}\left|\sum_a\hat{\mu}^t_h(s,a) - \sum_a\bar{\mu}^t_h(s,a)\right| \leq 2\|\hat{\mu}^t - \bar{\mu}^t\|_1.$$

Putting it all together we now arrive at

$$U(\mu^{\pi^*}) - U(\bar{\mu}^t) \leq L_U H\sum_{h,s}\bar{\mu}^t_h(s)\|\pi^k_{*,h}(\cdot|s) - \bar{\pi}^t_h(\cdot|s)\|_1 + L_U(2H+1)\|\hat{\mu}^t - \bar{\mu}^t\|_1.$$

By summing over t,

$$\sum_{t=1}^{T} U(\mu^{\pi^*}) - U(\bar{\mu}^t) \le L_U H \underbrace{\sum_{t=1}^{T}\sum_{h,s} \bar{\mu}_h^t(s) \|\pi_{*,h}^{k(t)}(\cdot|s) - \bar{\pi}_h^t(\cdot|s)\|_1}_{\Delta_{\text{pop}}} + L_U(2H+1)\underbrace{\sum_{t=1}^{T} \|\hat{\mu}^t - \bar{\mu}^t\|_1}_{\Delta_{\text{est}}}.$$

Using Lemma G.4, the estimation error term Δ_{est} can bounded by $12H^2 S\sqrt{\ln(THSA/\delta)AT}$ with probability at least $1-2\delta$.

To bound the population convergence term Δ_{pop}, we can use Lemma G.1:

$$\sum_{t=1}^{T}\sum_{h,s} \bar{\mu}_h^t(s) \|\pi_{*,h}^{k(t)}(\cdot|s) - \bar{\pi}_h^t(\cdot|s)\|_1 \le \sqrt{HSAT \underbrace{\sum_{t=1}^{T} \langle R_{\pi_*^{k(t)}}(\bar{\mu}^t), \mu_*^{k(t)} - \bar{\mu}^t\rangle}_{\texttt{AgentReg}}}$$

Furthermore, it can be easily seen that `AgentReg` is

$$\sum_{t=1}^{T} \langle R_{\text{z}}^t(\bar{\mu}^t), \mu_*^{k(t)} - \bar{\mu}^t\rangle = \sum_{k=1}^{K} \underbrace{\sum_{t=T_{k-1}+1}^{T_k} \langle R_{\text{z}}^t(\bar{\mu}^t), \mu_*^{k} - \bar{\mu}^t\rangle}_{\le \text{AdaReg}(T)} \le K \cdot \text{AdaReg}(T).$$

Finally, to bound the number of episodes K, note that K is also the number of times the condition in line 5 of the algorithm has been true. For each (h,s,a), this condition can be true at most $\log_2 T$ times. Thus, $K \le HSA\log_2 T$.

Proof for Steering Costs Note that for any reward function R,

$$\langle R(\mu) + \|R(\mu)\|_\infty \mathbf{1}, \mu\rangle = H\|R(\mu)\|_\infty + \langle R(\mu), \mu\rangle \le 2H\|R(\mu)\|_\infty.$$

Let $\pi^* = \pi_*^k$ for some k. Recall that $R_{\pi^*}(\mu) = -((W^{\pi^*} - I)\mu)^\top (W^{\pi^*} - I)$. By looking at the definition of W^{π^*} in (4), we see that

$$\left\|(W^{\pi^*} - I)^\top\right\|_\infty = \max_{h,s,a} \sum_{a'\neq a} |\pi_h(a'|s)| + |\pi_h(a|s) - 1| \le 2,$$

where the $\|\cdot\|_\infty$-matrix norm is defined as $\|M\|_\infty = \max_i \sum_j |M_{ij}|$. Using this, we can bound

$$\begin{aligned}\|R_{\pi^*}(\mu)\|_\infty &= \left\|((W^{\pi^*} - I)\mu)^\top (W^{\pi^*} - I)\right\|_\infty = \left\|(W^{\pi^*} - I)^\top (W^{\pi^*} - I)\mu\right\|_\infty \\ &\le \left\|(W^{\pi^*} - I)^\top\right\|_\infty \cdot \left\|(W^{\pi^*} - I)\mu\right\|_\infty \le 2\left\|(W^{\pi^*} - I)\mu\right\|_2.\end{aligned}$$

Finally, using Jensen's inequality and the fact that the agent regret is bounded by $K\,\text{AdaReg}(T)$, our steering cost can be bounded by

$$\begin{aligned}\sum_{t=1}^{T} \left\langle R_{\text{z}}^t(\bar{\mu}^t), \bar{\mu}^t\right\rangle &= \sum_{t=1}^{T} \left\langle R_{\pi_*^{k(t)}}(\bar{\mu}^t) + \|R_{\pi_*^{k(t)}}(\bar{\mu}^t)\|_\infty \mathbf{1}, \bar{\mu}^t\right\rangle \\ &\le 4H\sum_{t=1}^{T} \left\|(W^{\pi_*^{k(t)}} - I)\bar{\mu}^t\right\|_2 \le 4H\sqrt{T\sum_{t=1}^{T}\left\|(W^{\pi_*^{k(t)}} - I)\bar{\mu}^t\right\|_2^2} \\ &\le 4H\sqrt{T\sum_{t=1}^{T}\langle R_{\text{z}}^t(\bar{\mu}^t), \mu_*^{k(t)} - \bar{\mu}^t\rangle} \le 4H\sqrt{TK\,\text{AdaReg}(T)}\end{aligned}$$

□

H ELUDER DIMENSION

H.1 Example Function Classes

Here, we list some bounds of the eluder dimension for different function classes that are commonly considered. We see that in all these cases, the eluder dimension can be bounded logarithmically in T, if $\varepsilon = T^{-1}$.

Proposition H.1 (Linear functions, Russo and Van Roy (2013))**.** *Let $\mathcal{F} = \{f | f(x) = \theta^\top \phi(x), \theta \in \mathbb{R}^d, \|\theta\|_2 \leq C_\theta, \|\phi(x)\|_2 \leq C_\phi\}$.*

$$\dim_E(\mathcal{F}, \varepsilon) \leq 3d \frac{e}{e-1} \ln\left(3 + 3\left(\frac{2C_\theta}{\varepsilon}\right)^2\right) + 1.$$

Proposition H.2 (Quadratic functions, Osband and Roy (2014))**.** *Let $\mathcal{F} = \{f | f(x) = \phi(x)^\top \theta \phi(x), \theta \in \mathbb{R}^{p \times p}, \phi \in \mathbb{R}^p, \|\theta\|_2 \leq C_\theta, \|\phi\|_2 \leq C_\phi\}$.*

$$\dim_E(\mathcal{F}, \varepsilon) \leq p(4p-1) \frac{e}{e-1} \log\left(\left(1 + \left(\frac{2pC_\phi^2 C_\theta}{\varepsilon}\right)^2\right)(4p-1)\right) + 1.$$

Proposition H.3 (Generalized linear functions, Russo and Van Roy (2013))**.** *Let g be strictly increasing, differentiable and have derivatives bounded in $[\underline{h}, \overline{h}]$ with $\overline{h} > \underline{h} > 0$. Let $r = \overline{h}/\underline{h}$ and $\mathcal{F} = \{f | f(x) = g(\theta^\top \phi(x)), \theta \in \mathbb{R}^d, \|\theta\|_2 \leq C_\theta, \|\phi\|_2 \leq C_\phi\}$.*

$$\dim_E(\mathcal{F}, \varepsilon) \leq 3dr^2 \frac{e}{e-1} \log\left(3r^2 + 3r^2\left(\frac{2C_\theta \overline{h}}{\varepsilon}\right)^2\right) + 1.$$

Remark H.4 (Bounding β_T)**.** If we assume that the functions in $\mathcal{R}$ are parametrized by parameters in some set $\Theta \subset \mathbb{R}^d$ with constant diameter and the functions are L-Lipschitz in that parameter, we have $N(\mathcal{R}, \alpha, \|\cdot\|_\infty) \leq N(\Theta, \alpha/L, \|\cdot\|_\infty) \leq (1 + \mathcal{O}(L/\alpha))^d$. Then, we might choose $\alpha = T^{-1}$ such that

$$\beta_T = 8\sigma^2 \log(N(\mathcal{R}, \alpha, \|\cdot\|_\infty)/\delta) + 2\alpha T(8r_{\max} + \sqrt{8\sigma^2 \ln(4T^2/\delta)})$$

can also be bounded logarithmically in T.

H.2 Bounding The Width Of The Confidence Set

Notations and Definitions Here, we introduce some notation used in this section. We define the width function $w_\mathcal{F}(x) = \sup_{\underline{f}, \overline{f} \in \mathcal{F}} |\underline{f}(x) - \overline{f}(x)|$. Throughout this section, we use the notation x_{H_t+h} with $H_t = (t-1)H$ to describe elements of a sequence $x_1, ..., x_{HT}$. The idea behind it is that we can later define $x_{H_t+h} = (h, s_h^t, a_h^t, \bar{\mu}_{M^*,h}^t)$ and apply the results in this section to our setting. Furthermore, for any function g we write $\|g\|_{2,E_t}^2 = \sum_{i=1}^{t-1} \sum_{h=1}^{H} g^2(x_{H_t+h})$.

Lemma H.5 (Proposition 3 of Russo and Van Roy (2013))**.** *If $(\beta_t)_{t \in \mathbb{N}}$ is a positive non-decreasing sequence, $(\hat{f}_t)_t$ some function sequence and $\mathcal{F}_t := \{f \in \mathcal{F} : \|f - \hat{f}_t\|_{2,E_t} \leq \sqrt{\beta_t}\}$ then with probability 1, for all $T \in \mathbb{N}$,*

$$\sum_{t=1}^{T} \sum_{h=1}^{H} \mathbb{I}\{w_{\mathcal{F}_t}(x_{H_t+h}) > \varepsilon\} \leq \left(\frac{4\beta_T}{\varepsilon^2} + H\right) \dim_E(\mathcal{F}, \varepsilon)$$

for all $T \in \mathbb{N}$ and $\varepsilon > 0$.

Proof. First we show that for any $\tau = H_t + h < TH$, if $w_{\mathcal{F}_t}(x_\tau) > \varepsilon$ then x_τ is $(\mathcal{F}, \varepsilon)$-dependent on fewer than $4\beta_T/\varepsilon^2$ disjoint subsequences of $(x_1, ..., x_{H_t})$. Suppose $w_{\mathcal{F}_t}(x_\tau) > \varepsilon$. Then, there are $f, \tilde{f} \in \mathcal{F}_t$ such that $|f(x_\tau) - \tilde{f}(x_\tau)| > \varepsilon$. Furthermore, let $(x_{i_1}, ..., x_{i_k})$ be a subsequence of $(x_1, ..., x_{H_t})$ on which x_τ is $(\mathcal{F}, \varepsilon)$-dependent. This implies, by definition, that $\sum_{j=1}^{k} (f(x_{i_j}) - \tilde{f}(x_{i_j}))^2 > \varepsilon^2$. If x_τ is $(\mathcal{F}, \varepsilon)$-dependent on K disjoint subsequences of $(x_1, ..., x_{H_t})$ then we must have

$$\|f - \tilde{f}\|_{2,E_t}^2 = \sum_{i=1}^{t-1} \sum_{h=1}^{H} (f(x_{H_i+h}) - \tilde{f}(x_{H_i+h}))^2 \geq \sum_{l=1}^{K} \sum_{j=1}^{k_l} (f(x_{i_j^l}) - \tilde{f}(x_{i_j^l}))^2 > K\varepsilon^2.$$

By the triangle inequality, $\|f - \tilde{f}\|_{2,E_t} \leq \|f - \hat{f}_t\|_{2,E_t} + \|\tilde{f} - \hat{f}_t\|_{2,E_t} \leq 2\sqrt{\beta_t} \leq 2\sqrt{\beta_T}$. Combining these two inequalities, we get $K < 4\beta_T/\varepsilon^2$.

Next, we show that in any sequence $(y_1, ..., y_l)$ there is an element y_j which is $(\mathcal{F}, \varepsilon)$-dependent on at least $l/d - 1$ disjoint subsequences of $(y_1, ..., y_{j-1})$, where $d = \dim_E(\mathcal{F}, \varepsilon)$. Let K be an integer with $Kd + 1 \leq l \leq Kd + d$. We will construct K disjoint subsequences $B_1, ..., B_K$. First, $B_i = (y_i)$ for all $i \in [K]$. If y_{K+1} is already $(\mathcal{F}, \varepsilon)$-dependent on $B_1, ..., B_K$, we are done. Otherwise, select a B_i of which y_{K+1} is $(\mathcal{F}, \varepsilon)$-independent and append y_{K+1} to B_i. We repeat this for $y_{K+2}, y_{K+3}, ...$ until we find y_j that is $(\mathcal{F}, \varepsilon)$-dependent on each subsequence or until we have reached y_l. In the latter case, each element of a subsequence B_i is independent of its predecessors and hence $|B_i| = d$. Then, y_l must be $(\mathcal{F}, \varepsilon)$-dependent on each subsequence, by definition of the eluder dimension. In both cases we find an element in $(y_1, ..., y_l)$ that is $(\mathcal{F}, \varepsilon)$-dependent on $K \geq t/d - 1$ disjoint subsequences.

Finally, let $(y_1, ..., y_l) = (x_{i_1}, ..., x_{i_l})$ be a subsequence of $(x_1, ..., x_{TH})$ consisting of all elements x_{H_t+h} for which $w_{\mathcal{F}_t}(x_{H_t+h}) > \varepsilon$. From before, we know there is some y_j that is $(\mathcal{F}, \varepsilon)$-dependent on at least $l/d - 1$ disjoint subsequences of $(y_1, ..., y_{j-1})$. Let t, h be such that $y_j = x_{H_t+h}$. Note that in $(y_1, ..., y_{j-1})$ there are at most $H - 1$ elements $y_i = x_{H_t+h'}$ for some $h' < h$. From this follows that $y_j = x_{H_t+h}$ is $(\mathcal{F}, \varepsilon)$-dependent on at least $l/d - 1 - (H - 1) = l/d - H$ disjoint subsequences of $(y_1, ..., y_{j-H}) \subseteq (x_1, ..., x_{H_t})$. Now, as we have also shown, x_{H_t+h} is $(\mathcal{F}, \varepsilon)$-dependent on fewer than $4\beta_T/\varepsilon^2$ disjoint subsequences of $(x_1, ..., x_{H_t})$. Combining these two bounds, we get $l/d - H \leq 4\beta_T/\varepsilon^2$, and therefore $l \leq (4\beta_T/\varepsilon^2 + H)d$. □

Lemma H.6 (Variant of Lemma 2 in Russo and Van Roy (2013))**.** *Let $(\beta_t)_{t\in\mathbb{N}}$ be a positive non-decreasing sequence, $(\hat{f}_t)_t$ some function sequence and $\mathcal{F}_t := \{f \in \mathcal{F} : \|f - \hat{f}_t\|_{2,E_t} \leq \sqrt{\beta_t}\}$. Let $w_{\mathcal{F}}(x) \leq C$ for all x. Then, for all $T \in \mathbb{N}$ and $\varepsilon > 0$,*

$$\begin{aligned}\sum_{t=1}^{T}\sum_{h=1}^{H} w_{\mathcal{F}_t}(x_{H_t+h}) \leq \varepsilon HT + CH \dim_E(\mathcal{F}, \varepsilon)& \\ +4\sqrt{\beta_T H \dim_E(\mathcal{F}, \varepsilon) T}&.\end{aligned}$$

Proof. We abbreviate $w_{H_t+h} = w_{\mathcal{F}_t}(x_{H_t+h})$ and $d = \dim_E(\mathcal{F}, \varepsilon)$. Let $w_{i_1} \geq ... \geq w_{i_{HT}}$. Using this ordering of the sequence, $w_{i_k} > \varepsilon$ implies that $\sum_{j=1}^{T} \mathbb{I}\{w_j > \varepsilon\} \geq k$. By Lemma H.5, this would mean $k \leq (4\beta_T/\varepsilon^2 + H)d$ or, equivalently, $\varepsilon < \sqrt{4\beta_T d/(k - Hd)}$. Now, since $w_{i_k} > \varepsilon$ implies $\varepsilon < \sqrt{4\beta_T d/(k - Hd)}$, this means that $w_{i_k} < \sqrt{4\beta_T d/(k - Hd)}$.

In the following, we bound the first and largest widths $w_{i_1}, ..., w_{i_{Hd}}$ by C and the remaining widths (larger than ε) by the previously established bound.

$$\begin{aligned}\sum_{t=1}^{T}\sum_{h=1}^{H} w_{H_t+h} &= \sum_{k=1}^{HT} \mathbb{I}\{w_k \leq \varepsilon\} w_k + \sum_{k=1}^{HT} \mathbb{I}\{w_k > \varepsilon\} w_k \leq \varepsilon HT + \sum_{k=1}^{HT} \mathbb{I}\{w_k > \varepsilon\} w_k \\ &\leq \varepsilon HT + HdC + \sum_{k=Hd+1}^{HT} \mathbb{I}\{w_{i_k} > \varepsilon\} w_{k_t} \\ &\leq \varepsilon HT + HdC + \sum_{k=Hd+1}^{HT} \sqrt{4\beta_T d/(k - Hd)} \\ &\leq \varepsilon HT + HdC + \sqrt{4d\beta_T} \int_0^{HT} \frac{1}{\sqrt{x}} dx \\ &= \varepsilon HT + HdC + 4\sqrt{d\beta_T HT}\end{aligned}$$

□

I PROOF OF THEOREM 6.1

I.1 Algorithm Details

We present our full algorithm for the unknown reward setting in Alg. 4.

Algorithm 4 Steering reward design for Scenario 2

```
Initialize $\mathcal{P}^1$ := set of all possible transition functions, $\pi_*^1$ (arbitrarily), $k = 1, T_0 = 0$.
for $t = 1, ..., T$ do
    Update $\hat{\mathcal{R}}^t$ as in (10).
    Choose $R_{\mathrm{nz}}^t$ as in (11).
    Agents play $t$-th game with $r^* + R_{\mathrm{nz}}^t$.
    Obtain trajectory $((s_h^t, a_h^t, r_h^t))_{h=1}^H$.
    if $\exists(h, s, a),\ s.t.\ n_k(h, s, a) \geq N_k(h, s, a)$ then
        Update $\mathcal{P}^{k+1}$ as in (7).
        $T_k \leftarrow t$; $k \leftarrow k + 1$.
        $\pi_*^k, \hat{M}^k \leftarrow \arg\max_{\pi \in \Pi, \hat{M}: \mathbb{P}_{\hat{M}} \in \mathcal{P}^k} U(\mu_{\hat{M}}^\pi)$.
    end if
end for
```

I.2 Missing Proofs

Lemma I.1 (Proposition 2 in Russo and Van Roy (2013))**.** *Let $N(\mathcal{R}, \alpha, \|\cdot\|_\infty)$ be the α-covering number of $\mathcal{R}$ w.r.t. the $\|\cdot\|_\infty$-norm. Let $\delta > 0, \alpha > 0$, and for each t, $\beta_t = 8\sigma^2 \log(N(\mathcal{R}, \alpha, \|\cdot\|_\infty)/\delta) + 2\alpha t(8r_{\max} + \sqrt{8\sigma^2 \ln(4t^2/\delta)})$. With probability at least $1 - 2\delta$, $r^* \in \bigcap_{t=1}^\infty \hat{\mathcal{R}}^t$.*

Lemma I.2. *We abbreviate $\bar{\mu}^t = \bar{\mu}_{M^*}^t$. With probability at least $1 - \delta$,*

$$\sum_{t=1}^T \langle w_{\hat{\mathcal{R}}^t}(\bar{\mu}^t), \bar{\mu}^t \rangle \leq 3 \sum_{t=1}^T \sum_{h=1}^H w_{\hat{\mathcal{R}}^t}(h, s_h^t, a_h^t, \bar{\mu}^t) + r_{\max} H \ln(1/\delta).$$

Proof. Note that $\langle w_{\hat{\mathcal{R}}^t}(\bar{\mu}^t), \bar{\mu}^t \rangle = \mathbb{E}_{(s_h, a_h)_{h=1}^H \sim \bar{\mu}^t}[\sum_{h=1}^H w_{\hat{\mathcal{R}}^t}(h, s_h, a_h, \bar{\mu}^t)] =: Y_t$. Recall that $(s_h^t, a_h^t)_h \sim \bar{\mu}^t$ are the trajectories we gather from the population at step t. Therefore, we can define $X_t := \sum_{h=1}^H w_{\hat{\mathcal{R}}^t}(h, s_h^t, a_h^t, \bar{\mu}^t)$ with $\mathbb{E}[X_t | \bar{\mu}^t] = Y_t$. By the assumption that r^* is bounded in $[0, r_{\max}]$, we have that $w_{\hat{\mathcal{R}}}(h, s, a, \mu) \leq r_{\max}$ for any μ, h, s, a and $\hat{\mathcal{R}} \subseteq \mathcal{R}$. Therefore, $0 \leq X_t \leq r_{\max} H$. A direct application of Lemma D.4 from Huang et al. (2023) shows that with probability at least $1 - \delta$,

$$\sum_{t=1}^T Y_t \leq 3 \sum_{t=1}^T X_t + r_{\max} H \ln \frac{1}{\delta}.$$

□

Lemma I.3. *We abbreviate $\mu_*^k = \mu_{M^*}^{\pi_*^k}, \bar{\mu}^t = \bar{\mu}_{M^*}^t$. If the true r^* is contained in all $\hat{\mathcal{R}}^t$, then, with probability at least $1 - \delta$,*

$$\sum_{t=1}^T \langle R_{\pi_*^{k(t)}}(\bar{\mu}^t), \mu_*^{k(t)} - \bar{\mu}^t \rangle \leq \sum_{t=1}^T \langle r^*(\bar{\mu}^t) + R_{nz}^t(\bar{\mu}^t), \mu_*^{k(t)} - \bar{\mu}^t \rangle + 6 \sum_{t=1}^T \sum_{h=1}^H w_{\hat{\mathcal{R}}^t}(h, s_h^t, a_h^t, \bar{\mu}^t) + 2r_{\max} H \ln \frac{1}{\delta}.$$

Proof. Let $t \in [T]$ and $k = k(t)$. By Eq. (11),

$$\langle R_{\pi_*^k}(\bar{\mu}^t), \mu_*^k - \bar{\mu}^t \rangle = \langle r^*(\bar{\mu}^t) + R_{\mathrm{nz}}^t(\bar{\mu}^t), \mu_*^k - \bar{\mu}^t \rangle + \langle \bar{r}^t(\bar{\mu}^t) - r^*(\bar{\mu}^t) - w_{\hat{\mathcal{R}}^t}(\bar{\mu}^t), \mu_*^k - \bar{\mu}^t \rangle.$$

With that, we have already separated out the first term (agent regret). Using the assumption that $r^* \in \hat{\mathcal{R}}^t$ for all t, we can bound the second term as follows.

$$\begin{aligned}
&\langle \bar{r}^t(\bar{\mu}^t) - r^*(\bar{\mu}^t) - w_{\hat{\mathcal{R}}^t}(\bar{\mu}^t), \mu_*^k - \bar{\mu}^t \rangle \\
&= \langle r^*(\bar{\mu}^t) - \bar{r}^t(\bar{\mu}^t) + w_{\hat{\mathcal{R}}^t}(\bar{\mu}^t), \bar{\mu}^t \rangle + \langle \underbrace{\bar{r}^t(\bar{\mu}^t) - r^*(\bar{\mu}^t)}_{\le w_{\hat{\mathcal{R}}^t}(\bar{\mu}^t)} - w_{\hat{\mathcal{R}}^t}(\bar{\mu}^t), \mu_*^k \rangle \\
&\le \langle \underbrace{r^*(\bar{\mu}^t) - \bar{r}^t(\bar{\mu}^t)}_{\le w_{\hat{\mathcal{R}}^t}(\bar{\mu}^t)} + w_{\hat{\mathcal{R}}^t}(\bar{\mu}^t), \bar{\mu}^t \rangle \le 2\langle w_{\hat{\mathcal{R}}^t}(\bar{\mu}^t), \bar{\mu}^t \rangle
\end{aligned}$$

Finally, we can bound $\sum_{t=1}^T \langle w_{\hat{\mathcal{R}}^t}(\bar{\mu}^t), \bar{\mu}^t \rangle$ using Lemma I.2, which implies the result. □

Theorem 6.1. *Under Assump. A, B and C, if we run Alg. 4 with $0 < \delta < 1$, then with probability at least $1 - 6\delta$, $K \le HSA \log_2 T$, and*

$$\begin{aligned}
\Delta_T(\{\bar{\mu}_{M^*}^t\}_{t=1}^T) &\le L_U \sqrt{H^3 SAT(K\,\mathrm{AdaReg}(T) + D)} \\
&\quad + 36 L_U H^3 S \sqrt{AT \ln(THSA/\delta)}, \\
C_T(\{\bar{\mu}_{M^*}^t, R_{nz}^t - (r_{\max} \cdot \mathbf{1} - r^*)\}_{t=1}^T) & \\
&= 4H\sqrt{T(K\,\mathrm{AdaReg}(T) + D)} + D,
\end{aligned}$$

where $D = \tilde{O}(\sqrt{\beta_T H \dim_E(\mathcal{R}, T^{-1}) T})$.

Proof. We abbreviate $\mu_*^k = \mu_{M^*}^{\pi_*^k}, \bar{\mu}^t = \bar{\mu}_{M^*}^t$. We can use the exact same arguments as in the proof of Theorem 5.1, up until the point where we have to bound

$$\texttt{AgentReg} = \sum_{t=1}^T \langle R_{\pi_*^{k(t)}}(\bar{\mu}^t), \mu_*^{k(t)} - \bar{\mu}^t \rangle.$$

Combining Lemma I.1 and Lemma I.3 we have with probability at least $1 - 3\delta$ that

$$\sum_{t=1}^T \langle R_{\pi_*^{k(t)}}(\bar{\mu}^t), \mu_*^{k(t)} - \bar{\mu}^t \rangle \le \underbrace{\sum_{t=1}^T \langle r^*(\bar{\mu}^t) + R_{\mathrm{nz}}^t(\bar{\mu}^t), \mu_*^{k(t)} - \bar{\mu}^t \rangle}_{\texttt{NewAgentReg}} + 6\sum_{t=1}^T \sum_{h=1}^H w_{\hat{\mathcal{R}}^t}(h, s_h^t, a_h^t, \bar{\mu}^t) + 2r_{\max} H \ln \frac{1}{\delta}.$$

Now, we summarize $x_{H_t+h} = (h, s_h^t, a_h^t, \bar{\mu}_h^t)$, where $H_t = H(t-1)$, and with slight abuse of notation, we rewrite $\hat{r}(x_{H_t+h}) = \hat{r}_h(s_h^t, a_h^t, \bar{\mu}_h^t)$. With this rewriting of notation, we can apply Lemma H.6 with $\varepsilon = T^{-1}$ to show that

$$\sum_{t=1}^T \sum_{h=1}^H w_{\hat{\mathcal{R}}^t}(h, s_h^t, a_h^t, \bar{\mu}^t) \le H + r_{\max} H \dim_E(\mathcal{R}, T^{-1}) + 4\sqrt{\beta_T H \dim_E(\mathcal{R}, T^{-1}) T}.$$

Combining the with the previous results, we get that with probability at least $1 - 3\delta$,

$$\begin{aligned}
\sum_{t=1}^T \langle R_{\pi_*^{k(t)}}(\bar{\mu}^t), \mu_*^{k(t)} - \bar{\mu}^t \rangle &\le \texttt{NewAgentReg} \\
&\quad + \underbrace{6\left(H + r_{\max} H \dim_E(\mathcal{R}, T^{-1}) + 4\sqrt{\beta_T H \dim_E(\mathcal{R}, T^{-1}) T}\right) + 2r_{\max} H \ln \frac{1}{\delta}}_{=:D}.
\end{aligned}$$

The new agent regret term NewAgentReg can be bounded in the same way as in the proof of Theorem 5.1:

$$\sum_{t=1}^T \langle r^*(\bar{\mu}^t) + R_{\mathrm{nz}}^t(\bar{\mu}^t), \mu_*^{k(t)} - \bar{\mu}^t \rangle \le \sum_{k=1}^K \sum_{t=T_{k-1}+1}^{T_k} \langle r^*(\bar{\mu}^t) + R_{\mathrm{nz}}^t(\bar{\mu}^t), \mu_*^k - \bar{\mu}^t \rangle \le K\,\mathrm{AdaReg}(T).$$

For the steering cost, we have for any $\mu \in \Psi_M$,

$$\begin{aligned} C(\mu, R^t_{\text{nz}}) - C(\mu, r_{\max}\mathbf{1} - r^*) &= \langle r^*(\mu) - \bar{r}^t(\mu) + w_{\hat{\mathcal{R}}^t}(\mu) + R_{\pi_*^{k(t)}}(\mu) + \|R_{\pi_*^{k(t)}}(\mu)\|_\infty \mathbf{1}, \mu \rangle \\ &\leq 2\langle w_{\hat{\mathcal{R}}^t}(\mu), \mu \rangle + \langle R_{\pi_*^{k(t)}}(\mu) + \|R_{\pi_*^{k(t)}}(\mu)\|_\infty \mathbf{1}, \mu \rangle \end{aligned}$$

Then, summing over $t = 1, ..., T$,

$$C_T(\{\bar{\mu}^t, R^t_{\text{nz}} - (r_{\max}\mathbf{1} - r^*)\}_{t=1}^T) \leq 2\sum_{t=1}^T \langle w_{\hat{\mathcal{R}}^t}(\bar{\mu}^t), \bar{\mu}^t \rangle + \sum_{t=1}^T \langle R_{\pi_*^{k(t)}}(\bar{\mu}^t) + \|R_{\pi_*^{k(t)}}(\bar{\mu}^t)\|_\infty \mathbf{1}, \bar{\mu}^t \rangle.$$

Using Lemma I.2, we can bound the first term by $2(3\sum_{t=1}^T \sum_{h=1}^H w_{\hat{\mathcal{R}}^t}(x_{H_t+h}) + r_{\max} H \ln(1/\delta))$ with probability at least $1-\delta$. Using Lemma H.6 with $\varepsilon = T^{-1}$, we can further bound this by $D = 2r_{\max} H \ln(1/\delta) + 6(H + r_{\max} H \dim_E(\mathcal{F}, T^{-1}) + 4\sqrt{\beta_T H \dim_E(\mathcal{R}, T^{-1}) T})$. From the steering cost bound in Thm. 5.1 follows that the second term is bounded by

$$4H\sqrt{T\sum_{t=1}^T \langle R_{\pi_*^{k(t)}}(\bar{\mu}^t), \mu_*^{k(t)} - \bar{\mu}^t \rangle},$$

which is at most $4H\sqrt{T(K\,\text{AdaReg}(T) + D)}$, as we have already shown in this proof.

Lastly, we have to discuss the asymptotic bound for D. The term D is dependent on the Eluder dimension of $\mathcal{R}$ and β_T. In Appendix H.1, we show several common function classes with $\dim_E(\mathcal{R}, T^{-1}) \in \tilde{\mathcal{O}}(1)$. Furthermore, if we assume that the functions in $\mathcal{R}$ are parametrized by parameters in some set $\Theta \subset \mathbb{R}^d$ with constant diameter and L-Lipschitz in that parameter, we have $N(\mathcal{R}, \alpha, \|\cdot\|_\infty) \leq N(\Theta, \alpha/L, \|\cdot\|_\infty) \leq (1 + \mathcal{O}(L/\alpha))^d$ Then, we might choose $\alpha = T^{-1}$ such that β_T can also be bounded logarithmically in T. In the cases where the Eluder dimension and β_T are in $\tilde{\mathcal{O}}(1)$, we have $D \in \tilde{\mathcal{O}}(\sqrt{T})$ (ignoring other factors). □

J EXTENSION TO UNKNOWN UTILITY FUNCTION

In this section, we generalize our previous results to the setting where the mediator does not have prior knowledge of the utility function U. We consider non-zero intrinsic reward setting, i.e., Scenario 2 described in Section 3.4. Note that the results for Scenario 1 can be directly derived by setting $r^* = 0$.

Motivation for Unknown Utility Setting This setting makes sense, especially when U partially depends on the agents' intrinsic rewards r^*. As a motivating example, in financial markets, the government (mediator) gains benefits (utility U) from not only the impact on the society by the desired behaviors of the companies (the agents), but also the tax paid by them, which is directly related to the rewards r^* received by agents.[4] Due to the lack of knowledge of r^*, U should only be partially revealed to the mediator. This restricts the applicability of our methods to this setting. However, if U is unknown, we might infer it, for example, by estimating the true reward functions r^* through the online interaction with the agents. We can also generalize this setting as follows.

We consider a general setting, where the mediator does not have prior knowledge on U, but it can observe samples from U, perturbed by σ_U-sub-Gaussian noise, and get access to a function class $\mathcal{U}$ which contains U and whose functions are bounded in $[0, U_{\max}]$.

J.1 Algorithm

We can use the standard technique described in Russo and Van Roy (2013) to handle this case. We define

$$\bar{U}^k = \operatorname*{arg\,min}_{\hat{U} \in \mathcal{U}} \sum_{t=1}^{T_k} (\hat{U}(\bar{\mu}^t) - U(\bar{\mu}^t))^2, \tag{12}$$

$$\hat{\mathcal{U}}^k = \left\{ \hat{U} \in \mathcal{U} : \|\hat{U} - \bar{U}^k\|^2_{2, E_{T_k}} \leq \beta^U_k \right\}, \tag{13}$$

Algorithm 5 Steering reward design for Scenario 2 and unknown utility

```
1: Initialize $\mathcal{P}^1 :=$ set of all possible transition functions, $\pi_*^1$ (arbitrarily), $k = 1, T_0 = 0$.
2: for $t = 1, ..., T$ do
3:     Update $\hat{\mathcal{R}}^t$ as in (10).
4:     Choose $R_{\text{nz}}^t$ as in (11).
5:     Agents play $t$-th game with $r^* + R_{\text{nz}}^t$.
6:     Obtain trajectory $((s_h^t, a_h^t, r_h^t))_{h=1}^H$.
7:     if $\exists (h, s, a),\ s.t.\ n_k(h, s, a) \geq N_k(h, s, a)$ or $t - T_{k-1} \geq T_{epoch}$ then
8:         Update $\mathcal{P}^{k+1}$ as in (7).
9:         $T_k \leftarrow t$; $k \leftarrow k + 1$.
10:        Compute $\hat{\mathcal{U}}^k$ as in (13).
11:        $\hat{U}^k, \pi_*^k, \hat{M}^k \leftarrow \arg\max_{\hat{U} \in \hat{\mathcal{U}}^k, \pi \in \Pi, \hat{M}: \mathbb{P}_{\hat{M}} \in \mathcal{P}^k} \hat{U}(\mu_{\hat{M}}^\pi)$.
12:    end if
13: end for
```

where $\beta_k^U := 8\sigma_U^2 \log(N(\mathcal{U}, \alpha, \|\cdot\|_\infty)/\delta) + 2\alpha k(8U_{\max} + \sqrt{8\sigma_U^2 \ln(4k^2/\delta)})$ and, e.g., $\alpha = T^{-1}$.

Algorithm 5 differs from Algorithm 4 in the if-condition in line 7 as well as in lines 10 and 11.

The if-condition in line 7 now includes the case $t - T_{k-1} \geq T_{epoch}$, where T_{epoch} will be chosen later. We need this to guarantee $T_k - T_{k-1} \leq T_{epoch}$ for all k and thereby bound the estimation error of the utility function estimate. Intuitively, we need to keep the estimates of U somewhat up to date to be able to bound the estimation error. Meanwhile, we cannot update the estimate in each round (or too often) since then we would also have to change π_*^k in each round, which would lead to $K = T$.

Since we also need to estimate the utility function, we changed line 11 to also compute an optimistic estimate of the utility using the definition in (13).

J.2 Analysis

Theorem J.1. *Under Assump. A, B and C, if we run Alg. 5 with* $0 < \delta < 1$*, then with probability at least* $1 - 8\delta$*,* $K \leq T^{1/6} + HSA \log_2 T$*, and*

$$\begin{aligned}\Delta_T(\{\bar{\mu}_{M^*}^t\}_{t=1}^T) \leq{}& L_U \sqrt{H^3 SAT(K\,\text{AdaReg}(T) + D)} + 36 L_U H^3 S \sqrt{AT} \ln(THSA/\delta) \\ &+ \mathcal{O}\left(T^{5/6} U_{\max} \dim_E(\mathcal{U}, T^{-1}) + \sqrt{\beta_K^U \dim_E(\mathcal{U}, T^{-1}) T}\right), \\ C_T(\{\bar{\mu}_{M^*}^t, R_{nz}^t&-(r_{\max} \cdot \mathbf{1} - r^*)\}_{t=1}^T) \\ &= 4H\sqrt{T(K\,\text{AdaReg}(T) + D)} + D,\end{aligned}$$

where $D = \tilde{O}(\sqrt{\beta_T H \dim_E(\mathcal{R}, T^{-1}) T})$.

Comparing with Theorem 6.1, we see that the steering gap has an additional term originating from the estimation of U. Furthermore, the bound of the number of epochs K has an additional $T^{1/6}$. Similar to the discussion about $\mathcal{R}$ in Section H.1, we can also bound β_K^U and $\dim_E(\mathcal{U}, T^{-1})$ under suitable assumptions about $\mathcal{U}$. If $\beta_K^U, \dim_E(\mathcal{U}, T^{-1}) \in \tilde{\mathcal{O}}(1)$ and $\text{AdaReg}(T) = \tilde{\mathcal{O}}(\sqrt{T})$, both the steering cap and steering cost are in $\tilde{\mathcal{O}}(T^{5/6})$ (ignoring all other constants).

Proof. We can adapt the proof of Theorem 6.1 by choosing the following regret decomposition.

$$U(\mu^{\pi^*}) - U(\bar{\mu}^t) = \left(U(\mu^{\pi^*}) - \hat{U}^k(\hat{\mu}_*^k)\right) + \left(\hat{U}^k(\hat{\mu}_*^k) - \hat{U}^k(\bar{\mu}^t)\right) + \left(\hat{U}^k(\bar{\mu}^t) - U(\bar{\mu}^t)\right)$$

[4] Another way to interpret this scenario is that the mediator's utility $U = \alpha U_{\text{mediator}} + (1 - \alpha) U_{\text{agents};r^*}$ can be decomposed to a known function U_{mediator} representing its intrinsic utility, and another unknown part $U_{\text{agents};r^*}$, which reflects the agents' interests and depends on r^*. Here α serves as a parameter to trade-off the interests between two parties.

Using Lemma I.1 (and replacing $\mathcal{R}$ by $\mathcal{U}$ in the Lemma), we have $U \in \bigcap_{k=1}^K \hat{\mathcal{U}}^k$ with probability at least $1-2\delta$. Thus, with probability at least $1-2\delta$, the first term can be bounded by 0 using optimism. The second term can be bounded in the same way as in the proof of Theorem 6.1. Summing over all t, the last term accumulates to

$$\sum_{t=1}^{T} \hat{U}^k(\bar{\mu}^t) - U(\bar{\mu}^t) = \sum_{k=1}^{K} \sum_{t=T_{k-1}+1}^{T_k} \hat{U}^k(\bar{\mu}^t) - U(\bar{\mu}^t) \le \sum_{k=1}^{K} \sum_{t=T_{k-1}+1}^{T_k} w_{\hat{\mathcal{U}}^k}(\bar{\mu}^t).$$

Using the fact that $T_k - T_{k-1} \le T_{epoch}$ and Lemma H.6 with $\varepsilon = T^{-1}$, the sum above is at most

$$\frac{KT_{epoch}}{T} + U_{\max} T_{epoch} \dim_E(\mathcal{U}, T^{-1}) + 4\sqrt{\beta_T^U K T_{epoch} \dim_E(\mathcal{U}, T^{-1})}.$$

We also have to find a new bound for K. As before, we can enter the if block at most $HSA \log_2 T$ times because of the first condition. In addition, we can enter the if block at most T/T_{epoch} times due to the condition $T_k \ge T_{epoch}$. Therefore, $K \le T/T_{epoch} + HSA \log_2 T$ and $KT_{epoch} \le T + HSAT_{epoch} \log_2 T$.

Now, we set $T_{epoch} = T^{5/6}$. Then, $K \le T^{1/6} + HSA \log_2 T$ and

$$\sum_{t=1}^{T} \hat{U}^k(\bar{\mu}^t) - U(\bar{\mu}^t) \le \mathcal{O}\left(T^{5/6} U_{\max} \dim_E(\mathcal{U}, T^{-1}) + \sqrt{\beta_T^U \dim_E(\mathcal{U}, T^{-1}) T} \right).$$

Finally, the steering gap is the previous bound of Theorem 6.1 plus the above term.

With regard to the steering cost, the only change is the bound of K.

□