Accommodating Picky Customers: Regret Bound and Exploration Complexity for Multi-Objective Reinforcement Learning

Jingfeng Wu
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218
uuujf@jhu.edu

Vladimir Braverman
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218
vova@cs.jhu.edu

Lin F. Yang
Department of Electrical and Computer Engineering
University of California, Los Angeles
Los Angeles, CA 90095
linyang@ee.ucla.edu

Abstract

In this paper we consider multi-objective reinforcement learning where the objectives are balanced using preferences. In practice, the preferences are often given in an adversarial manner, e.g., customers can be picky in many applications. We formalize this problem as an episodic learning problem on a Markov decision process, where transitions are unknown and a reward function is the inner product of a preference vector with pre-specified multi-objective reward functions. We consider two settings. In the online setting, the agent receives a (adversarial) preference every episode and proposes policies to interact with the environment. We provide a model-based algorithm that achieves a nearly minimax optimal regret bound $\widetilde{\mathcal{O}}\big(\sqrt{\min\{d, S\} \cdot H^2 SAK}\big)$, where d is the number of objectives, S is the number of states, A is the number of actions, H is the length of the horizon, and K is the number of episodes. Furthermore, we consider preference-free exploration, i.e., the agent first interacts with the environment without specifying any preference and then is able to accommodate arbitrary preference vector up to ϵ error. Our proposed algorithm is provably efficient with a nearly optimal trajectory complexity $\widetilde{\mathcal{O}}\big(\min\{d, S\} \cdot H^3 SA/\epsilon^2\big)$. This result partly resolves an open problem raised by Jin et al. [2020].

1 Introduction

In single-objective reinforcement learning (RL), a *scalar reward* is pre-specified and an agent learns a policy to maximize the long-term cumulative reward [Sutton and Barto, 2018]. However, in many real-world applications, we need to optimize multiple objectives for the same (unknown) environment, even when these objectives are possibly contradicting [Roijers et al., 2013]. For example, in an autonomous driving application, each passenger may have a different preference of driving styles: some of the passengers prefer a very steady riding experience while other passengers enjoy the fast acceleration of the car. Therefore, traditional single-objective RL approach may fail to be applied in such scenarios. One way to tackle this issue is the *multi-objective reinforcement learning* (MORL) [Roijers et al., 2013, Yang et al., 2019, Natarajan and Tadepalli, 2005, Abels et al., 2018]

method, which models the multiple objectives by a *vectorized reward*, and an additional *preference vector* to specify the relative importance of each objective. The agent of MORL needs to find policies to optimize the cumulative preference-weighted rewards.

If the preference vector is fixed or drawn from a fixed distribution, MORL is no more challenging than single-objective RL, as we can predict the objective to be optimized and apply (variants of) single-objective RL algorithms (e.g., [Azar et al., 2017]). However, more often in practice, the preference vector (under which the weighted objective needs to be optimized) is:

  (i) adversarially provided,
  (ii) or even not available in the learning phase.

Once more taking the autonomous driving application as example: (i) the intelligent system needs to adapt driving configurations to accommodate every customer, even though the next customer can have picky preference that is unpredictable; (ii) when the intelligent system is under development, the future customer cannot be known in advance, nonetheless, when the system is deployed, it has to be capable to accommodate any potential customer who rides the car. Such MORL examples are common, to name a few beyond the autonomous driving one: medical treatment must take care of every patient even in very rare health conditions; an education system should accommodate every student according to his/her own characteristics; and emergency response systems have to be responsible in all extreme cases. Due to these exclusive challenges that have not appeared in single-objective RL, new *sample-efficient* algorithms for MORL need to be developed, as well as their *theoretical grounds* need to be established.

In this work, we study *provable* sample-efficient algorithms for MORL that resolve the aforementioned issues. Specifically, we consider MORL on a finite-horizon Markov decision process (MDP) with an unknown transition kernel, S states, A actions, H steps, and d reward functions that represent the d difference objectives. We investigate MORL problems in two paradigms: (i) *online MORL* (where the preferences are adversarially presented), and (ii) *preference-free exploration* (where the preferences are not available in the learning/exploration phase). The two settings and our contributions are explained respectively in the following.

Setting (i): Online MORL. We first consider an online learning problem to capture the challenge that preference vectors could be adversarially provided in MORL. In the beginning of each episode, the MORL agent is provided (potentially adversarially) with a preference vector, and the agent interacts with the unknown environment to collect data and rewards (that is specified by the provided preference). The performance of an algorithm is measured by the *regret*, i.e., the difference between the rewards collected by the agent and those would be collected by a theoretically optimal agent (who could use varying policies that adapt to the preferences). This setting generalizes the classical online single-objective RL problem [Azar et al., 2017] (where $d = 1$).

Contribution (i). For online MORL, we provide a provably efficient algorithm with a regret upper bound $\widetilde{\mathcal{O}}\big(\sqrt{\min\{d, S\} \cdot H^2 SAK}\big)$[1], where K is the number of episodes for interacting with the environment. Furthermore, we show an information-theoretic lower bound $\Omega\big(\sqrt{\min\{d, S\} \cdot H^2 SAK}\big)$ for online MORL. These bounds together show that, ignoring logarithmic factors, our algorithm resolves the online MORL problems optimally.

Setting (ii): Preference-Free Exploration. We further consider an unsupervised MORL problem to capture the issue that preferences could be hard to obtain in the training phase. The MORL agent first interacts with the unknown environment in the absence of preference vectors; afterwards, the agent is required to use the collected data to compute near-optimal policies for an arbitrarily specified preference vector. The performance is measured by the *sample complexity*, i.e., the minimum amount of trajectories that an MORL agent needs to collect during exploration in order to be near-optimal during planning. This setting extends the recent proposed *reward-free exploration* problem [Jin et al., 2020] to MORL.

Contribution (ii). For preference-free exploration, we show that a simple variant of the proposed online algorithm can achieve nearly optimal sample complexity. In particular, the algorithm achieves a sample complexity upper bound $\widetilde{\mathcal{O}}\big(\min\{d, S\} \cdot H^3 SA/\epsilon^2\big)$ where ϵ is the tolerance of the planning error; and we also show a sample complexity lower bound, $\Omega\big(\min\{d, S\} \cdot H^2 SA/\epsilon^2\big)$, for any

---

[1]We use $\widetilde{\mathcal{O}}(\cdot)$ to hide (potential) polylogarithmic factors in $\mathcal{O}(\cdot)$, i.e., $\widetilde{\mathcal{O}}(n) := \mathcal{O}(n \log^k n)$ for sufficiently large n and some absolute constant $k > 0$.

algorithm. These bounds suggest that our algorithm is optimal in terms of d, S, A, ϵ up to logarithmic factors. It is also worth noting that our results for preference-free exploration partly answer an open question raised by Jin et al. [2020]: reward-free RL is easier when the unknown reward functions enjoy low-dimensional representations (as in MORL).

Paper Layout. The remaining paper is organized as follows: the preliminaries are summarized in Section 2; then in Section 3, we formally introduce the problem of online MORL, our algorithm and its regret upper and lower bounds; then in Section 4, we turn to study the preference-free exploration problem in MORL, where we present an exploration algorithm with sample complexity analysis (with both upper bound and lower bound), and compare our results with existing results for related problems; finally, the related literature is reviewed in Section 5 and the paper is concluded by Section 6.

2 Preliminaries

We specify a finite-horizon Markov decision process (MDP) by a tuple of $(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, \boldsymbol{r}, \mathcal{W})$. $\mathcal{S}$ is a finite *state* set where $|\mathcal{S}| = S$. $\mathcal{A}$ is a finite *action* set where $|\mathcal{A}| = A$. H is the length of the horizon. $\mathbb{P}(\cdot \mid x, a)$ is a *stationary, unknown transition probability* to a new state for taking action a at state x. $\boldsymbol{r} = \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_H\}$, where $\boldsymbol{r}_h : \mathcal{S} \times \mathcal{A} \to [0, 1]^d$ represents a d-dimensional *vector rewards function* that captures the d objectives[2]. $\mathcal{W} := \{w \in [0, 1]^d, \|w\|_1 = 1\}$ specifies the set of all possible *preference vectors*[3], where each preference vector $w \in \mathcal{W}$ induces a scalar reward function by[4] $r_h(\cdot, \cdot) = \langle w, \boldsymbol{r}_h(\cdot, \cdot)\rangle$ for $h = 1, \ldots, H$. A *policy* is represented by $\pi := \{\pi_1, \ldots, \pi_H\}$, where each $\pi_h(\cdot)$ maps a state to a distribution over the action set. Fixing a policy π, we will consider the following generalized *Q-value function* and generalized *value function*:

$$Q_h^\pi(x, a; w) := \mathbb{E}\big[\textstyle\sum_{j=h}^H \langle w, \boldsymbol{r}_j(x_j, a_j)\rangle \big| x_h = x, a_h = a\big], \quad V_h^\pi(x; w) := Q_h^\pi(x, \pi_h(x); w),$$

where $x_j \sim \mathbb{P}(\cdot \mid x_{j-1}, a_{j-1})$ and $a_j \sim \pi_j(x_j)$ for $j > h$. Note that compared with the typical Q-value function (or the value function) used in single-objective RL, here the generalized Q-value function (or the generalized value function) takes the preference vector as an additional input, besides the state-action pair. Fixing a preference $w \in \mathcal{W}$, the optimal policy under w is defined as $\pi_w^* := \arg\max_\pi V_1^\pi(x_1; w)$. For simplicity, we denote

$$Q_h^*(x, a; w) := Q_h^{\pi_w^*}(x, a; w), \quad V_h^*(x; w) := V_h^{\pi_w^*}(x; w) = \max_a Q_h^*(x, a; w).$$

The following abbreviation is adopted for simplicity:

$$\mathbb{P}V_h^\pi\big|_{x,a,w} := \textstyle\sum_{y \in \mathcal{S}} \mathbb{P}(y|x, a) V_h^\pi(y; w),$$

and similar abbreviations will be adopted for variants of probability transition (e.g., the empirical transition probability) and variants of value function (e.g., the estimated value function). Finally, we remark that the following Bellman equations hold for the generalized Q-value functions

$$\begin{cases} Q_h^\pi(x, a; w) = \langle w, \boldsymbol{r}_h(x, a)\rangle + \mathbb{P}V_{h+1}^\pi\big|_{x,a,w}, \\ V_h^\pi(x; w) = Q_h^\pi(x, \pi_h(x); w); \end{cases} \quad \begin{cases} Q_h^*(x, a; w) = \langle w, \boldsymbol{r}_h(x, a)\rangle + \mathbb{P}V_{h+1}^*\big|_{x,a,w}, \\ V_h^*(x; w) = \max_a Q_h^*(x, a; w). \end{cases}$$

With the above preparations, we are ready to discuss our algorithms and theory for online MORL (Section 3) and preference-free exploration (Section 4).

3 Online MORL

Problem Setups. The online setting captures the first difficulty in MORL, where the preference vectors can be adversarially provided to the MORL agent. Formally, the MORL agent interacts with

---

[2]For the sake of presentation, we discuss bounded deterministic reward functions in this work. The techniques can be readily extended to stochastic reward settings.

[3]The condition that $w \in [0, 1]^d$ is only assumed for convenience. Our results naturally generalize to preference vectors that are entry-wisely bounded by absolute constants.

[4]The linear scalarization method can be generalized. See more discussions in Section 3.2, Remark 2.

an unknown environment through Protocol 1: at the beginning of the k-th episode, an adversary selects a preference vector w^k and reveals it to the agent; then starting from a fixed initial state[5] x_1, the agent draws a trajectory from the environment by recursively taking an action and observing a new state, and collects rewards from the trajectory, where the rewards are scalarized from the vector rewards by the given preference. The agent's goal is to find the policies that maximize the cumulative rewards. Its performance will be measured by the following *regret*: suppose the MORL agent has interacted with the environment through Protocol 1 for K episodes, where the provided preferences are $\{w^k\}_{k=1}^K$ and the adopted policies are $\{\pi^k\}_{k=1}^K$ correspondingly, we consider the regret of the collected rewards (in expectation) competing with the theoretically maximum collected rewards (in expectation):

$$\texttt{regret}(K) := \sum_{k=1}^K V_1^*(x_1; w^k) - V_1^{\pi^k}(x_1; w^k). \tag{1}$$

Clearly, the regret (1) is always non-negative, and a smaller regret implies a better online performance. We would like to highlight that the regret (1) allows the theoretically optimal agent to adopt varying policies that adapt to the preferences.

Protocol 1 Environment Interaction Protocol
Require: MDP$(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, \boldsymbol{r}, \mathcal{W})$ and online agent
1: agent observes $(\mathcal{S}, \mathcal{A}, H, \boldsymbol{r}, \mathcal{W})$
2: **for** episode $k = 1, 2, \ldots, K$ **do**
3:     agent receives an initial state $x_1^k = x_1$, and a preference w^k (from an adversary)
4:     **for** step $h = 1, 2, \ldots, H$ **do**
5:         agent chooses an action a_h^k, and collects reward $\langle w^k, \boldsymbol{r}_h(x_h^k, a_h^k) \rangle$
6:         agent transits to a new state $x_{h+1}^k \sim \mathbb{P}(\cdot \mid x_h^k, a_h^k)$
7:     **end for**
8: **end for**

Connections to Online Single-Objective RL. The online regret minimization problems are well investigated in the context of single-objective RL (see, e.g., [Azar et al., 2017, Jin et al., 2018]). In both online single-objective RL and our studied online MORL, it is typical to assume that the transition probability $\mathbb{P}$ is the only unknown information about the environment since estimating a stochastic reward is relatively easy (see, e.g., [Azar et al., 2017, Jin et al., 2018]). In particular, single-objective RL is a special case of MORL in the online setting, where the preference is fixed during the entire learning process (i.e., $w^k := w$ for all k) — therefore an online MORL algorithm naturally applies to single-objective RL. However, the reverse is not true as that in MORL the preference vectors can change over time and are potentially adversarial.

Figure 1: A regret comparison of `MO-UCBVI` vs. the best-in-hindsight policy in a simulated random multi-objective MDP. Note that the best-in-hindsight policy is the optimal policy for single-objective RL. The plots show that the best-in-hindsight policy will incur linear regret in online MORL, and the proposed `MO-UCBVI` achieves sublinear regret as predicted by Theorem 1. See Appendix A for details.

Comparison with Single-Objective Stochastic and Adversarial Reward Setting. The essential difficulty of MORL is further reflected in the regret (1). Specifically, the regret (1) compares the performance of the agent's policies to a sequence of *optimal* policy under *each* given preference, which could *vary* over time. However, in online single-objective RL, either with known/stochastic rewards [Azar et al., 2017, Jin et al., 2018] or adversarial rewards [Rosenberg and Mansour, 2019, Jin et al., 2019], the benchmark policy is supposed to be *fixed* over time (the best policy in the hindsight). This difference suggests that online MORL could be more challenging than online single-objective RL, as a harder performance measurement is adopted. More specifically, when measured by regret
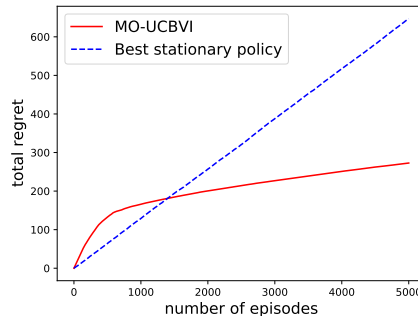
(1), existing algorithms for single-objective RL (e.g., [Azar et al., 2017]) easily suffer a $\propto \Theta(K)$ regret when the sequence of preferences is adversarially designed; in contrast, we will show an MORL algorithm that experiences at most $\propto \widetilde{\mathcal{O}}(\sqrt{K})$ regret under any sequence of preferences. A numerical simulation of this issue is presented in Figure 1.

3.1 A Sample-Efficient Online Algorithm

In this part we introduce an algorithm for online MORL, called *multi-objective upper confidence bound value iteration* (`MO-UCBVI`). A simplified verison of `MO-UCBVI` is presented as Algorithm 1, and a more advanced version (Algorithm 4) can be found in Appendix B. Our algorithm is inspired by `UCBVI` [Azar et al., 2017] that achieves minimax regret bound in online single-objective RL.

Algorithm 1 `MO-UCBVI`

1: initialize history $\mathcal{H}^0 = \emptyset$
2: **for** episode $k = 1, 2, \ldots, K$ **do**
3:      $N^k(x, a), \widehat{\mathbb{P}}^k(y \mid x, a) \leftarrow$ `Empi-Prob`$(\mathcal{H}^{k-1})$
4:      compute bonus $b^k(x, a) := c \cdot \sqrt{\frac{\min\{d, S\}H^2\iota}{N^k(x,a)}}$, where $\iota = \log(HSAK/\delta)$ and c is a constant
5:      receive a preference w^k
6:      $\{Q_h^k(x, a; w^k)\}_{h=1}^H \leftarrow$ `UCB-Q-Value`$(\widehat{\mathbb{P}}^k, w^k, b^k)$
7:      receive initial state $x_1^k = x_1$
8:      **for** step $h = 1, 2, \ldots, H$ **do**
9:          take action $a_h^k = \arg\max_a Q_h^k(x_h^k, a; w^k)$, and obtain a new state x_{h+1}^k
10:      **end for**
11:      update history $\mathcal{H}^k = \mathcal{H}^{k-1} \cup \{x_h^k, a_h^k\}_{h=1}^H$
12: **end for**

13: **Function** `Empi-Prob`
14: **Require:** history $\mathcal{H}^{k-1}$
15: **for** $(x, a, y) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ **do**
16:      $N^k(x, a, y) := \#\{(x, a, y) \in \mathcal{H}^{k-1}\}$, and $N^k(x, a) := \sum_y N^k(x, a, y)$
17:      **if** $N^k(x, a) > 0$ **then**
18:          $\widehat{\mathbb{P}}^k(y \mid x, a) = N^k(x, a, y)/N^k(x, a)$
19:      **else**
20:          $\widehat{\mathbb{P}}^k(y \mid x, a) = 1/S$
21:      **end if**
22: **end for**
23: **return** $N^k(x, a)$ and $\widehat{\mathbb{P}}^k(y \mid x, a)$

24: **Function** `UCB-Q-Value`
25: **Require:** empirical transition $\widehat{\mathbb{P}}^k$, preference w^k, and bonus $b^k(x, a)$
26: set $V_{H+1}^k(x; w^k) = 0$
27: **for** step $h = H, H-1, \ldots, 1$ **do**
28:      **for** $(x, a) \in \mathcal{S} \times \mathcal{A}$ **do**
29:          $Q_h^k(x, a; w^k) = \min\left\{H, \langle w^k, \boldsymbol{r}_h(x, a)\rangle + b^k(x, a) + \widehat{\mathbb{P}}_h^k V_{h+1}^k\big|_{x,a,w^k}\right\}$
30:          $V_h^k(x; w^k) = \max_{a \in \mathcal{A}} Q_h^k(x, a; w^k)$
31:      **end for**
32: **end for**
33: **return** $\left\{Q_h^k(x, a; w^k)\right\}_{h=1}^H$

In Algorithm 1, the agent interacts with the environment according to Protocol 1, and use an *optimistic policy* to explore the unknown environment and collect rewards. The optimistic policy is a greedy policy with respect to an optimistic estimation to the value function, `UCB-Q-Value` (lines 6 and 9). Specifically, `UCB-Q-Value` is constructed to maximize the cumulative reward, which is scalarized by the current preference, through dynamic programming over an empirical transition probability (line 24). The empirical transition probability is inferred from the data collected so far (lines 3 and

13), which might not be accurate if a state-action pair has not yet been visited for sufficient times. To mitigate this inaccuracy, `UCB-Q-Value` utilizes an extra *exploration bonus* (lines 4 and 29) so that: (i) `UCB-Q-Value` never under-estimates the optimal true value function, for *whatever preference vector* (and with high probability); and (ii) the added bonus *shrinks quickly* enough (as the corresponding state-action pair continuously being visited) so that the over-estimation is under control. The overall consequence is that: `MO-UCBVI` explores the environment via a sequence of optimistic policies, in order to collect rewards under a sequence of adversarially provided preferences; since the policies are optimistic for any preference, the incurred regret would not exceed the total amount of added bonus; and since the bonus decays fast, their sum up would be sublinear (with respect to the number of episodes played). Therefore `MO-UCBVI` only suffers a sublinear regret, even when the preferences are adversarially presented. The above intuition is formalized in the next part.

3.2 Theoretical Analysis

The next two theorems justify regret upper bounds for `MO-UCBVI` and a regret lower bound for online MORL problems, respectively.

Theorem 1 (Regret bounds for `MO-UCBVI`). *Suppose K is sufficiently large. Then for any sequence of the incoming preferences $\{w^k\}_{k=1}^{K}$, with probability at least $1 - \delta$:*

- *the regret (1) of* `MO-UCBVI` *(see Algorithm 1) satisfies*

$$\texttt{regret}(K) \leq \mathcal{O}\big(\sqrt{\min\{d, S\} \cdot H^3 SAK \log(HSAK/\delta)}\big);$$

- *and the regret (1) of a Bernstein-variant* `MO-UCBVI` *(see Algorithm 4 in Appendix B) satisfies*

$$\texttt{regret}(K) \leq \mathcal{O}\big(\sqrt{\min\{d, S\} \cdot H^2 SAK \log^2(HSAK/\delta)}\big).$$

Theorem 2 (A regret lower bound for MORL). *There exist some absolute constants $c, K_0 > 0$, such that for any $K > K_0$, any MORL algorithm that runs K episodes, there is a set of MOMDPs and a sequence of (necessarily adversarially chosen) preferences vectors such that*

$$\mathbb{E}[\texttt{regret}(K)] \geq c \cdot \sqrt{\min\{d, S\} \cdot H^2 SAK},$$

where the expectation is taken with respect to the randomness of choosing MOMDPs and the randomness of the algorithm for collecting dataset.

Remark 1. When $d = 1$, MORL recovers the single-objective RL setting, and Theorem 1 recovers existing nearly minimax regret bounds for single-objective RL [Azar et al., 2017, Zanette and Brunskill, 2019]. Moreover, the lower bound in Theorem 2 implies that our upper bound in Theorem 1 is tight ignoring logarithm terms. Interestingly, the lower bound suggests MORL with $d > 2$ is truly harder than single objective RL (corresponding to $d = 1$) as the sequence of preferences can be adversarially chosen.

Remark 2. Theorem 1 (as well as our other theorems) applies to general scalarization methods besides the linear one as adopted by Algorithm 1 and many other MORL papers [Yang et al., 2019, Roijers et al., 2013, Natarajan and Tadepalli, 2005, Abels et al., 2018]. In particular, our results apply to scalarization functions $r_h(\cdot, \cdot) = f(\boldsymbol{r}_h(\cdot, \cdot); w)$ that are (1) deterministic, (2) Lipschitz continuous for w, and (3) bounded between $[0, 1]$ (which can be relaxed). This will be clear from proofs in Appendix B, where we treat the potentially adversarially given preferences by a covering argument and union bound, and these techniques are not dedicated to linear scalarization function and can be easily extended to more general cases.

The proof of Theorem 1 leverages standard analysis procedures for single-objective RL [Azar et al., 2017, Zanette and Brunskill, 2019], and a covering argument with an union bound to tackle the challenge of adversarial preferences. Th rigorous proof is included in Appendix B.

Specifying an adversarial process of providing preferences is the key challenge for proving Theorem 2. To handle this issue, we use reduction techniques and utilize a lower bound that we will show shortly for preference-free exploration problems. We refer the readers to Theorem 4 and Appendix E for more details.

4 Preference-Free Exploration

Problem Setups. *Preference-free exploration* (PFE) captures the second difficulty in MORL: the preference vector might not be observable when the agent explores the environment. Specifically, PFE

6

consists of an exploration phase and a planning phase. Similarly as in the online setting, the transition probability is hidden from the MORL agent. In the exploration phase, the agent interacts with the unknown environment to collect samples, however the agent has no information about the preference vectors at this point. Afterwards PFE switches to the planning phase, where the agent is prohibited to obtain new data, and is required to compute near-optimal policy for any preference-weighted reward functions. Since this task is no longer in an online fashion, we turn to measure the performance of a PFE algorithm by the minimum number of required trajectories (in the exploration phase) so that the algorithm can behave near-optimally in the planning phase. This is made formal as follows: a PFE algorithm is called (ϵ, δ)-PAC (Probably Approximately Correct), if

$$\mathbb{P}\big\{\forall w \in \mathcal{W}, \ V_1^*(x_1; w) - V_1^{\pi_w}(x_1; w) \leq \epsilon\big\} \geq 1 - \delta,$$

where π_w is the policy outputted by the PFE algorithm for input preference w, then the *sample complexity* of a PFE algorithm is defined by the least amount of trajectories it needs to collect in the exploration phase for being (ϵ, δ)-PAC in the planning phase.

Connections to Reward-Free Exploration. PFE problem is a natural extension to the recent proposed *reward-free exploration* (RFE) problem [Jin et al., 2020, Kaufmann et al., 2020, Wang et al., 2020, Ménard et al., 2020, Zhang et al., 2020b]. Both problems consist of an exploration phase and a planning phase; the difference is in the planning phase: in RFE, the agent needs to be able to compute near-optimal policies for *all reward functions*, while in PFE, the agent only needs to achieve that for *all preference-weighted reward functions*, i.e., the reward functions that can be represented as the inner product of a d-dimensional preference vectors and the d-dimensional vector rewards functions (i.e., the d objectives in MORL). A PFE problem reduces to a RFE problem if $d = SA$ such that every reward function can be represented as a preference-weighted reward function. However, if $d \ll SA$, it is conjectured by Jin et al. [2020] that PFE can be solved with a much smaller sample complexity than RFE. Indeed, in the following part we show an algorithm that improves a $\propto \widetilde{\mathcal{O}}(S^2)$ dependence for RFE to $\propto \widetilde{\mathcal{O}}(\min\{d, S\} \cdot S)$ dependence for PFE, in terms of sample complexity.

4.1 A Sample-Efficient Exploration Algorithm

We now present a simple variant of `MO-UCBVI` that is sample-efficient for PFE. The algorithm is called *preference-free upper confidence bound exploration* (`PF-UCB`), and is discussed separately as in the exploration phase (Algorithm 2) and in the planning phase (Algorithm 3) in below.

Algorithm 2 `PF-UCB` (Exploration)

1: initialize history $\mathcal{H}^0 = \emptyset$
2: **for** episode $k = 1, 2, \ldots, K$ **do**
3:    $N^k(x, a), \widehat{\mathbb{P}}^k(y \mid x, a) \leftarrow$ `Empi-Prob`$(\mathcal{H}^{k-1})$
4:    compute bonus $c^k(x, a) := \frac{H^2 S}{2N^k(x,a)} + 2b^k(x, a)$ for $b^k(x, a)$ defined in Algorithms 1 or 3
5:    $\{\overline{Q}_h^k(x, a)\}_{h=1}^H \leftarrow$ `UCB-Q-Value`$(\widehat{\mathbb{P}}^k, w = 0, c^k)$
6:    receive initial state $x_1^k = x_1$
7:    **for** step $h = 1, 2, \ldots, H$ **do**
8:       take action $a_h^k = \arg\max_a \overline{Q}_h^k(x_h^k, a)$, and obtain a new state x_{h+1}^k
9:    **end for**
10:   update history $\mathcal{H}^k = \mathcal{H}^{k-1} \cup \{x_h^k, a_h^k\}_{h=1}^H$
11: **end for**

Algorithm 2 presents our PFE algorithm in the exploration phase. Indeed, Algorithm 2 is a modified `MO-UCBVI` (Algorithm 1) by setting the preference to be zero (Algorithm 2, line 5), and slightly enlarging the exploration bonus (Algorithm 2, line 4). The intention of an increased exploration bonus will be made clear later when we discuss the planning phase. With a zero preference vector, the `UCB-Q-Value` in Algorithm 2 will identify a trajectory along which the cumulative bonus (instead of the cumulative rewards) is maximized (with respect to the empirical transition probability). Also note that the bonus function (Algorithm 2, line 4) is negatively correlated with the number of visits to a state-action pair. Hence the greedy policy with respect to the zero-preference `UCB-Q-Value` tends to visit the state-actions pairs that are associated with large bonus, i.e., those that have been visited

---

Algorithm 3 `PF-UCB` (Planning)

Require: history $\mathcal{H}^K$, preference vector w

1: **for** $k = 1, 2, \ldots, K$ **do**

2:     $N^k(x, a), \widehat{\mathbb{P}}^k(y|x, a) \leftarrow$ `Empi-Prob`$(\mathcal{H}^{k-1})$

3:     compute bonus $b^k(x, a) := c \cdot \sqrt{\frac{\min\{d, S\} H^2 \iota}{N^k(x, a)}}$ where $\iota = \log(HSAK/\delta)$ and c is a constant

4:     $\{Q_h^k(\cdot, \cdot; w)\}_{h=1}^H \leftarrow$ `UCB-Q-Value`$(\widehat{\mathbb{P}}^k, w, b^k)$

5:     infer greedy policy $\pi_h^k(x) = \arg\max_a Q_h^k(x, a; w)$

6: **end for**

7: **return** π drawn uniformly from $\{\pi^k\}_{k=1}^K$

---

for less times. In sum, Algorithm 2 explores the unknown environment "uniformly", without the guidance of preference vectors.

Then Algorithm 3 shows our PFE algorithm in the planning phase. Given any preference vector, Algorithm 3 computes a sequence of optimistically estimated value functions based on the data collected from the exploration phase, and then outputs a greedy policy with respect to a randomly drawn optimistic value estimation. Note that the bonus in Algorithm 3 is set as the one in Algorithm 1, and recall that the bonus in Algorithm 2 is an enlarged one. The relatively large bonus in the exploration phase guarantees that the regret in the planning phase never exceeds that in the exploration phase. On the other hand, based on Theorem 1 for `MO-UCBVI` (Algorithm 1), we have already known that the exploration algorithm (Algorithm 2), a modified Algorithm 1, suffers at most $\widetilde{\mathcal{O}}(\sqrt{K})$ regret, hence the planning algorithm (Algorithm 3) experiences at most $\widetilde{\mathcal{O}}(1/\sqrt{K})$ error. The next part rigorously justifies these discussions.

4.2 Theoretic Analysis

We first provide Theorem 3 to justify the trajectory complexity of Algorithms 2 and 3 in the PFE setting; then we present Theorem 4 that gives an information-theoretic lower bound on the trajectory complexity for any PFE algorithm.

Theorem 3 (A trajectory complexity of `PF-UCB`). *Suppose $\epsilon > 0$ is sufficiently small. Then for* `PF-UCB` *(Algorithm 2) run for*

$$K = \mathcal{O}\big(\min\{d, S\} \cdot H^3 SA\iota/\epsilon^2\big), \quad \text{where } \iota := \log(HSA/(\delta\epsilon)),$$

episodes, `PF-UCB` *(Algorithm 3) is (ϵ, δ)-PAC.*

Theorem 4 (A lower bound for PFE). *There exist absolute constants $c, \epsilon_0 > 0$, such that for any $0 < \epsilon < \epsilon_0$, there exists a set of MOMDPs such that any PFE algorithm that is $(\epsilon, 0.1)$-PAC on them, it needs to collect at least*

$$K \geq \Omega\big(\min\{d, S\} \cdot H^2 SA/\epsilon^2\big)$$

trajectories in expectation (with respect to the randomness of choosing MOMDPs and the exploration algorithm).

Remark 3. According to Theorems 3 and 4, the trajectory complexity of `PF-UCB` is optimal for d, S, A, ϵ ignoring logarithmic factors, but is an H factor loose compared with the lower bound. This is because the current Algorithm 2 utilizes a *preference-independent*, Hoeffding-type bonus since the preference vector is not available during exploration. We leave it as an open problem to further remove this gap about H.

Proof Sketch of Theorem 3. Theorem 3 is obtained in three procedures. (1) We first observe the total regret incurred by Algorithm 2 is $\widetilde{\mathcal{O}}\big(\sqrt{\min\{d, S\} H^3 SAK}\big)$ according to Theorem 1. (2) Then utilizing the enlarged exploration bonus, we show that in each episode, the planning error is at most constant times of the incurred exploration error. (3) Thus the averaged planning error is at most $\widetilde{\mathcal{O}}\big(\sqrt{\min\{d, S\} H^3 SA/K}\big)$ as claimed. A complete proof is deferred to Appendix C.

Note that the second argument is motivated by [Zhang et al., 2020a, Wang et al., 2020]. However in their original paper, a brute-force union bound over all possible value functions are required to obtain similar effect, due to the limitation of model-free algorithm [Zhang et al., 2020a] (see Appendix F

for more details) or linear function approximation [Wang et al., 2020]. This will cause a loose, $\propto \widetilde{\mathcal{O}}(S^2)$ complexity in the obtained bound. Different from their approach, we carefully manipulate a lower order term to avoid union bounding all value functions during the second argument. As a consequence we obtain a near-tight $\propto \widetilde{\mathcal{O}}(\min\{d, S\} \cdot S)$ dependence in the final bound. We believe this observation has broader application in the analysis of similar RL problems.

Proof Sketch of Theorem 4. We next introduce the idea of constructing the hard instance that witnesses the lower bound in Theorem 4. A basic ingredient is the hard instance given by Jin et al. [2020]. However, this hard instance is invented for RFE, where the corresponding lower bound is $K \geq \Omega\big(H^2 S^2 A / \epsilon^2\big)$. Note this lower bound cannot match the upper bound in Theorem 3 in terms of d and S. In order to develop a dedicated lower bound for PFE, we utilize Johnson–Lindenstrauss Lemma [Johnson and Lindenstrauss, 1984] to refine the hard instance in [Jin et al., 2020], and successfully reduce a factor S to $\min\{d, S\}$ in their lower bound, which gives the result in Theorem 4. We believe that the idea to refine RL hard instance by Johnson–Lindenstrauss Lemma is of broader interests. A rigorous proof is deferred to Appendix D.

Application in Reward-Free Exploration. By setting $d = SA$ and allowing arbitrary reward functions, PFE problems reduce to RFE problems. Therefore as a side product, Theorem 3 implies the following results for RFE problems on stationary or non-stationary MDPs[6]:

Corollary 5. *Suppose $\epsilon > 0$ is sufficiently small. Consider the reward-free exploration problems on a* stationary *MDP. Suppose* `PF-UCB` *(Algorithm 2) is run for*

$$K = \mathcal{O}\big(H^3 S^2 A \iota / \epsilon^2\big), \quad \text{where } \iota := \log(HSA/(\delta\epsilon)),$$

episodes, then `PF-UCB` *(Algorithm 3) is (ϵ, δ)-PAC. Moreover, if the MDP is* non-stationary, *the above bound will be revised to $K = \mathcal{O}\big(H^4 S^2 A \iota / \epsilon^2\big)$.*

Remark 4. When applied to RFE, `PF-UCB` matches the rate shown in [Kaufmann et al., 2020] and improves an H factor compared with [Jin et al., 2020] (for both stationary and non-stationary MDPs), but is an H factor loose compared with the current best rates, [Ménard et al., 2020] (for non-stationary MDPs) and [Zhang et al., 2020b] (for stationary MDPs). However we highlight that our results are superior in the context of PFE, since `PF-UCB` adapts with the structure of the rewards. In specific, if rewards admit a d-dimensional feature for $d < S$, `PF-UCB` only needs $\propto \widetilde{\mathcal{O}}(\min\{d, S\} \cdot S)$ samples, but the above methods must explore the whole environment with a high precision which consumes $\propto \widetilde{\mathcal{O}}(S^2)$ samples.

Application in Task-Agnostic Exploration. PFE is also related to *task-agnostic exploration* (TAE) [Zhang et al., 2020a]: in PFE, the agent needs to plan for an arbitrary reward function from a fixed and bounded d-dimensional space; and in TAE, the agent needs to plan for N fixed reward functions. Due to the nature of the problem setups, PFE algorithms (that do not exploit the given reward basis r during exploration, e.g., ours) and TAE algorithms can be easily applied to solve the other problem through a covering argument and a union bound, and with a modification of $\min\{d, S\} \leftrightarrow \log(N)$ in the obtained trajectory complexity bounds. For TAE on a non-stationary MDP, Zhang et al. [2020a] show an algorithm which takes $\widetilde{\mathcal{O}}\big(\log(N) \cdot H^5 SA / \epsilon^2\big)$ episodes for TAE[7]. In comparison, when applied to TAE on a non-stationary MDP, Theorem 3 implies `PF-UCB` only takes $\widetilde{\mathcal{O}}\big(\log(N) \cdot H^4 SA / \epsilon^2\big)$ episodes[8], which improves [Zhang et al., 2020a].

---

[6]Our considered MDP is *stationary* as the transition probability $\mathbb{P}$ is fixed (across different steps). An MDP is called *non-stationary*, if the transition probability varies at different steps, i.e., replacing $\mathbb{P}$ by $\{\mathbb{P}_h\}_{h=1}^H$.

[7]This bound is copied from [Zhang et al., 2020a], which is erroneously stated due to a technical issue in the proof of Lemma 2 in their original paper. The issue can be fixed by a covering argument and union bound on the value functions, but then the obtained bound should be $\widetilde{\mathcal{O}}\big(\log(N) H^5 S^2 A / \epsilon^2\big)$. See Appendix F for details.

[8]The conversion holds as follows. First set $d = 1$ in our algorithm to yield an algorithm for TAE with a single agnostic task, where we have $\min\{d, S\} = 1$. Then one can extend this algorithm to TAE with N agnostic tasks using a union bound to have the algorithm succeed simultaneously for all N tasks, which adds a $\log N$ multiplicative factor in the sample complexity bound. In this way, we obtain a TAE algorithm with a sample complexity bound in Theorem 3 where $\min\{d, S\}$ is replaced with $\log N$.

5 Related Works

MORL. MORL receives extensive attention from RL applications [Yang et al., 2019, Natarajan and Tadepalli, 2005, Mossalam et al., 2016, Abels et al., 2018, Roijers et al., 2013]. However, little theoretical results are known, especially in terms of the sample efficiency and adversarial preference settings. A large amount of MORL works focus on identifying (a cover for) the policies belonging to the *Pareto front* [Yang et al., 2019, Roijers et al., 2013, Cheung, 2019], but the correspondence between a preference and an optimal policy is ignored. Hence they cannot "accommodate picky customers" as our algorithms do. To our knowledge, this paper initiates the theoretical study of the sample efficiency for MORL in the setting of adversarial preferences.

Adversarial MDP. Similarly to the online MORL problem studied in this paper, the adversarial MDP problem [Neu et al., 2012, Rosenberg and Mansour, 2019, Jin et al., 2019] also allows the reward function to change adversarially over time. However, we study a totally different regret (see (1)). In the adversarial MDP problem, the regret is measured against a *fixed* policy that is best in the hindsight; but in online MORL problem, we study a regret against a sequence of optimal policies with respect to the sequence of incoming preferences, i.e., our benchmark policy may *vary* over time. Therefore our regret bound for online MORL problems is orthogonal to those for adversarial MDP problems [Neu et al., 2012, Rosenberg and Mansour, 2019, Jin et al., 2019].

Constrained MDP. In the problem of the *constrained MDP*, d constraints are enforced to restrict the policy domain, and the agent aims to find a policy that belongs to the domain and maximizes the cumulative rewards [Achiam et al., 2017, Efroni et al., 2020, El Chamie et al., 2018, Fisac et al., 2018, Wachi and Sui, 2020, Garcıa and Fernández, 2015, Brantley et al., 2020]. Constrained MDP is related to MORL as when the constraints are soft, they can be formulated as "objectives" with negative weights. However, there is a fundamental difference: constrained MDP aims to optimize only one (and known) objective, where in MORL studied in this paper, we aim to be able to find near optimal policies for any preference-weighted objective (to accommodate picky customers).

Reward-Free Exploration. The proposed preference-free exploration problem generalizes the problems of reward-free exploration [Jin et al., 2020]. Compared with existing works for RFE [Jin et al., 2020, Kaufmann et al., 2020, Ménard et al., 2020, Zhang et al., 2020b], our method has the advantage of adapting with rewards that admit low-dimensional structure — this partly answers an open question raised by Jin et al. [2020]. We also note that Wang et al. [2020] study RFE with linear function approximation on the value functions; in contrast our setting can be interpreted as RFE with linear function approximation on the reward functions, which is orthogonal to their setting.

6 Conclusion

In this paper we study provably sample-efficient algorithms for multi-objective reinforcement learning in both online and preference-free exploration settings. For both settings, sample-efficient algorithms are proposed and their sample complexity analysis is provided; moreover, two information-theoretic lower bounds are proved to justify the near-tightness of the proposed algorithms, respectively. Our results extend existing theory for single-objective RL and reward-free exploration, and resolve an open question raised by Jin et al. [2020].

Acknowledgement

References

Axel Abels, Diederik M Roijers, Tom Lenaerts, Ann Nowé, and Denis Steckelmacher. Dynamic weights in multi-objective deep reinforcement learning. *arXiv preprint arXiv:1809.07803*, 2018.

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning*, pages 22–31. PMLR, 2017.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 263–272. JMLR. org, 2017.

Kianté Brantley, Miroslav Dudik, Thodoris Lykouris, Sobhan Miryoosefi, Max Simchowitz, Aleksandrs Slivkins, and Wen Sun. Constrained episodic reinforcement learning in concave-convex and knapsack settings. *arXiv preprint arXiv:2006.05051*, 2020.

Wang Chi Cheung. Exploration-exploitation trade-off in reinforcement learning on online markov decision processes with global concave rewards. *arXiv preprint arXiv:1905.06466*, 2019.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *arXiv preprint arXiv:1703.07710*, 2017.

Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

Mahmoud El Chamie, Yue Yu, Behçet Açıkmeşe, and Masahiro Ono. Controlled markov processes with safety state constraints. *IEEE Transactions on Automatic Control*, 64(3):1003–1018, 2018.

Jaime F Fisac, Anayo K Akametalu, Melanie N Zeilinger, Shahab Kaynama, Jeremy Gillula, and Claire J Tomlin. A general safety framework for learning-based control in uncertain robotic systems. *IEEE Transactions on Automatic Control*, 64(7):2737–2752, 2018.

Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial mdps with bandit feedback and unknown transition, 2019.

Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. *arXiv preprint arXiv:2002.02794*, 2020.

William B Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Edouard Leurent, and Michal Valko. Adaptive reward-free exploration. *arXiv preprint arXiv:2006.06294*, 2020.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. *arXiv preprint arXiv:2007.13442*, 2020.

Hossam Mossalam, Yannis M Assael, Diederik M Roijers, and Shimon Whiteson. Multi-objective deep reinforcement learning. *arXiv preprint arXiv:1610.02707*, 2016.

Sriraam Natarajan and Prasad Tadepalli. Dynamic preferences in multi-criteria reinforcement learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 601–608, 2005.

Gergely Neu, Andras Gyorgy, and Csaba Szepesvári. The adversarial stochastic shortest path problem with unknown transition probabilities. In *Artificial Intelligence and Statistics*, pages 805–813, 2012.

Diederik M Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113, 2013.

Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial markov decision processes. *arXiv preprint arXiv:1905.07773*, 2019.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Akifumi Wachi and Yanan Sui. Safe reinforcement learning in constrained markov decision processes. In *International Conference on Machine Learning*, pages 9797–9806. PMLR, 2020.

Ruosong Wang, Simon S Du, Lin F Yang, and Ruslan Salakhutdinov. On reward-free reinforcement learning with linear function approximation. *arXiv preprint arXiv:2006.11274*, 2020.

Runzhe Yang, Xingyuan Sun, and Karthik Narasimhan. A generalized algorithm for multi-objective reinforcement learning and policy adaptation. In *Advances in Neural Information Processing Systems*, pages 14610–14621, 2019.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312. PMLR, 2019.

Xuezhou Zhang, Yuzhe Ma, and Adish Singla. Task-agnostic exploration in reinforcement learning, 2020a.

Zihan Zhang, Simon S Du, and Xiangyang Ji. Nearly minimax optimal reward-free reinforcement learning. *arXiv preprint arXiv:2010.05901*, 2020b.

Checklist

1. For all authors...

    (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]

    (b) Did you describe the limitations of your work? [Yes]

    (c) Did you discuss any potential negative societal impacts of your work? [N/A]

    (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

    (a) Did you state the full set of assumptions of all theoretical results? [Yes]

    (b) Did you include complete proofs of all theoretical results? [Yes]

3. If you ran experiments...

    (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]

    (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]

    (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]

    (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

    (a) If your work uses existing assets, did you cite the creators? [Yes]

    (b) Did you mention the license of the assets? [N/A]

    (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

    (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

    (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

    (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]
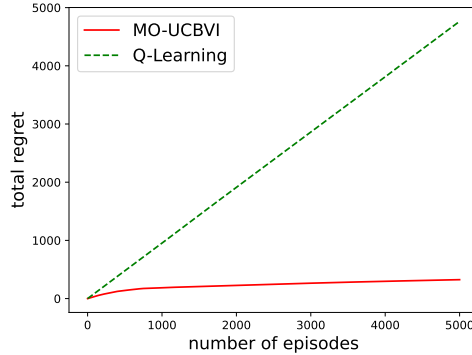
Figure 2: `MO-UCBVI` vs. Q-learning [Jin et al., 2018] in a simulated random multi-objective MDP. The plot suggests that `MO-UCBVI` achieves sublinear regret but Q-learning incurs linear regret.

A Numerical Simulations

Code for simulations is available at `https://github.com/uuujf/MORL`.

Comparison with the Optimal Single-Objective RL Algorithm. Figure 1 shows a regret comparison between the proposed `MO-UCBVI` and the best-in-hindsight policy. In the experiment we simulate a random multi-objective MDP with $S = 20, A = 5, H = 10, d = 15$, and run the algorithms for $K = 5000$ episodes. The best-in-hindsight policy refers to a policy that is fixed across episodes and achieves maximum cumulative rewards in the whole game, i.e., $\arg\max_\pi \sum_{k=1}^K V_1^k(x_1; w^k)$. Note that the best-in-hindsight policy is the optimal policy in the single-objective RL setting; however it could be much worse than a time-varying policy in the multi-objective RL setting. Since the information-theoretically adversarial preferences are computationally infeasible to compute, in Figure 1 we simply use a randomly generated set of preferences, for which the "best-in-hindsight" policy already performs poorly. In sum, the plot shows that our algorithm achieves sublinear regret in online MORL, but the best single-objective RL algorithm will incur linear regret in online MORL.

Comparison with Q-Learning. Since MORL is connected to task-agnostic exploration [Zhang et al., 2020a] in the exploration setting (see discussions in Section 4), it is tempted to think that the Q-learning method studied in Jin et al. [2018], Zhang et al. [2020a] could also work in the setting of online MORL. However this is refuted by Figure 2. We simulate a random multi-objective MDP with $S = 20, A = 5, H = 10, d = 15$, and run the algorithms for $K = 5000$ episodes. The Q-learning algorithm is specified by Algorithm 1 in Jin et al. [2018], excepted that the reward is now a linear scalarization of a preference vector and a multi-objective reward vector. Figure 2 shows that Q-learning cannot achieve a sublinear regret in the setting of online MORL.

The Effect of Number of Objectives. Figure 3 shows the performance of `MO-UCBVI` with different number of objectives. In the experiment we simulate a random multi-objective MDP with $S = 20, A = 5, H = 10$, and number of objectives $d \in \{1, 5, 15, 20, 30\}$, and run `MO-UCBVI` for $K = 5000$ episodes. As shown in Figure 3, the regret will increase as the number of objectives increases; moreover, in all settings `MO-UCBVI` achieves a sublinear regret.

B Proof of Theorem 1

Our proof utilizes techniques from [Zanette and Brunskill, 2019].

Notations. For two functions $f(x) \geq 0$ and $g(x) \geq 0$ defined for $x \in [0, \infty)$, we write $f(x) \lesssim g(x)$ if $f(x) \leq c \cdot g(x)$ for some absolute constant $c > 0$; we write $f(x) \gtrsim g(x)$ if $g(x) \lesssim f(x)$; and we write $f(x) \asymp g(x)$ if $f(x) \lesssim g(x) \lesssim f(x)$. Moreover, we write $f(x) = \mathcal{O}(g(x))$ if $\lim_{x \to \infty} f(x)/g(x) < c$ for some absolute constant $c > 0$; we write $f(x) = \Omega(g(x))$ if $g(x) = \mathcal{O}(f(x))$; and we write $f(x) = \Theta(g(x))$ if $f(x) = \mathcal{O}(g(x))$ and $g(x) = \mathcal{O}(f(x))$. To hide
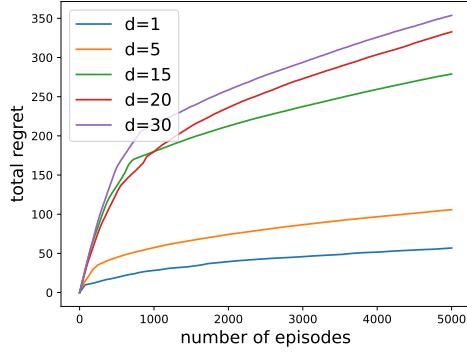
Figure 3: The effect of number of objectives. The plot shows the performance of `MO-UCBVI` in a simulated random multi-objective MDP with different number of objectives. The plot suggests that regret increases as number of objectives increases, but `MO-UCBVI` achieves sublinear regret in all cases, as predicted by Theorem 1.

the logarithmic factors, we write $f(x) = \widetilde{\mathcal{O}}(g(x))$ if $f(x) = \mathcal{O}(g(x) \log^d x)$ for some absolute constant $d > 0$. For $a, b \in \mathbb{R}$, we write $a \wedge b := \min\{a, b\}$ and $a \vee b := \max\{a, b\}$. We will use ι to denote a general logarithmic factor. Let π^k be the planning policy at the k-th episode, i.e., a greedy policy that maximizes $Q_h^k(x, a)$. Let $w_h^k(x, a) := \mathbb{P}\left\{(x_h, a_h) = (x, a) \mid \pi^k, \mathbb{P}\right\}$ and $w^k(x, a) := \sum_h w_h^k(x, a)$.

B.1 Good Events and Useful Lemmas

Fix ϵ to be a constant to be determined later (we will set $\epsilon = 1/K$). Consider the following events:

$$G_H := \left\{ \forall x, a, h, k, w, \ \left|\left(\widehat{\mathbb{P}}^k - \mathbb{P}\right) V_{h+1}^*\Big|_{x,a,w}\right| \leq 2\epsilon + \sqrt{\frac{(d \wedge S) H^2}{2 N^k(x,a)} \log \frac{6 H^2 SAK}{\delta \epsilon}} \right\}, \quad (\text{G}_\text{H})$$

$$G_{\widehat{B}} := \left\{ \forall x, a, h, k, w, \ \left|\left(\widehat{\mathbb{P}}^k - \mathbb{P}\right) V_{h+1}^*\Big|_{x,a,w}\right| \leq \right.$$

$$\left. 2\epsilon + \sqrt{\frac{2(d \wedge S) \left\|V_{h+1}^*\right\|_{\widehat{\mathbb{P}}^k}^2}{N^k(x,a)} \log \frac{6 H^2 SAK}{\delta \epsilon}} + \frac{7(d \wedge S) H}{3 N^k(x,a)} \log \frac{6 H^2 SAK}{\delta \epsilon} \right\}, \quad (\text{G}_{\widehat{B}})$$

$$G_V := \left\{ \forall x, a, h, k, w, \ \left| \|V_h^*\|_{\widehat{\mathbb{P}}^k} - \|V_h^*\|_{\mathbb{P}} \Big|_{x,a,w}\right| \leq 2\epsilon + \sqrt{\frac{4(d \wedge S) H^2}{N^k(x,a)} \log \frac{6 H^2 SAK}{\delta \epsilon}} \right\}, \quad (\text{G}_\text{V})$$

$$G_P := \left\{ \forall x, a, y, k, \ \left|\widehat{\mathbb{P}}^k(y \mid x, a) - \mathbb{P}(y \mid x, a)\right| \leq \right.$$

$$\left. \sqrt{\frac{2 \mathbb{P}(y \mid x, a)}{N^k(x,a)} \log \frac{2 S^2 AK}{\delta}} + \frac{2}{3 N^k(x,a)} \log \frac{2 S^2 AK}{\delta} \right\}, \quad (\text{G}_\text{P})$$

$$G_{\widehat{P}} := \left\{ \forall x, a, y, k, \ \left|\widehat{\mathbb{P}}^k(y \mid x, a) - \mathbb{P}(y \mid x, a)\right| \leq \right.$$

$$\left. \sqrt{\frac{2 \widehat{\mathbb{P}}^k(y \mid x, a)}{N^k(x,a)} \log \frac{2 S^2 AK}{\delta}} + \frac{7}{3 N^k(x,a)} \log \frac{2 S^2 AK}{\delta} \right\}, \quad (\text{G}_{\widehat{P}})$$

$$G_N := \left\{ \forall x, a, k, \ N^k(x,a) \geq \frac{1}{2} \sum_{j<k} w^j(x,a) - H \log \frac{HSA}{\delta} \right\}. \quad (\text{G}_\text{N})$$

15

Lemma 1 (Probability of good events.)**.** *Each of the good event holds with probability at least* $1 - \delta$. *In particular, they hold simultaneously with probability at least*

$$\mathbb{P}\left\{G_H \cap G_{\widehat{B}} \cap G_V \cap G_P \cap G_{\widehat{P}} \cap G_N\right\} \geq 1 - 6\delta.$$

Proof. We only need to prove that each of the good event holds with probability at least $1 - \delta$.

We first show that $\mathbb{P}\left\{G_H\right\} \geq 1 - \delta$ by Hoeffding's inequality and a covering argument. In particular by Hoeffding's inequality we have that for fixed x, a, h, k, w,

$$\left|(\widehat{\mathbb{P}}^k - \mathbb{P})V_{h+1}^*\Big|_{x,a,w}\right| \leq \sqrt{\frac{H^2}{2N^k(x,a)}\log\frac{2}{\delta}}$$

holds with probability at least $1 - \delta$. We first apply a covering argument for the preferences set. We consider an $\frac{\epsilon}{H}$-covering $\mathcal{C}$ for the unit ball $\left\{w \in \mathbb{R}^d : \|w\|_1 \leq 1\right\}$, then $|\mathcal{C}| \leq \left(\frac{3H}{\epsilon}\right)^d$, and for any w in the ball, there exists $w' \in \mathcal{C}$ such that $\|w - w'\|_1 \leq \frac{\epsilon}{H}$. Then a union bound on $w \in \mathcal{C}$ and $(x, a, k, h) \in \mathcal{S} \times \mathcal{A} \times [K] \times [H]$ yields that with probability at least $1 - \delta$, the following holds for every $(x, a, k, h, w) \in \mathcal{S} \times \mathcal{A} \times [K] \times [H] \times \mathcal{C}$:

$$\left|(\widehat{\mathbb{P}}^k - \mathbb{P})V_{h+1}^*\Big|_{x,a,w}\right| \leq \sqrt{\frac{dH^2}{2N^k(x,a)}\log\frac{6H^2SAK}{\delta\epsilon}}$$

Now consider an arbitrary $w \in \mathcal{W}$ (hence in the unit ball), let $w' \in \mathcal{C}$ be such that $\|w - w'\|_1 \leq \frac{\epsilon}{H}$, then by Lemma 2 we have that with probability at least $1 - \delta$,

$$\left|(\widehat{\mathbb{P}}^k - \mathbb{P})V_{h+1}^*\Big|_{x,a,w}\right| \leq \left|\widehat{\mathbb{P}}^k V_{h+1}^*\Big|_{x,a,w} - \widehat{\mathbb{P}}^k V_{h+1}^*\Big|_{x,a,w'}\right|$$

$$+ \left|(\widehat{\mathbb{P}}^k - \mathbb{P})V_{h+1}^*\Big|_{x,a,w'}\right| + \left|\mathbb{P}V_{h+1}^*\Big|_{x,a,w'} - \mathbb{P}V_{h+1}^*\Big|_{x,a,w}\right|$$

$$\leq 2\epsilon + \sqrt{\frac{dH^2}{2N^k(x,a)}\log\frac{6H^2SAK}{\delta\epsilon}}$$

holds for every $(x, a, k, h, w) \in \mathcal{S} \times \mathcal{A} \times [K] \times [H] \times \mathcal{W}$. Similarly, we can apply the above covering argument for the value function set by considering an ϵ-covering $\mathcal{C}$ for the H-ball $\left\{v \in \mathbb{R}^S : \|v\|_\infty \leq H\right\}$, then $|\mathcal{C}| \leq \left(\frac{3H}{\epsilon}\right)^S$, then we obtain that with probability at least $1 - \delta$,

$$\left|(\widehat{\mathbb{P}}^k - \mathbb{P})V_{h+1}^*\Big|_{x,a,w}\right| \leq 2\epsilon + \sqrt{\frac{SH^2}{2N^k(x,a)}\log\frac{6H^2SAK}{\delta\epsilon}}$$

holds for every x, a, k, h and every value functions, hence for every preference-induced value functions. The two inequalities together show that $\mathbb{P}\left\{G_H\right\} \geq 1 - \delta$.

Similarly, we can apply the covering arguments with the empirical Bernstein's inequalities [Maurer and Pontil, 2009] to show that $\mathbb{P}\left\{G_{\widehat{B}}\right\} \geq 1 - \delta$.

$\mathbb{P}\left\{G_V\right\} \geq 1 - \delta$ is proved by the covering arguments with Theorem 10 from [Maurer and Pontil, 2009].

$\mathbb{P}\left\{G_P\right\} \geq 1 - \delta$ and $\mathbb{P}\left\{G_{\widehat{P}}\right\} \geq 1 - \delta$ are proved by Bernstein and empirical Bernstein's inequalities [Maurer and Pontil, 2009], respectively.

Finally, $\mathbb{P}\left\{G_N\right\} \geq 1 - \delta$ is due to Lemma F.4 from [Dann et al., 2017].

$\square$

Lemma 2 (Continuity)**.** *For every h, we have $|V_h^*(x; w) - V_h^*(x; w')| \leq (H - h + 1)\|w - w'\|_1$.*

Proof. We prove it by induction. For $H + 1$, we have $V_{H+1}^*(x; w) = V_{H+1}^*(x; w') = 0$. Now suppose

$$\left|V_{h+1}^*(x; w) - V_{h+1}^*(x; w')\right| \leq (H - h)\|w - w'\|_1,$$

16

and consider h. Without loss of generality, suppose $V_h^*(x; w) \geq V_h^*(x; w')$. Then

$$
\begin{aligned}
|V_h^*(x; w) - V_h^*(x; w')| &= V_h^*(x; w) - V_h^*(x; w') \\
&= \max_a Q_h^*(x, a; w) - \max_{a'} Q_h^*(x, a; w') \\
&\leq Q_h^*(x, a; w) - Q_h^*(x, a; w') \qquad (\text{let } a = \arg\max_a Q_h^*(x, a; w)) \\
&\leq |\langle w - w', \boldsymbol{r}_h(x, a) \rangle| + \sum_{y \in \mathcal{S}} \mathbb{P}(y \mid x, a) \left(V_{h+1}^*(y; w) - V_{h+1}^*(y; w') \right) \\
&\leq \|w - w'\|_1 + (H - h) \|w - w'\|_1 \\
&= (H - h + 1) \|w - w'\|_1 .
\end{aligned}
$$

Hence the conclusion holds. $\qquad\qquad\square$

Lemma 3 (Bounds for lower order terms). *If events* $(\mathrm{G_P})$, $(\mathrm{G_{\widehat{P}}})$ *hold, then for every* V_1, V_2 *such that* $0 \leq V_1(x; w) \leq V_2(x; w) \leq H$, *and for every* x, a, h, k, w, *the following inequalities hold:*

1. $\left| (\widehat{\mathbb{P}}^k - \mathbb{P})(V_2 - V_1) \right|_{x,a,w} \leq \frac{1}{H} \left. \mathbb{P}(V_2 - V_1) \right|_{x,a,w} + \frac{2H^2 S\iota}{N^k(x,a)}$;

2. $\left| (\widehat{\mathbb{P}}^k - \mathbb{P})(V_2 - V_1) \right|_{x,a,w} \leq \frac{1}{H} \left. \widehat{\mathbb{P}}(V_2 - V_1) \right|_{x,a,w} + \frac{3H^2 S\iota}{N^k(x,a)}$;

3. $\left| (\widehat{\mathbb{P}}^k - \mathbb{P})(V_2 - V_1)^2 \right|_{x,a,w} \leq \left. \mathbb{P}(V_2 - V_1)^2 \right|_{x,a,w} + \frac{2H^2 S\iota}{N^k(x,a)}$;

4. $\left. \|V_2 - V_1\|_{\widehat{\mathbb{P}}^k}^2 \right|_{x,a,w} \leq 2 \left. \mathbb{P}(V_2 - V_1)^2 \right|_{x,a,w} + \frac{2H^2 S\iota}{N^k(x,a)}$;

Proof. For simplicity in this proof we denote $p(y) := \mathbb{P}(y \mid x, a)$ and $\hat{p}(y) := \widehat{\mathbb{P}}^k(y \mid x, a)$.

For the first inequality,

$$
\begin{aligned}
\left| (\widehat{\mathbb{P}}^k - \mathbb{P})(V_2 - V_1) \right|_{x,a,w} &\leq \sum_{y \in \mathcal{S}} |\hat{p}^k(y) - p(y)| \, (V_2(y; w) - V_1(y; w)) \\
&\leq \sum_{y \in \mathcal{S}} \left(\sqrt{\frac{2p(y)\iota}{N^k(x,a)}} + \frac{2\iota}{3N^k(x,a)} \right) (V_2(y; w) - V_1(y; w)) \qquad (\text{since } (\mathrm{G_P}) \text{ holds}) \\
&\leq \sum_{y \in \mathcal{S}} \left(\frac{p(y)}{H} + \frac{H\iota}{2N^k(x,a)} + \frac{2\iota}{3N^k(x,a)} \right) (V_2(y; w) - V_1(y; w)) \qquad (\text{use } \sqrt{ab} \leq \tfrac{1}{2}(a + b)) \\
&\leq \sum_{y \in \mathcal{S}} \frac{p(y)}{H} (V_2(y; w) - V_1(y; w)) + \frac{2H^2 S\iota}{N^k(x,a)} \\
&= \frac{1}{H} \left. \mathbb{P}(V_2 - V_1) \right|_{x,a,w} + \frac{2H^2 S\iota}{N^k(x,a)}.
\end{aligned}
$$

The second inequality is proved in a same way as the first one, except that in the second step we use event $(\mathrm{G_{\widehat{P}}})$ instead of $(\mathrm{G_P})$.

17

For the third inequality,

$$
\left| (\widehat{\mathbb{P}}^k - \mathbb{P})(V_2 - V_1)^2 \Big|_{x,a,w} \right| \le \sum_y |\hat{p}(y) - p(y)| \left(V_2(y;w) - V_1(y;w) \right)^2
$$

$$
\le \sum_y \left(\sqrt{\frac{2p(y)\iota}{N^k(x,a)}} + \frac{2\iota}{3N^k(x,a)} \right) (V_2(y;w) - V_1(y;w))^2 \qquad \text{(by (}G_P\text{))}
$$

$$
\le \sum_y \left(p(y) + \frac{\iota}{2N^k(x,a)} + \frac{2\iota}{3N^k(x,a)} \right) (V_2(y;w) - V_1(y;w))^2 \qquad \text{(use } \sqrt{ab} \le a/2 + b/2 \text{)}
$$

$$
\le \mathbb{P}(V_2 - V_1)^2 \Big|_{x,a,w} + \frac{2H^2 S\iota}{N^k(x,a)}.
$$

For the fourth inequality, notice the following

$$
\|V_2 - V_1\|_{\widehat{\mathbb{P}}^k}^2 \le \widehat{\mathbb{P}}^k (V_2 - V_1)^2 \le \mathbb{P}(V_2 - V_1)^2 + (\widehat{\mathbb{P}}^k - \mathbb{P})(V_2 - V_1)^2,
$$

and use the third inequality.

$\square$

B.2 Proof of the Hoeffding Variant.

In this part, we follow notations in Algorithm 1 and prove the first claim in Theorem 1.

Bonus. We set the bonus function in Algorithm 1 to be

$$
b^k(x,a) := 2\epsilon + \sqrt{\frac{(d \wedge S)H^2\iota}{2N^k(x,a)}}, \qquad \iota := \log \frac{6H^2 SAK}{\delta\epsilon} \tag{2}
$$

where ϵ will be set as $\epsilon = 1/K$.

Lemma 4 (Optimistic value function estimation). *If event* (G$_H$) *holds, then* $V_h^*(x;w) \le V_h^k(x;w)$.

Proof. We prove the lemma by induction over $h \in [H]$. For $H + 1$, $V_{H+1}^k(x;w) = V_{H+1}^*(x;w) = 0$. Now suppose $V_{h+1}^k(x;w) \ge V_{h+1}^*(x;w)$ holds for all $(x,k,w) \in \mathcal{S} \times [K] \times \mathcal{W}$, and consider h. Let $a = \arg\max_a Q_h^*(x,a;w)$, then by definition we have

$$
V_h^*(x;w) = Q_h^*(x,a;w) = \langle w, \boldsymbol{r}_h(x,a) \rangle + \mathbb{P}V_{h+1}^* \big|_{x,a,w},
$$

$$
V_h^k(x;w) \ge Q_h^k(x,a;w) = H \wedge \left(b^k(x,a) + \langle w, \boldsymbol{r}_h(x,a) \rangle + \widehat{\mathbb{P}}^k V_{h+1}^k \big|_{x,a,w} \right).
$$

If $Q_h^k(x,a;w) = H$, then $V_h^k(x;w) \ge H \ge V_h^*(x;w)$; otherwise we have

$$
V_h^k(x;w) - V_h^*(x;w) \ge b^k(x,a) + \widehat{\mathbb{P}}^k V_{h+1}^k \big|_{x,a,w} - \mathbb{P}V_{h+1}^* \big|_{x,a;w}
$$

$$
\ge b^k(x,a) + \widehat{\mathbb{P}}^k V_{h+1}^* \big|_{x,a,w} - \mathbb{P}V_{h+1}^* \big|_{x,a;w} \qquad \text{(by induction hypothesis)}
$$

$$
\ge 0, \qquad \text{(by (}G_H\text{))}
$$

which completes the induction and finishes the proof. $\square$

Lemma 5 (Per-episode regret decomposition). *If events* (G$_H$), (G$_P$) *hold, then for every k, it holds:*

$$
V_1^*(x_1;w^k) - V_1^{\pi^k}(x_1;w^k) \lesssim \mathbb{E}_{\mathbb{P},\pi^k} \sum_{h=1}^H H \wedge \left(\epsilon + \sqrt{\frac{(d \wedge S)H^2\iota}{N^k(x,a)}} + \frac{H^2 S\iota}{N^k(x_h,a_h)} \right).
$$

Proof. We first note the following recursion,

$$
V_h^k(x;w) - V_h^{\pi^k}(x;w) = Q_h^k(x,a;w) - Q_h^{\pi^k}(x,a;w) \qquad \text{(set } a = \pi_h^k(x) \text{)}
$$

$$\leq b^k + \widehat{\mathbb{P}}^k V_{h+1}^k - \mathbb{P}V_{h+1}^{\pi^k}\Big|_{x,a,w}$$

$$= b^k + (\widehat{\mathbb{P}}^k - \mathbb{P})V_{h+1}^* + (\widehat{\mathbb{P}}^k - \mathbb{P})(V_{h+1}^k - V_{h+1}^*) + \mathbb{P}(V_{h+1}^k - V_{h+1}^{\pi^k})\Big|_{x,a,w}$$

$$\leq 2b^k + (\widehat{\mathbb{P}}^k - \mathbb{P})(V_{h+1}^k - V_{h+1}^*) + \mathbb{P}(V_{h+1}^k - V_{h+1}^{\pi^k})\Big|_{x,a,w} \qquad \text{(by (G}_\mathrm{H}\text{))}$$

$$\leq 2b^k + \frac{2H^2 S\iota}{N^k} + \left(1 + \frac{1}{H}\right)\mathbb{P}(V_{h+1}^k - V_{h+1}^{\pi^k})\Big|_{x,a,w} \qquad \text{(by Lemma 3),}$$

and by solving the recursion and noting that $(1 + 1/H)^H \leq e$ we have

$$V_1^k(x_1; w^k) - V_1^{\pi^k}(x_1; w^k) \leq e \cdot \mathbb{E}_{\mathbb{P},\pi^k} \sum_{h=1}^{H} \left(2b^k(x_h, a_h) + \frac{2H^2 S\iota}{N^k(x_h, a_h)}\right)$$

$$\lesssim \mathbb{E}_{\mathbb{P},\pi^k} \sum_{h=1}^{H} \left(\epsilon + \sqrt{\frac{(d \wedge S)H^2\iota}{N^k(x, a)}} + \frac{H^2 S\iota}{N^k(x_h, a_h)}\right).$$

The claim is proved by using Lemma 4 and noting that the value functions are bounded by H. $\quad\square$

Now we are ready to prove the first part of Theorem 1.

Theorem 6 (Restatement of Theorem 1, Hoeffding part). *Consider Algorithm 1 with bonus function* (2) *and* $\epsilon = 1/K$, *then for any sequence of the incoming preferences* $\{w^1, w^2, \ldots, w^K\}$, *the regret* (1) *of Algorithm 1 is bounded by*

$$\mathtt{regret}(K) \lesssim \sqrt{(d \wedge S) \cdot H^3 SAK \cdot \log(HSAK/\delta)} + H^2 S^2 A \log^2(HSAK/\delta)$$

with probability at least $1 - \delta$.

Proof. First by Lemma 5 we have

$$\mathtt{regret}(K) = \sum_{k=1}^{K} V_1^*(x_1; w^k) - V_1^{\pi^k}(x_1; w^k)$$

$$\lesssim \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{x,a} w_h^k(x, a) \cdot H \wedge \left(\epsilon + \frac{H^2 S\iota}{N^k(x, a)} + \sqrt{\frac{(d \wedge S)H^2\iota}{N^k(x, a)}}\right)$$

$$\lesssim \underbrace{\sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{x,a \notin L^k} w_h^k(x, a) \cdot H}_{(i)} + \underbrace{\sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{x,a \in L^k} w_h^k(x, a) \cdot \epsilon}_{(ii)}$$

$$+ \underbrace{\sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{x,a \in L^k} w_h^k(x, a) \cdot \frac{HS\iota}{N^k(x, a)}}_{(iii)} + \underbrace{\sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{x,a \in L^k} w_h^k(x, a) \cdot \sqrt{\frac{(d \wedge S)H^2\iota}{N^k(x, a)}}}_{(iv)},$$

where we set

$$L^k := \Big\{(x, a) : \sum_{j<k} w^j(x, a) \geq 2H\iota\Big\}. \qquad (3)$$

We next bound each terms separately.

Term (i). By (3) we have that (i) $= \sum_{k=1}^{K} \sum_{x,a \notin L^k} w^k(x, a) \cdot H \lesssim H^2 SA\iota$.

Term (ii). By definition we have that (ii) $= \sum_{k=1}^{K} \sum_{x,a \in L^k} w^k(x, a) \cdot \epsilon \lesssim HSAK\epsilon$.

Term (iii). The third term is bounded as follows:

$$(iii) = \sum_{k=1}^{K} \sum_{x,a} w^k(x, a) \cdot \frac{H^2 S\iota}{N^k(x, a)} \cdot \mathbb{1}\left[\sum_{j<k} w^j(x, a) \geq 2H\iota\right] \qquad \text{(by (3))}$$

19

$$\lesssim \sum_{x,a} \sum_{k=1}^{K} w^k(x,a) \cdot \frac{H^2 S \iota}{\sum_{j<k} w^j(x,a) - H\iota} \cdot \mathbb{1}\left[\sum_{j<k} w^j(x,a) \geq 2H\iota\right] \qquad \text{(by (G}_\text{N}\text{))}$$

$$\lesssim H^2 S\iota \cdot SA \cdot \iota = H^2 S^2 A \iota^2. \qquad \text{(integration trick)}$$

Term (iv). The fourth term is bounded as follows:

$$(\text{iv}) = \sum_{k=1}^{K} \sum_{x,a} w^k(x,a) \cdot \sqrt{\frac{(d \wedge S)H^2\iota}{N^k(x,a)}} \cdot \mathbb{1}\left[\sum_{j<k} w^j(x,a) \geq 2H\iota\right] \qquad \text{(by (3))}$$

$$\lesssim \sum_{x,a} \sum_{k=1}^{K} w^k(x,a) \cdot \sqrt{\frac{(d \wedge S)H^2\iota}{\sum_{j<k} w^j(x,a) - H\iota}} \cdot \mathbb{1}\left[\sum_{j<k} w^j(x,a) \geq 2H\iota\right] \qquad \text{(by (G}_\text{N}\text{))}$$

$$\lesssim \sqrt{(d \wedge S)H^2\iota} \cdot SA \cdot \sqrt{HK} = \sqrt{(d \wedge S)H^3 SAK\iota}. \qquad \text{(integration trick)}$$

Summing up these terms, setting $\epsilon = 1/K$ and applying Lemma 1, we complete the proof.

$\square$

Algorithm 4 `MO-UCBVI` (Bernstein Variant)

1: initialize history $\mathcal{H}^0 = \emptyset$, $\iota = \log(HSAK/\delta)$
2: **for** episode $k = 1, 2, \ldots, K$ **do**
3:     $N^k(x,a)$, $\widehat{\mathbb{P}}^k(y \mid x,a) \leftarrow$ `Empi-Prob`$(\mathcal{H}^{k-1})$
4:     receive a preference w^k
5:     set $V_{H+1}^k(x; w^k) = \underline{V}_{H+1}^k(x; w^k) = 0$
6:     **for** step $h = H, H-1, \ldots, 1$ **do**
7:         **for** $(x,a) \in \mathcal{S} \times \mathcal{A}$ **do**
8:             compute bonus $b_h^k(x,a) :\eqsim \sqrt{\frac{(d \wedge S)\iota}{N^k(x,a)}} \cdot \left(\left\|V_{h+1}^k\right\|_{\widehat{\mathbb{P}}^k} + \left\|V_{h+1}^k - \underline{V}_{h+1}^k\right\|_{\widehat{\mathbb{P}}^k}\right) + \frac{(d \wedge S)H\iota}{N^k(x,a)}$
9:             compute bonus $a_h^k(x,a) :\eqsim \sqrt{\frac{(d \wedge S)\iota}{N^k(x,a)}} \cdot \left(\left\|\underline{V}_{h+1}^k\right\|_{\widehat{\mathbb{P}}^k} + \left\|V_{h+1}^k - \underline{V}_{h+1}^k\right\|_{\widehat{\mathbb{P}}^k}\right) + \frac{(d \wedge S)H\iota}{N^k(x,a)}$
10:             $Q_h^k(x,a; w^k) = \min\left\{H, \langle w^k, \boldsymbol{r}_h(x,a)\rangle + b^k(x,a) + \widehat{\mathbb{P}}_h^k V_{h+1}^k\big|_{x,a,w^k}\right\}$
11:             $V_h^k(x; w^k) = \max_{a \in \mathcal{A}} Q_h^k(x,a; w^k)$
12:             $\pi_h^k(x) = \arg\max_a Q_h^k(x,a; w^k)$
13:             $\underline{Q}_h^k(x,a; w^k) = \max\left\{0, \langle w^k, \boldsymbol{r}_h(x,a)\rangle - a^k(x,a) + \widehat{\mathbb{P}}_h^k \underline{V}_{h+1}^k\big|_{x,a,w^k}\right\}$
14:             $\underline{V}_h^k(x; w^k) = \underline{Q}_h^k(x, \pi_h^k(x); w^k)$
15:         **end for**
16:     **end for**
17:     receive initial state $x_1^k = x_1$
18:     **for** step $h = 1, 2, \ldots, H$ **do**
19:         take action $a_h^k = \pi_h^k(x_h^k)$, and obtain a new state x_{h+1}^k
20:     **end for**
21:     update history $\mathcal{H}^k = \mathcal{H}^{k-1} \cup \{x_h^k, a_h^k\}_{h=1}^H$
22: **end for**

B.3 Proof of Bernstein Variant

In this part, we follow notations in Algorithm 4 and prove the second claim in Theorem 1.

For a (estimated) transition kernel $\mathbb{P}$ and a (estimated) value function $V_h(y; w)$, define its one-step variance as

$$\|V_h\|_{\mathbb{P}}^2\Big|_{x,a,w} := \sum_{y \in \mathcal{S}} \mathbb{P}(y \mid x,a)\left(V_h(y; w) - \mathbb{P}V_h\big|_{x,a,w}\right)^2.$$

Moreover, we often omit $\big|_{x,a,w}$ and simply write $\|V_h\|_{\mathbb{P}}^2$ when x, a, w are clear from the context.

Bonus. Recall that

$$a_h^k(x,a,w) := 2\epsilon + \sqrt{\frac{2(d \wedge S)\iota}{N^k(x,a)}} \cdot \left(\left\| \underline{V}_h^k \right\|_{\widehat{\mathbb{P}}^k} + \left\| V_h^k - \underline{V}_h^k \right\|_{\widehat{\mathbb{P}}^k} \right) + \frac{7(d \wedge S)H\iota}{3N^k(x,a)},$$

$$b_h^k(x,a,w) := 2\epsilon + \sqrt{\frac{2(d \wedge S)\iota}{N^k(x,a)}} \cdot \left(\left\| \overline{V}_h^k \right\|_{\widehat{\mathbb{P}}^k} + \left\| \overline{V}_h^k - \underline{V}_h^k \right\|_{\widehat{\mathbb{P}}^k} \right) + \frac{7(d \wedge S)H\iota}{3N^k(x,a)},$$

$$(4)$$

where ϵ is a parameter to be determined (we will set $\epsilon = 1/K$).

Lemma 6 (Optimistic value estimation). *Under event* $(\mathrm{G}_{\widehat{\mathrm{B}}})$, *we have that for every* x, a, k, h, w,

- $\left| (\widehat{\mathbb{P}}^k - \mathbb{P})V_{h+1}^* \Big|_{x,a,w} \right| \le b_h^k(x,a,w) \wedge a_h^k(x,a,w);$

- $\underline{V}_h^k(x;w) \le V_h^*(x;w) \le \overline{V}_h^k(x;w)$

Proof. We prove the claims by induction. For $H+1$, $V_{H+1}^*(x,a;w) = \overline{V}_{H+1}^k(x,a;w) = \underline{V}_{H+1}^k(x,a;w) = 0$, hence the hypotheses hold. Now suppose the hypotheses hold for $h+1$, i.e.,

$$\left| (\widehat{\mathbb{P}}^k - \mathbb{P})V_{h+1}^* \Big|_{x,a,w} \right| \le b_h^k(x,a,w) \wedge a_h^k(x,a,w), \tag{5}$$

$$\underline{V}_{h+1}^k(x;w) \le V_{h+1}^*(x;w) \le \overline{V}_{h+1}^k(x;w) \tag{6}$$

and consider h. First we have the following two sets of inequalities:

$$\begin{aligned}
\overline{V}_h^k(x;w) - V_h^*(x;w) &\ge \overline{Q}_h^k(x,a;w) - Q_h^*(x,a;w) && (\text{set } a = \pi_h^*(x)) \\
&\ge b_h^k(x,a,w) + \widehat{\mathbb{P}}^k \overline{V}_{h+1}^k \Big|_{x,a,w} - \mathbb{P}V_{h+1}^* \Big|_{x,a,w} \\
&\ge b_h^k(x,a,w) + (\widehat{\mathbb{P}}^k - \mathbb{P})V_{h+1}^* \Big|_{x,a,w} && (\text{by (6)}) \\
&\ge 0, && (\text{by (5)})
\end{aligned}$$

and

$$\begin{aligned}
V_h^*(x;w) - \underline{V}_h^k(x;w) &\ge Q_h^*(x,a;w) - \underline{Q}_h^k(x,a;w) && (\text{set } a = \underline{\pi}_h^*(x)) \\
&\ge a_h^k(x,a,w) + \mathbb{P}V_{h+1}^* \Big|_{x,a,w} - \widehat{\mathbb{P}}^k \underline{V}_{h+1}^k \Big|_{x,a,w} \\
&\ge a_h^k(x,a,w) + (\mathbb{P} - \widehat{\mathbb{P}}^k)V_{h+1}^* \Big|_{x,a,w} && (\text{by (6)}) \\
&\ge 0. && (\text{by (5)})
\end{aligned}$$

These justify that

$$\underline{V}_h^k(x;w) \le V_h^*(x;w) \le \overline{V}_h^k(x;w). \tag{7}$$

Second, the above inequalities imply that

$$\|V_h^*\|_{\widehat{\mathbb{P}}^k} \le \left\| \overline{V}_h^k \right\|_{\widehat{\mathbb{P}}^k} + \left\| \overline{V}_h^k - V_h^* \right\|_{\widehat{\mathbb{P}}^k} \le \left\| \overline{V}_h^k \right\|_{\widehat{\mathbb{P}}^k} + \left\| \overline{V}_h^k - \underline{V}_h^k \right\|_{\widehat{\mathbb{P}}^k}, \tag{8}$$

$$\|V_h^*\|_{\widehat{\mathbb{P}}^k} \le \left\| \underline{V}_h^k \right\|_{\widehat{\mathbb{P}}^k} + \left\| V_h^* - \underline{V}_h^k \right\|_{\widehat{\mathbb{P}}^k} \le \left\| \underline{V}_h^k \right\|_{\widehat{\mathbb{P}}^k} + \left\| \overline{V}_h^k - \underline{V}_h^k \right\|_{\widehat{\mathbb{P}}^k}, \tag{9}$$

therefore,

$$\begin{aligned}
\left| (\widehat{\mathbb{P}}^k - \mathbb{P})V_h^* \Big|_{x,a,w} \right| &\le 2\epsilon + \sqrt{\frac{2(d \wedge S)\iota}{N^k(x,a)}} \cdot \|V_h^*\|_{\widehat{\mathbb{P}}^k} + \frac{7(d \wedge S)H\iota}{3N^k(x,a)} && (\text{by } (\mathrm{G}_{\widehat{\mathrm{B}}})) \\
&\le a_{h-1}^k(x,a;w) \wedge b_{h-1}^k(x,a;w). && (\text{by (8), (9) and (4)})
\end{aligned}$$

This completes our induction. $\qquad\square$

Lemma 7 (Bonus upper bound)*. If events* ($G_{\widehat{B}}$)*,* (G_V) *hold, then we have that*

$$b_h^k \vee a_h^k\big|_{x,a,w} \le 3\epsilon + \sqrt{\frac{2(d \wedge S)\iota}{N^k}}\left(\left\|V_{h+1}^*\right\|_{\mathbb{P}} + 2\sqrt{\mathbb{P}(V_{h+1}^k - \underline{V}_{h+1}^k)^2}\right) + \frac{15HS\iota}{N^k}\bigg|_{x,a,w}.$$

Proof. We first prove the bound for $b_h^k(x, a, w)$. Note that

$$\left\|V_{h+1}^k\right\|_{\widehat{\mathbb{P}}^k} + \left\|V_{h+1}^k - \underline{V}_{h+1}^k\right\|_{\widehat{\mathbb{P}}^k} \le \left\|V_{h+1}^*\right\|_{\widehat{\mathbb{P}}^k} + \left\|V_{h+1}^k - V_{h+1}^*\right\|_{\widehat{\mathbb{P}}^k} + \left\|V_{h+1}^k - \underline{V}_{h+1}^k\right\|_{\widehat{\mathbb{P}}^k}$$

$$\le \left\|V_{h+1}^*\right\|_{\widehat{\mathbb{P}}^k} + 2\left\|V_{h+1}^k - \underline{V}_{h+1}^k\right\|_{\widehat{\mathbb{P}}^k} \qquad \text{(by Lemma 6)}$$

$$\le 2\epsilon + \left\|V_{h+1}^*\right\|_{\mathbb{P}} + \sqrt{\frac{4(d \wedge S)H^2\iota}{N^k(x,a)}} + 2\left\|V_{h+1}^k - \underline{V}_{h+1}^k\right\|_{\widehat{\mathbb{P}}^k} \qquad \text{(by (G_V))}$$

$$\le 2\epsilon + \left\|V_{h+1}^*\right\|_{\mathbb{P}} + \sqrt{\frac{4(d \wedge S)H^2\iota}{N^k}} + 2\sqrt{2\mathbb{P}(V_{h+1}^k - \underline{V}_{h+1}^k)^2 + \frac{2H^2S\iota}{N^k}}\bigg|_{x,a,w} \qquad \text{(by Lemma 3)}$$

$$\le 2\epsilon + \left\|V_{h+1}^*\right\|_{\mathbb{P}} + 2\sqrt{\mathbb{P}\left(\overline{V}_{h+1}^k - \underline{V}_{h+1}^k\right)^2} + \sqrt{\frac{50H^2S\iota}{N^k}}\bigg|_{x,a,w}.$$

Then the bound for $b_h^k(x, a, w)$ is obtained by substituting the above into (4) and use

$$\sqrt{\frac{2(d \wedge S)\iota}{N^k(x,a)}} \cdot 2\epsilon \le \epsilon^2 + \frac{2(d \wedge S)\iota}{N^k(x,a)} \le \epsilon + \frac{2HS\iota}{N^k(x,a)}.$$

Similarly we can obtain the bound for $a_h^k(x, a, w)$. $\qquad \square$

Lemma 8 (Per-episode regret decomposition)*. If events* ($G_{\widehat{B}}$)*,* (G_V)*,* (G_P) *hold, then*

$$V_1^*(x_1; w^k) - V_1^{\pi^k}(x_1; w^k) \lesssim$$

$$\mathbb{E}_{\mathbb{P},\pi^k}\sum_{h=1}^{H} H \wedge \left(\epsilon + \sqrt{\frac{(d \wedge S)\iota}{N^k}}\left(\left\|V_{h+1}^*\right\|_{\mathbb{P}} + \sqrt{\mathbb{P}(V_{h+1}^k - \underline{V}_{h+1}^k)^2}\right) + \frac{H^2S\iota}{N^k}\bigg|_{x_h,a_h,w^k}\right).$$

Proof. We first note the following recursion,

$$V_h^k(x; w) - V_h^{\pi^k}(x; w) = Q_h^k(x, a; w) - Q_h^{\pi^k}(x, a; w) \qquad \text{(set } a = \pi_h^k(x)\text{)}$$

$$\le b_h^k + \widehat{\mathbb{P}}^k V_{h+1}^k - \mathbb{P}V_{h+1}^{\pi^k}\bigg|_{x,a,w}$$

$$= b_h^k + (\widehat{\mathbb{P}}^k - \mathbb{P})V_{h+1}^* + (\widehat{\mathbb{P}}^k - \mathbb{P})(V_{h+1}^k - V_{h+1}^*) + \mathbb{P}(V_{h+1}^k - V_{h+1}^{\pi^k})\bigg|_{x,a,w}$$

$$\le 2b_h^k + (\widehat{\mathbb{P}}^k - \mathbb{P})(V_{h+1}^k - V_{h+1}^*) + \mathbb{P}(V_{h+1}^k - V_{h+1}^{\pi^k})\bigg|_{x,a,w} \qquad \text{(by Lemma 6)}$$

$$\le 2b_h^k + \frac{2H^2S\iota}{N^k} + \left(1 + \frac{1}{H}\right)\mathbb{P}(V_{h+1}^k - V_{h+1}^{\pi^k})\bigg|_{x,a,w} \qquad \text{(by Lemma 3)},$$

and by solving the recursion and noting that $(1 + 1/H)^H \le e$ we have

$$V_1^k(x_1; w^k) - V_1^{\pi^k}(x_1; w^k) \le e \cdot \mathbb{E}_{\mathbb{P},\pi^k}\sum_{h=1}^{H}\left(2b_h^k(x_h, a_h) + \frac{2H^2S\iota}{N^k(x_h, a_h)}\right)$$

$$\lesssim \mathbb{E}_{\mathbb{P},\pi^k}\sum_{h=1}^{H}\left(\epsilon + \sqrt{\frac{(d \wedge S)\iota}{N^k}} \cdot \left(\left\|V_{h+1}^*\right\|_{\mathbb{P}} + \sqrt{\mathbb{P}(V_{h+1}^k - \underline{V}_{h+1}^k)^2}\right) + \frac{H^2S\iota}{N^k}\bigg|_{x_h,a_h,w^k}\right),$$

where the last inequality is by Lemma 7. The claim is proved by using Lemma 6 and noting that the value functions are bounded by H.

$\qquad \square$

Lemma 9 (Lower order regrets). *By setting $\epsilon = 1/K$, we have that*

$$\sum_{k=1}^{K} \mathbb{E}_{\mathbb{P},\pi^k}\left[\sum_{h=1}^{H} \mathbb{P}(V_{h+1}^k - \underline{V}_{h+1}^k)^2\Big|_{x_h,a_h,w^k}\right] \lesssim H^5 S^2 A \iota^2.$$

Proof. We first note the following recursion,

$$V_h^k(x;w) - \underline{V}_h^k(x;w) = Q_h^k(x,a;w) - \underline{Q}_h^k(x,a;w) \quad (\text{set } a = \pi_h^k(x))$$

$$\leq a_h^k + b_h^k + \widehat{\mathbb{P}}^k(V_{h+1}^k - \underline{V}_{h+1}^k)\Big|_{x,a,w}$$

$$= a_h^k + b_h^k + (\widehat{\mathbb{P}}^k - \mathbb{P})(V_{h+1}^k - \underline{V}_{h+1}^k) + \mathbb{P}(V_{h+1}^k - \underline{V}_{h+1}^k)\Big|_{x,a,w}$$

$$\leq a_h^k + b_h^k + \frac{2H^2 S\iota}{N^k} + \left(1 + \frac{1}{H}\right)\mathbb{P}(V_{h+1}^k - \underline{V}_{h+1}^k)\Big|_{x,a,w} \qquad (\text{by Lemma 3})$$

$$\leq 6\epsilon + 6H\sqrt{\frac{2(d \wedge S)\iota}{N^k}} + \frac{30HS\iota}{N^k} + \frac{2H^2 S\iota}{N^k} + \left(1 + \frac{1}{H}\right)\mathbb{P}(V_{h+1}^k - \underline{V}_{h+1}^k)\Big|_{x,a,w}$$

    (by Lemmas 6 and 7)

$$\leq 6\epsilon + \sqrt{\frac{100H^2 S\iota}{N^k}} + \frac{32H^2 S\iota}{N^k} + \left(1 + \frac{1}{H}\right)\mathbb{P}(V_{h+1}^k - \underline{V}_{h+1}^k)\Big|_{x,a,w},$$

by solving which and noting that $(1 + 1/H)^H \leq e$ and that the value functions are bounded by H, we obtain that

$$V_h^k(x;w) - \underline{V}_h^k(x;w) \lesssim \mathbb{E}_{\mathbb{P},\pi^k}\sum_{t \geq h} H \wedge \left(\epsilon + \sqrt{\frac{100H^2 S\iota}{N^k(x_h,a_h)}} + \frac{H^2 S\iota}{N^k(x_h,a_h)}\right),$$

therefore

$$\left(V_h^k(x;w) - \underline{V}_h^k(x;w)\right)^2 \lesssim \left(\mathbb{E}_{\mathbb{P},\pi^k}\sum_{t \geq h} H \wedge \left(\epsilon + \sqrt{\frac{H^2 S\iota}{N^k(x_h,a_h)}} + \frac{H^2 S\iota}{N^k(x_h,a_h)}\right)\right)^2$$

$$\lesssim H \cdot \mathbb{E}_{\mathbb{P},\pi^k}\sum_{t \geq h}\left(H \wedge \left(\epsilon + \sqrt{\frac{H^2 S\iota}{N^k(x_h,a_h)}} + \frac{H^2 S\iota}{N^k(x_h,a_h)}\right)\right)^2$$

$$\lesssim H \cdot \mathbb{E}_{\mathbb{P},\pi^k}\sum_{t \geq h} H^2 \wedge \left(\epsilon^2 + \frac{H^2 S\iota}{N^k(x_h,a_h)} + \frac{H^4 S^2 \iota^2}{(N^k(x_h,a_h))^2}\right).$$

Then

$$\text{LHS} := \sum_{k=1}^{K} \mathbb{E}_{\mathbb{P},\pi^k}\sum_{h=1}^{H} \mathbb{P}(V_{h+1}^k - \underline{V}_{h+1}^k)^2\Big|_{x_h,a_h,w^k}$$

$$= \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{x,a} w_h^k(x,a)\,\mathbb{P}(V_{h+1}^k - \underline{V}_{h+1}^k)^2\Big|_{x,a,w^k}$$

$$\lesssim H \cdot \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{x,a} w_h^k(x,a) \cdot \mathbb{E}_{\mathbb{P},\pi^k}\sum_{t \geq h} H^2 \wedge \left(\epsilon^2 + \frac{H^2 S\iota}{N^k(x_h,a_h)} + \frac{H^4 S^2 \iota^2}{(N^k(x_h,a_h))^2}\right)$$

$$\lesssim H^2 \cdot \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{x,a} w_h^k(x,a) \cdot H^2 \wedge \left(\epsilon^2 + \frac{H^2 S\iota}{N^k(x_h,a_h)} + \frac{H^4 S^2 \iota^2}{(N^k(x_h,a_h))^2}\right)$$

$$\lesssim \sum_{k,h}\sum_{x,a \notin M^k} w_h^k(x,a)H^4 + \sum_{k,h}\sum_{x,a \in M^k} w_h^k(x,a)\left(H^2\epsilon^2 + \frac{H^4 S\iota}{N^k(x,a)} + \frac{H^6 S^2 \iota^2}{(N^k(x,a))^2}\right),$$

where we set
$$M^k := \big\{(x,a) : \sum_{j<k} w^j(x,a) \geq 2HS\iota\big\}.$$

Then by (G_N) and the integration tricks (see the proof of Theorem 7), we obtain

$$\text{LHS} \lesssim H^4 \cdot SA \cdot HS\iota + H^2\epsilon^2 \cdot SA \cdot HK + H^4 S\iota \cdot SA \cdot \iota + H^6 S^2 \iota^2 \cdot SA/(HS\iota)$$
$$\lesssim H^5 S^2 A\iota^2 + H^3 SAK\epsilon^2 \lesssim H^5 S^2 A\iota^2,$$

where the last inequality is because $\epsilon = 1/K$. $\qquad\square$

Now we are ready to prove the second part of Theorem 1.

Theorem 7 (Restatement of Theorem 1, Bernstein part). *Consider Algorithm 4 with bonus function* (4) *and* $\epsilon = 1/K$, *then for any sequence of the incoming preferences* $\{w^1, w^2, \ldots, w^K\}$, *the regret* (1) *of Algorithm 4 is bounded by*

$$\texttt{regret}(K) \lesssim \sqrt{(d \wedge S)H^2 SAK\iota^2} + H^{2.5}S^2 A\iota^2, \quad \iota := \log(HSAK/\delta),$$

with probability at least $1 - \delta$.

Proof. First by Lemma 8 we have

$$\texttt{regret}(K) = \sum_{k=1}^{K} V_1^*(x_1; w^k) - V_1^{\pi^k}(x_1; w^k)$$

$$\lesssim \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{x,a} w_h^k(x,a) H \wedge \Big(\epsilon + \frac{H^2 S\iota}{N^k} + \sqrt{\frac{(d \wedge S)\iota}{N^k}}\Big(\sqrt{\mathbb{P}(V_{h+1}^k - \underline{V}_{h+1}^k)^2} + \big\|V_{h+1}^*\big\|_{\mathbb{P}}\Big)\Big)$$

$$\lesssim \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{x,a\notin L^k} w_h^k(x,a) H}_{(i)} + \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{x,a\in L^k} w_h^k(x,a)\epsilon}_{(ii)} + \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{x,a\in L^k} w_h^k(x,a)\frac{H^2 S\iota}{N^k(x,a)}}_{(iii)}$$

$$+ \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{x,a\in L^k} w_h^k(x,a) \cdot \sqrt{\frac{(d \wedge S)\iota}{N^k(x,a)}} \cdot \sqrt{\mathbb{P}(V_{h+1}^k - \underline{V}_{h+1}^k)^2}\Big|_{x,a,w^k}}_{(iv)}$$

$$+ \underbrace{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{x,a\in L^k} w_h^k(x,a) \cdot \sqrt{\frac{(d \wedge S)\iota}{N^k(x,a)}} \cdot \big\|V_{h+1}^*\big\|_{\mathbb{P}}\Big|_{x,a,w^k}}_{(v)},$$

where we set
$$L^k := \big\{(x,a) : \sum_{j<k} w^j(x,a) \geq 2H\iota\big\}. \tag{10}$$

We next bound each of these terms separately.

Term (i). By (10) we have that (i) $= \sum_{k=1}^{K}\sum_{x,a\notin L^k} w^k(x,a)H \lesssim H^2 SA\iota$.

Term (ii). By definition we have that (ii) $= \sum_{k=1}^{K}\sum_{x,a\in L^k} w^k(x,a)\epsilon \lesssim HSAK\epsilon$.

Term (iii). The third term is bounded as follows:

$$(iii) = \sum_{k=1}^{K}\sum_{x,a} w^k(x,a) \cdot \frac{H^2 S\iota}{N^k(x,a)} \cdot \mathbb{1}\Big[\sum_{j<k} w^j(x,a) \geq 2H\iota\Big] \qquad \text{(by (10))}$$

$$\lesssim \sum_{x,a}\sum_{k=1}^{K} w^k(x,a) \cdot \frac{H^2 S\iota}{\sum_{j<k} w^j(x,a) - H\iota} \cdot \mathbb{1}\Big[\sum_{j<k} w^j(x,a) \geq 2H\iota\Big] \qquad \text{(by (G_N))}$$

$$\lesssim H^2 S\iota \cdot SA \cdot \iota = H^2 S^2 A\iota^2. \qquad \text{(integration trick)}$$

24

Term (iv). By Lemma 9 and the integration tricks, we bound the fourth term as follows:

$$\text{(iv)} \le \sqrt{(d \wedge S)\iota} \cdot \sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{x,a \in L^k} \frac{w_h^k(x,a)}{N^k(x,a)}} \cdot \sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{x,a \in L^k} w_h^k(x,a) \cdot \mathbb{P}(V_{h+1}^k - \underline{V}_{h+1}^k)^2}$$

$$\lesssim \sqrt{S\iota} \cdot \sqrt{SA\iota} \cdot \sqrt{H^5 S^2 A \iota^2} \lesssim H^{2.5} S^2 A \iota^2.$$

Term (v). The fifth term is the leading term. We proceed to bound this term by noting

$$\text{(v)} \le \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{x,a \in L^k} w_h^k(x,a) \cdot \left(\sqrt{\frac{(d \wedge S)\iota}{N^k(x,a)}} \cdot \left(\left\| V_{h+1}^{\pi^k} \right\|_{\mathbb{P}} + \left\| V_{h+1}^* - V_{h+1}^{\pi^k} \right\|_{\mathbb{P}} \right) \right)$$

$$\le \sqrt{(d \wedge S)\iota} \cdot \sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{x,a \in L^k} \frac{w_h^k(x,a)}{N^k(x,a)}}.$$

$$\left(\underbrace{\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{x,a \in L^k} w_h^k(x,a) \cdot \left\| V_{h+1}^{\pi^k} \right\|_{\mathbb{P}}^2}}_{\text{(v1)}} + \underbrace{\sqrt{\sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{x,a \in L^k} w_h^k(x,a) \cdot \left\| V_{h+1}^* - V_{h+1}^{\pi^k} \right\|_{\mathbb{P}}^2}}_{\text{(v2)}} \right),$$

where

$$\text{(v1)} = \sum_{k=1}^{K} \text{Var}_{\pi^k, \mathbb{P}} \left[\sum_{h=1}^{H} \langle w^k, \boldsymbol{r}_h(x_h, a_h) \rangle \right] \le H^2 K$$

by the law of total variance and that the cumulative reward cannot exceed H, and

$$\text{(v2)} \le \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{x,a \in L^k} w_h^k(x,a) \cdot \mathbb{P}(V_{h+1}^* - V_{h+1}^{\pi^k})^2 \Big|_{x,a,w^k}$$

$$\le H \cdot \sum_{k=1}^{K}\sum_{h=1}^{H}\sum_{x,a \in L^k} w_h^k(x,a) \cdot \mathbb{P}(V_{h+1}^* - V_{h+1}^{\pi^k}) \Big|_{x,a,w^k}$$

$$\le H^2 \cdot \sum_{k=1}^{K} \left(V_1^*(x_1; w^k) - V_1^{\pi^k}(x_1; w^k) \right) = H^2 \cdot \texttt{regret}(K).$$

Hence,

$$\text{(v)} \lesssim \sqrt{(d \wedge S)\iota} \cdot \sqrt{SA\iota} \cdot \left(\sqrt{H^2 K} + H \cdot \sqrt{\texttt{regret}(K)} \right)$$

$$\lesssim \sqrt{(d \wedge S) H^2 SAK\iota^2} + \sqrt{(d \wedge S) H^2 SA\iota^2} \cdot \sqrt{\texttt{regret}(K)}.$$

Summing up terms (i) to (v) and choosing $\epsilon = 1/K$, we obtain

$$\texttt{regret}(K) \lesssim H^2 SA\iota + HSAK\epsilon + H^2 S^2 A \iota^2 + H^{2.5} S^2 A \iota^2$$

$$+ \sqrt{(d \wedge S) H^2 SAK\iota^2} + \sqrt{(d \wedge S) H^2 SA\iota^2} \cdot \sqrt{\texttt{regret}(K)}$$

$$\lesssim H^{2.5} S^2 A \iota^2 + \sqrt{(d \wedge S) H^2 SAK\iota^2} + \sqrt{(d \wedge S) H^2 SA\iota^2} \cdot \sqrt{\texttt{regret}(K)},$$

solving which we obtain

$$\texttt{regret}(K) \lesssim \sqrt{(d \wedge S) H^2 SAK\iota^2} + H^{2.5} S^2 A \iota^2.$$

Applying Lemma 1 completes the proof. □

C Proof of Theorem 3 and Corollary 5

In this section, we follow notations in Algorithm 2 and Algorithm 3 and prove Theorem 3. Let $\overline{V}_h^k(x; w) := \max_a \overline{Q}_h^k(x, a; w)$. Let π^k be the planning policy at the k-th episode and $\bar{\pi}^k$ be the exploration policy at the k-th episode.

Bonus. We set the bonus functions in Algorithm 3 and Algorithm 2 to be

$$b^k(x,a) := 2\epsilon + \sqrt{\frac{(d \wedge S)H^2\iota}{2N^k(x,a)}}, \quad c^k(x,a) := \frac{3H^2S\iota}{N^k(x,a)} + 2b^k(x,a), \quad \iota := \log\frac{6H^2SAK}{\delta\epsilon}, \quad (11)$$

where ϵ is a parameter to be decided later (we will set $\epsilon = 1/K$).

We note Lemma 4 applies for the planning phase. The follows lemma relates planning error with exploration regret.

Lemma 10 (Planning error). *If events* (G_H)*,* ($G_{\widehat{P}}$) *hold, then for every* x, a, h, k, w,

$$V_h^k(x,w) - V_h^{\pi^k}(x,w) \leq \left(1 + \frac{1}{H}\right)^{H-h+1} \cdot \overline{V}_h^k(x).$$

In particular $V_1^k(x_1; w) - V_1^{\pi^k}(x_1; w) \leq e \cdot \overline{V}_1^k(x)$.

Proof. We prove the conclusion by induction. For $H + 1$, the conclusion holds since both the left high side and the right hand side are zero. Next we assume the conclusion holds for $h + 1$, i.e.,

$$V_{h+1}^k(x,w) - V_{h+1}^{\pi^k}(x,w) \leq \left(1 + \frac{1}{H}\right)^{H-h} \cdot \overline{V}_{h+1}^k(x), \quad (12)$$

and consider h. Recall the definitions in Algorithm 2 and Algorithm 3. Set $a = \pi_h^k(x)$, then we have

$$V_h^k(x;w) = Q_h^k(x,a;w) = H \wedge \left(\langle w, \boldsymbol{r}_h(x,a)\rangle + \widehat{\mathbb{P}}^k V_{h+1}^k\Big|_{x,a,w} + b^k(x,a)\right), \quad (13)$$

$$V_h^{\pi^k}(x;w) = Q_h^{\pi^k}(x,a;w) = \langle w, \boldsymbol{r}_h(x,a)\rangle + \mathbb{P}V_{h+1}^{\pi^k}\Big|_{x,a,w}, \quad (14)$$

and

$$\overline{V}_h^k(x) = \max_{a'} \overline{Q}_h^k(x,a') \geq \overline{Q}_h^k(x,a) = H \wedge \left(\widehat{\mathbb{P}}^k \overline{V}_{h+1}^k\Big|_{x,a} + c^k(x,a)\right).$$

If $\overline{Q}_h^k(x,a) = H$, clearly we have $V_h^k(x;w) - V_h^{\pi^k}(x;w) \leq H \leq \overline{V}_h^k(x)$, thus the conclusion holds. In the following suppose

$$\overline{V}_h^k(x) = \max_{a'} \overline{Q}_h^k(x,a') \geq \overline{Q}_h^k(x,a) = \widehat{\mathbb{P}}^k \overline{V}_{h+1}^k\Big|_{x,a} + c^k(x,a). \quad (15)$$

Then (13) and (14) yield

$$V_h^k(x,w) - V_h^{\pi^k}(x,w) \leq \widehat{\mathbb{P}}^k V_{h+1}^k\Big|_{x,a,w} - \mathbb{P}V_{h+1}^{\pi^k}\Big|_{x,a,w} + b^k(x,a)$$

$$= \widehat{\mathbb{P}}^k(V_{h+1}^k - V_{h+1}^{\pi^k}) + (\mathbb{P} - \widehat{\mathbb{P}}^k)(V_{h+1}^* - V_{h+1}^{\pi^k}) + (\widehat{\mathbb{P}}^k - \mathbb{P})V_{h+1}^*\Big|_{x,a,w} + b^k(x,w)$$

$$\leq \widehat{\mathbb{P}}^k(V_{h+1}^k - V_{h+1}^{\pi^k}) + (\mathbb{P} - \widehat{\mathbb{P}}^k)(V_{h+1}^* - V_{h+1}^{\pi^k})\Big|_{x,a,w} + 2b^k(x,a). \quad \text{(by Lemma 4)}$$

$$\leq \left(1 + \frac{1}{H}\right)\widehat{\mathbb{P}}^k\left(V_{h+1}^k - V_{h+1}^{\pi^k}\right)\Big|_{x,a,w} + \frac{3H^2S\iota}{N^k(x,a)} + 2b^k(x,a) \quad \text{(by Lemma 3)}$$

$$\leq \left(1 + \frac{1}{H}\right)\widehat{\mathbb{P}}^k\left(\left(1 + \frac{1}{H}\right)^{H-h}\overline{V}_{h+1}^k\right)\Big|_{x,a} + \frac{3H^2S\iota}{N^k(x,a)} + 2b^k(x,a) \quad \text{(by (12))}$$

$$= \left(1 + \frac{1}{H}\right)^{H-h+1}\widehat{\mathbb{P}}^k\overline{V}_{h+1}^k\Big|_{x,a} + c^k(x,a) \quad \text{(by (11))}$$

$$\leq \left(1 + \frac{1}{H}\right)^{H-h+1}\overline{Q}_h^k(x,a) \leq \left(1 + \frac{1}{H}\right)^{H-h+1}\overline{V}_h^k(x) \quad \text{(by (15))}.$$

Thus the conclusion also holds for h. By induction we complete the proof. $\quad\square$

The following lemma is a consequence of Theorem 6 with zero reward.

Lemma 11 (Zero reward regret). *In* (11) *set* $\epsilon = 1/K$, *then with probability at least* $1 - \delta$, *the total value (regret) of Algorithm* 2 *is bounded by*

$$\sum_{k=1}^{K} \overline{V}_1^k(x_1) \lesssim \sqrt{(d \wedge S) \cdot H^3 SAK\iota} + H^2 S^2 A\iota^2, \quad \iota := \log(HSAK/\delta).$$

Proof. Use Theorem 6 with preference/reward set to be zero. □

We are ready to prove Theorem 3.

Theorem 8 (Restatement of Theorem 3). *Consider Algorithm* 2 *and Algorithm* 3 *with bonus function* (11) *and* $\epsilon = 1/K$, *and suppose that Algorithm* 2 *is run for*

$$K \approx \frac{(d \wedge S) \cdot H^3 SA\iota}{\epsilon^2} + \frac{H^2 S^2 A\iota^2}{\epsilon}, \quad \iota := \log \frac{HSA}{\delta\epsilon}$$

episodes, then Algorithm 3 *outputs an* (ϵ, δ)-*PAC policy for preference-free exploration.*

Proof. Recall that in the planning phase, we uniformly sample one of the K policies as our final policy. Thus, with probability at least $1 - \delta$,

$$
\begin{aligned}
V_1^*(x_1; w) - V_1^\pi(x_1; w) &= \frac{1}{K} \sum_{k=1}^{K} \left(V_1^*(x_1; w) - V_1^{\pi^k}(x_1; w) \right) \\
&\leq \frac{1}{K} \sum_{k=1}^{K} \left(V_1^k(x_1; w) - V_1^{\pi^k}(x_1; w) \right) \quad \text{(by Lemma 4)} \\
&\leq \frac{e}{K} \sum_{k=1}^{K} \overline{V}_1^k(x_1) \quad \text{(by Lemma 10)} \\
&\lesssim \sqrt{\frac{(d \wedge S) \cdot H^3 SA\iota}{K}} + \frac{H^2 S^2 A\iota^2}{K}, \quad \text{(by Lemma 11)}
\end{aligned}
$$

where $\iota = \log(HSAK/\delta)$, for the right hand side to be bounded by ϵ, it suffices to set

$$K \approx \frac{(d \wedge S) \cdot H^3 SA}{\epsilon^2} \log \frac{HSA}{\delta\epsilon} + \frac{H^2 S^2 A}{\epsilon} \log^2 \frac{HSA}{\delta\epsilon}.$$

□

C.1 Proof of Corollary 5

For the first conclusion, we simply set $d = SA$ and apply Theorem 3.

For the second conclusion about non-stationary MDPs, the proof logic follows with small revisions.

First, as the MDP is non-stationary, we replace $\mathbb{P}$ by $\{\mathbb{P}_h\}_{h=1}^H$. Similar in the algorithms, we need to replace $N^k(x, a)$ by $\{N_h^k(x, a)\}_{h=1}^H$, which represents the number of visits to (x, a) at step h upto episode k. Then the empirical transition probability will be estimated by $\widehat{\mathbb{P}}_h^k(y \mid x, a) = N_h^k(x, a, y)/N_h^k(x, a)$, for $N_h^k(x, a) \geq 1$.

Second, note that for stationary MDP, we have $\sum_{(x,a)} N^K(x, a) = HK$, but for non-stationary MDP, this needs to be replaced by $\sum_{(x,a)} N_h^K(x, a) = K$ for every h. Then for non-stationary MDP, the integration tricks used in the proof of Theorem 6 needs to be revised accordingly. Let us take

term (iv) as an example, the revised analysis should be:

$$(\text{iv}) = \sum_{k=1}^{K} \sum_{h=1}^{H} \sum_{x,a} w_h^k(x,a) \cdot \sqrt{\frac{(d \wedge S)H^2 \iota}{N_h^k(x,a)}} \cdot \mathbb{1}\left[\sum_{j<k} w_h^j(x,a) \geq 2\iota\right] \qquad \text{(by a revised (3))}$$

$$\lesssim \sum_{h=1}^{H} \sum_{x,a} \sum_{k=1}^{K} w_h^k(x,a) \cdot \sqrt{\frac{(d \wedge S)H^2 \iota}{\sum_{j<k} w_h^j(x,a) - \iota}} \cdot \mathbb{1}\left[\sum_{j<k} w^j(x,a) \geq 2\iota\right] \quad \text{(by a revised (G_{N}))}$$

$$\lesssim \sqrt{(d \wedge S)H^2 \iota} \cdot HSA \cdot \sqrt{K} = \sqrt{(d \wedge S)H^4 SAK\iota}. \qquad \text{(integration trick)}$$

Therefore the obtained bound has an enlarged H dependence.

Last, repeating the proof of Theorem 3 with the revised notations and the above revised inequalities, we obtain the exploration trajectory complexity for non-stationary MDP, where the order of H is enlarged due to the revised Theorem 6.

D Proof of Theorem 4

In this section, we follow notations in Algorithm 2 and Algorithm 3 and prove Theorem 4. Our construction of hard instance is based on the results from Jin et al. [2020].

The following two definitions are migrated from reward-free exploitation (RFE) [Jin et al., 2020] to preference-free exploitation (PFE) in MORL.

Definition 1 (Set of MOMDPs). Fix $\mathcal{S}, \mathcal{A}, H, \boldsymbol{r}, \mathcal{W}$ as the state sets, action sets, length of horizon, reward vector, preferences set, respectively. Then a set of transition probabilities $\mathscr{P}$ induces *a set of MOMDPs*, denoted as $(\mathcal{S}, \mathcal{A}, H, \mathscr{P}, \boldsymbol{r}, \mathcal{W}) := \{(\mathcal{S}, \mathcal{A}, H, \mathbb{P}, \boldsymbol{r}, \mathcal{W}) : \mathbb{P} \in \mathscr{P}\}$.

Definition 2 ((ϵ, p)-correctness). We say a PFE algorithm is (ϵ, p)-correct for a set of MOMDPs $(\mathcal{S}, \mathcal{A}, H, \mathscr{P}, \boldsymbol{r}, \mathcal{W})$, if with probability at least $1 - p$,

$$V_1^{\pi_w}(x_1; w; \mathbb{P}) \geq V_1^*(x_1; w; \mathbb{P}) - \epsilon, \quad \text{for all } w \in \mathcal{W} \text{ and } \mathbb{P} \in \mathscr{P},$$

where π_w is the policy given by the PFE algorithm for preference $w \in \mathcal{W}$.

Basic hard instance. We specify the following set of MOMDPs $(\mathcal{S}, \mathcal{A}, 2, \mathscr{P}_{\texttt{basic}}(s, \epsilon), \boldsymbol{r}, \mathcal{W})$ as a basic hard instance:

- A state set $\mathcal{S} := \{s\} \cup [d]$, where s is the initial state, and $|\mathcal{S}| = d + 1$.
- An action set $\mathcal{A} := [A]$.
- A horizon length $H = 2$.
- A d-dimensional action-independent reward vector, described by a matrix $\boldsymbol{r} := (0, I_{d\times d}) \in \mathbb{R}^{d \times (d+1)}$, i.e., the i-th row $\boldsymbol{r}^{(i)}(\cdot) \in \mathbb{R}^{d+1}$ sets reward 1 at state i, and reward 0 elsewhere.
- A preference set $\mathcal{W} := \left\{w \in \mathbb{R}^d, \|w\|_1 = 1\right\}$. For $w \in \mathcal{W}$, $w^\top \boldsymbol{r} \in \mathbb{R}^{d+1}$ gives a scalarized reward.
- A set of transition probabilities

$$\mathscr{P}_{\texttt{basic}}(s, \epsilon) := \left\{\mathbb{P} : \forall a \in [A], \, i \in [d], \, \left|\mathbb{P}(i \,|\, s, a) - \frac{1}{d}\right| \leq \frac{\epsilon}{d}, \, \mathbb{P}(i \,|\, i, a) = 1\right\}.$$

For $\mathbb{P} \in \mathscr{P}_{\texttt{basic}}(s, \epsilon)$, the initial state is always set to be s; then the agent takes an action and transits to $[d]$ with a near uniform distribution; finally, $[d]$ are all absorbing states.

For the above basic hard instance, Jin et al. [2020] gives the following lemma to characterizes its learning complexity.

Lemma 12 (Lower bound on the sample complexity of PFE for the basic hard instance). *Fix $\epsilon \leq 1, p \leq \frac{1}{2}, A \geq 2$, suppose $d \geq \mathcal{O}(\log A)$. There exists a distribution $\mathcal{D}$ over $\mathscr{P}_{\texttt{basic}}(s, \epsilon)$, such that any $(\epsilon/12, p)$-correct PFE algorithm* `alg` *for $(\mathcal{S}, \mathcal{A}, 2, \mathscr{P}_{\texttt{basic}}, \boldsymbol{r}, \mathcal{W})$ must satisfy*

$$\mathbb{E}_{\mathbb{P} \sim \mathcal{D}} \mathbb{E}_{\mathbb{P}, \texttt{alg}}[K] \geq \Omega\left(\frac{dA}{\epsilon^2}\right),$$

where K is the number of trajectories collected by `alg` *in the exploration phase.*

Proof. See Jin et al. [2020], Lemma D.2. Note that by the construction of the basic hard instance, the preference-weighted reward recovers any reward that is action-independent, thus for the basic hard instance defined in the language of PFE, it equals to the basic hard instance studied by Jin et al. [2020], Lemma D.2 for RFE. □

Hard instance in full version. Based on the basic hard instance, we next build the hard instance in full version that witnesses the lower bound in Theorem 4. Let $n = 2^{\ell_0}$. The hard instance is a set of MOMDPs $(\mathcal{S}, \mathcal{A}, H, \mathscr{P}, \boldsymbol{r}, \mathcal{W})$ specified as follows:

- A state set
$$\mathcal{S} := \left\{ (s, \ell) : s \in [2^\ell], \; \ell = 0, \ldots, \ell_0 \right\} \bigcup \left\{ (s, \ell_0 + 1) : s \in [d] \right\},$$
where $(0, 0)$ is the initial state. The states can be viewed as a ℓ_0-layer binary tree with d leaves attached to the last layer. Clearly, $|\mathcal{S}| = 2^{\ell_0+1} - 1 + d = 2n - 1 + d = \Theta(n)$.

- An action set $\mathcal{A} := [A]$.
- A horizon length H.
- A $2d$-dimensional action-independent reward vector, described by a matrix
$$\boldsymbol{r} := \begin{pmatrix} 0_{d \times (n-1)} & A_{d \times n} & 0_{d \times d} \\ 0_{d \times (n-1)} & 0_{d \times n} & I_{d \times d} \end{pmatrix} \in \mathbb{R}^{2d \times (2n-1+d)},$$
where $A \in \mathbb{R}^{d \times n}$ is a fixed matrix that we will specify later.

- A preference set $\mathcal{W} := \mathcal{W}_A \oplus \mathcal{W}_{\texttt{basic}} \subset \mathbb{R}^{2d}$. Here $\mathcal{W}_{\texttt{basic}} \subset \mathbb{R}^d$ is the preference set for the basic hard instance and gives the set of the last d-dimensions of a preference vector. And we define $\mathcal{W}_A := \{ Ae_i, \; e_i \in \mathbb{R}^n, \; i = 1, \ldots, n \} \subset \mathbb{R}^d$, which gives the set of the first d-dimensions of a preference vector. Here e_i is the i-th standard coordinate vector for $\mathbb{R}^n$. Then $w^\top \boldsymbol{r} \in \mathbb{R}^{2n-1+d}$ gives a scalarized reward. The first $2n - 1$ dimensions of the scalarized reward $w^\top \boldsymbol{r}$ are always zero, i.e., the reward for the states in the $0, \ldots, \ell_0$-th layers is always zero. The last d dimensions of the scalarized reward is sampled from $\{ A^\top A e_i, \; i = 1, \ldots, n \}$, which specifies the reward for states in the $(\ell_0 + 1)$-th layer.

- A set of transition probabilities
$$\mathscr{P} := \big\{ \; \mathbb{P}\left(x_{\ell+1} = (s, \ell + 1) \,\middle|\, x_\ell = (s, \ell), a = 1 \right) = 1, \; \ell = 0, \ldots, \ell_0 - 1,$$
$$\mathbb{P}\left(x_{\ell+1} = (2^\ell + s, \ell + 1) \,\middle|\, x_\ell = (s, \ell), a > 1 \right) = 1, \; \ell = 0, \ldots, \ell_0 - 1;$$
$$\mathbb{P}\left(x_{\ell_0+1} \,\middle|\, x_{\ell_0} = (s, \ell_0), a \right) = \mathbb{P}^s_{\texttt{basic}}, \; \mathbb{P}^s_{\texttt{basic}} \in \mathscr{P}_{\texttt{basic}}(s, \epsilon) \big\}.$$
Notice the transition probabilities up to the ℓ_0-th layer are fixed. And the transition probability from the ℓ_0-th layer to the $(\ell_0 + 1)$-th layer is specified by n multiples of the transition probabilities that are defined earlier for the basic hard instance. In sum, a transition probability $\mathbb{P} \in \mathscr{P}$ corresponds to n multiples of transition probabilities $\mathbb{P}^s_{\texttt{basic}} \in \mathscr{P}_{\texttt{basic}}(s, \epsilon)$, $s \in [n]$.

The following lemma is the key ingredient for obtaining lower bound for PFE.

Lemma 13. *Let $e_1, \ldots, e_n$ be the standard coordinate vector for $\mathbb{R}^n$. Let e_0 be the zero vector. Suppose $d \geq \frac{200}{\epsilon_1^2} \log(n+1)$. Then there exists a matrix $A \in \mathbb{R}^{d \times n}$ such that*
$$\left\| A^\top A e_i - e_i \right\|_\infty \leq \epsilon_1, \quad i = 1, \ldots, n.$$

Proof. For $d \geq \frac{8}{\epsilon^2} \log(n+1)$, by Johnson-Lindenstrauss lemma [Johnson and Lindenstrauss, 1984], there exists matrix $A \in \mathbb{R}^{d \times n}$, such that
$$(1 - \epsilon) \| e_i - e_j \|_2 \leq \| A e_i - A e_j \|_2 \leq (1 + \epsilon) \| e_i - e_j \|_2, \quad \forall i, j \in \{0, 1, \ldots, n\}.$$
Thus for $i \neq j$,
$$\langle A e_i, A e_j \rangle = \frac{1}{2} \left(\| A(e_i - e_0) \|_2^2 + \| A(e_j - e_0) \|_2^2 - \| A(e_i - e_j) \|_2^2 \right)$$
$$\begin{cases} \leq \frac{1}{2} \left(2(1+\epsilon)^2 - 2(1-\epsilon)^2 \right) = 4\epsilon, \\ \geq \frac{1}{2} \left(2(1-\epsilon)^2 - 2(1+\epsilon)^2 \right) = -4\epsilon. \end{cases}$$

And for $i = j$,

$$\langle Ae_i, Ae_i \rangle = \|A(e_i - e_0)\|_2^2 \begin{cases} \leq (1+\epsilon)^2 \leq 1 + 3\epsilon, \\ \geq (1-\epsilon)^2 \geq 1 - 3\epsilon. \end{cases}$$

In sum

$$\delta_{i,j} - 4\epsilon \leq \langle A^\top Ae_i,\ e_j \rangle \leq \delta_{i,j} + 4\epsilon.$$

Note that $A^\top Ae_i = \sum_{j=1}^n \langle A^\top Ae_i, e_j \rangle$. Thus the above inequality implies $\left\| A^\top Ae_i - e_i \right\|_\infty \leq 4\epsilon$. A rescale of ϵ completes the proof. $\qquad\square$

We can then extend results from Jin et al. [2020] to PFE using Lemma 13.

Lemma 14 (Generalized Lemma D.3 of Jin et al. [2020]). *Fix $\epsilon < \frac{1}{4}$, $\epsilon_0 < \frac{1}{16H}$ and $\epsilon_1 < \frac{1}{16}$. Consider $\mathbb{P} \in \mathscr{P}$ and $w = (Ae_s,\ v) \in \mathcal{W}$, where $v \in \mathbb{R}^d$ is nearly uniform, i.e.,*

$$\left\| v - \frac{1}{d}\mathbb{1} \right\|_\infty \leq \frac{\epsilon_0}{d}.$$

Then if policy π is ϵ-optimal, i.e., $V_1^\pi(\mathbb{P}, w) \geq V_1^*(\mathbb{P}, w) - \epsilon$, it must visit state (s, ℓ_0) with probability at least $\frac{1}{2}$, i.e.,

$$\mathbb{P}^\pi[x_{\ell_0} = (s, \ell_0)] \geq \frac{1}{2}.$$

Proof. From the preference vector we compute the scalarized reward as $r = \left(0, A^\top Ae_s, v \right)$, i.e., states in the $0, \ldots, (\ell_0 - 1)$-th layers has zero reward, states in the ℓ_0-th layer has reward as $A^\top Ae_s$, which takes value nearly 1 at state (s, ℓ_0) and value nearly 0 at other states, and states in the last layer has nearly uniform reward v.

Let $\bar{v} := \frac{1}{d}\sum_{y=1}^d v[y]$ be the expected reward in the final layer under a uniform visiting distribution. Then by construction we have $|v(y) - \bar{v}| \leq \epsilon_0$, which implies $|v(y) - v(y')| \leq 2\epsilon_0$. Therefore,

$$|\mathbb{P}v(\cdot) - \mathbb{P}'v(\cdot)| \leq 2\epsilon_0, \text{ for any two transition probability } \mathbb{P} \text{ and } \mathbb{P}'.$$

Now we can compute the value of a policy π:

$$
\begin{aligned}
V_1^\pi &= \langle \mathbb{P}^\pi[(\cdot, \ell_0)], A^\top Ae_s(\cdot) \rangle + (H - \ell_0 - 1)\langle \mathbb{P}^\pi[(\cdot, \ell_0 + 1)], v(\cdot) \rangle \\
&\quad \text{(since the last layer is absorbing)} \\
&\leq (1 + \epsilon_1)\mathbb{P}^\pi[(s, \ell_0)] + \epsilon_1(1 - \mathbb{P}^\pi[(s, \ell_0)]) + (H - \ell_0 - 1)\langle \mathbb{P}^\pi[(\cdot, \ell_0 + 1)], v(\cdot) \rangle \\
&\quad \text{(by Lemma 13)} \\
&\leq \epsilon_1 + \mathbb{P}^\pi[(s, \ell_0)] + (H - \ell_0 - 1)\langle \mathbb{P}^\pi[(\cdot, \ell_0 + 1)], v(\cdot) \rangle.
\end{aligned}
$$

On the other hand we can lower bound the value of the optimal policy by a policy such that $x_{\ell_0} = (s, \ell_0)$ is taken with probability 1 (this is doable since the transition probability before the ℓ_0-th layer is deterministic):

$$V_1^* \geq A^\top Ae_s(s) + (H - \ell_0 - 1)\langle \mathbb{P}^*[(\cdot, \ell_0 + 1)], v(\cdot) \rangle \geq 1 - \epsilon_1 + (H - \ell_0 - 1)\langle \mathbb{P}^*[(\cdot, \ell_0 + 1)], v(\cdot) \rangle.$$

In sum

$$
\begin{aligned}
\frac{1}{4} &\geq \epsilon \geq V_1^* - V_1^\pi \geq 1 - \mathbb{P}^\pi[(s, \ell_0)] - 2\epsilon_1 - 2(H - \ell_0 - 1)\epsilon_0 \\
&\geq 1 - 2\epsilon_1 - 2H\epsilon_0 - \mathbb{P}^\pi[(s, \ell_0)] \geq \frac{3}{4} - \mathbb{P}^\pi[(s, \ell_0)].
\end{aligned}
$$

$\qquad\square$

Lemma 15 (Lemma D.4 of Jin et al. [2020]). *Suppose $H \geq 2(\ell_0 + 1)$. Then a PFE algorithm* `alg` *that is (ϵ, p)-correct for the hard instance induces n PFE algorithms* `alg`^{s}, $s \in [n]$, *which are all $\left(\frac{\epsilon}{4H}, p \right)$-correct for the basic hard instance.*

Proof. Suppose policy π satisfies $V_1^\pi(\mathbb{P}, w) \geq V_1^*(\mathbb{P}, w) - \epsilon$, for all $\mathbb{P} \in \mathscr{P}$ and $w \in \mathcal{W}$. Then by setting $w = \left(0, A^T A e_s, v\right)$ for $s = 1, \ldots, n$, we obtain n sub-policies, each of which visits a corresponding state from $(1, \ell_0), \ldots, (n, \ell_0)$ with probability at least $\frac{1}{2}$ by Lemma 14 and induces a near-optimal policy for the basic hard instance. The last claim is since

$$
\begin{aligned}
\epsilon &\geq V_1^*(\mathbb{P}, w) - V_1^\pi(\mathbb{P}, w) \\
&\geq \mathbb{P}^\pi\left[x_{\ell_0} = (s, \ell_0)\right] \cdot (H - \ell_0 - 1) \cdot (V_1^*(s; \mathbb{P}_{\mathtt{basic}}^s, v) - V_1^\pi(s; \mathbb{P}_{\mathtt{basic}}^s, v)) \\
&\geq \frac{1}{2}(H - \ell_0 - 1) \cdot (V_1^*(s; \mathbb{P}_{\mathtt{basic}}^s, v) - V_1^\pi(s; \mathbb{P}_{\mathtt{basic}}^s, v)) \qquad \text{(by Lemma 14)} \\
&\geq \frac{H}{4} \cdot (V_1^*(s; \mathbb{P}_{\mathtt{basic}}^s, v) - V_1^\pi(s; \mathbb{P}_{\mathtt{basic}}^s, v)), \qquad \text{(since } H \geq 2(\ell_0 + 1)\text{)}
\end{aligned}
$$

which implies $V_1^\pi(s; \mathbb{P}_{\mathtt{basic}}^s, v) \geq V_1^*(s; \mathbb{P}_{\mathtt{basic}}^s, w) - \frac{\epsilon}{4H}$ holds with probability p.

$\square$

We are now ready to state Theorem 4 formally and deliver the proof.

Theorem 9 (Restatement of Theorem 4). *Fix ϵ, p. There exists a set of MOMDPs induced by a set of transition probabilities $\mathscr{P}$, and a distribution $\mathcal{D}$ over $\mathscr{P}$, such that if a PFE algorithm (alg) is (ϵ, p)-correct for the set of MOMDPs, then the number of trajectories K that alg needs to collect in the exploitation phase must satisfy*

$$
\mathbb{E}_{\mathbb{P} \sim \mathcal{D}} \mathbb{E}_{\mathbb{P}, \mathtt{alg}}[K] \geq \Omega\left(\frac{\min\{d, S\} \cdot H^2 S A}{\epsilon^2}\right).
$$

Proof. Let K^s be the number of visits to state (s, ℓ_0) by alg. Then according to Lemma 15, there are n induced algorithms $\mathtt{alg}^s, s \in [n]$ that is $(\frac{\epsilon}{4H}, p)$-correct for the basic hard instance, and each of them collects k^s number of trajectories.

However, by Lemma 12 there exists a distribution $\mathcal{D}^s$ over $\mathscr{P}_{\mathtt{basic}}(s, \frac{\epsilon}{4H})$ such that

$$
\mathbb{E}_{\mathbb{P} \sim \mathcal{D}^s} \mathbb{E}_{\mathbb{P}, \mathtt{alg}^s}[K^s] \geq \Omega\left(\frac{dAH^2}{\epsilon^2}\right).
$$

Notice that during each episode, alg can and only can visit one of (s, ℓ_0) in the hard instance, thus we have $K = \sum_{s \in [n]} K^s$. Making a summation we obtain

$$
\mathbb{E}[K] \geq \Omega\left(\frac{d \cdot nAH^2}{\epsilon^2}\right) = \Omega\left(\frac{dSAH^2}{\epsilon^2}\right).
$$

On the other hand, clearly in our construction, we can set $d = n$ where Lemma 13 holds trivially. Then by the same procedure we obtain

$$
\mathbb{E}[K] \geq \Omega\left(\frac{n \cdot nAH^2}{\epsilon^2}\right) = \Omega\left(\frac{S^2 AH^2}{\epsilon^2}\right).
$$

In sum we have

$$
\mathbb{E}[K] \geq \Omega\left(\frac{\min\{d, S\} \cdot SAH^2}{\epsilon^2}\right).
$$

$\square$

E Proof of Theorem 2

We now prove Theorem 2 based on Theorem 9.

Theorem 10 (Restatement of Theorem 2). *Fix S, A, H. Suppose $K > 0$ is sufficiently large. Then for any algorithm that runs for K episodes, there exists a set of MDPs $\mathscr{P}$ and a sequence of (necessarily adversarially chosen) preferences $\{w^1, \ldots, w^K\}$, such that*

$$
\mathbb{E}_{\mathbb{P} \sim \mathscr{P}, \mathtt{alg}}[\mathtt{regret}(K)] \geq \frac{1}{c} \cdot \sqrt{\min\{d, S\} H^2 SAK}
$$

for some absolute constant $c > 1$, where the expectation is taken with respect to the randomness of drawing an MDP and an algorithm collecting a dataset from the chosen MDP during the exploration phase.

Proof. Let us fix d, S, A, H and a MDP structure as described in the proof of Theorem 9. Let K to be a large number, and $\epsilon := \sqrt{\min\{d, S\} H^2 SA/K_0}$ to be a small number.

We only consider the randomness of (i) choosing an MDP with transition $\mathbb{P}$ and (ii) an algorithm collecting a dataset $\mathcal{H}^K$ from the chosen MDP during the exploration phase. In order words, we will take expectation over all the randomness during the planning phase, in particular, the considered regret is understood as

$$\texttt{regret}(K) := \sum_{k=1}^{K} \mathbb{E}\left[V_1^*(x_1; w^k) - V_1^{\pi^k}(x_1; w^k) \mid \mathbb{P}, \mathcal{H}^K \right].$$

With this in mind, let us consider the following probability measures:

- Let $\boldsymbol{P}_0(\cdot)$ be the probability measure induced by the randomness of drawing a transition kernel $\mathbb{P}$ from the set of transitions $\mathscr{P}$.

- Let $\boldsymbol{P}(\cdot \mid \mathbb{P})$ to denote the conditional probability measure induced by the randomness of running an algorithm $\texttt{alg}$ on a MDP with a fixed transition kernel $\mathbb{P}$, i.e., the probability measure of collecting a dataset.

- Let $\boldsymbol{P}(\cdot)$ be the joint probability measure induced by the randomness of drawing a transition kernel $\mathbb{P}$, and the randomness of collecting a dataset.

Converting online-MORL algorithms into PFE algorithms. Let $\texttt{alg}$ be an MORL algorithm determined by any fixed rules (but could contain random bits), and let $\texttt{adv}$ be an (arbitrary) adversary that provides preferences for $\texttt{alg}$ in the online MORL game. Then we inductively define a preference-free exploration algorithm $\texttt{ALG}(K, \mathbb{P})$, which runs for at most K episodes on an MDP with transition kernel $\mathbb{P}$:

- At episode 1, $\texttt{adv}$ chooses a preference w^1 based on the fixed rule of $\texttt{alg}$; and $\texttt{alg}$ interacts with $\mathbb{P}$ under the guidance of w^1 to collect dataset $\mathcal{H}^1$;

- At episode $k \leq K$, $\texttt{adv}$ chooses a preference w^k based on the fixed rule of $\texttt{alg}$ and the history $\mathcal{H}^{k-1}$; and $\texttt{alg}$ interacts with $\mathbb{P}$ under the guidance of w^k and collect dataset $\mathcal{H}^k = \mathcal{H}^{k-1} \cup \{\text{new empirical observations}\}$.

- At episode $k \leq K$, if the agent receives a sequence of planning preferences $w_1, \ldots, w_n$, the agent outputs a sequence of policies $\pi_1, \ldots, \pi_n$ respectively according to the planning rule that $\texttt{alg}$ adopts at the current episode (which is based on history $\mathcal{H}^{k-1}$). For simplicity, we denote $\texttt{ALG}(\mathbb{P}, k)[\cdot] : \{w\} \to \{\pi\}$ be such a planning process, i.e., $\texttt{ALG}(\mathbb{P}, k)[w_i] = \pi_i$.

We now extend $\texttt{ALG}$ into a "correct" PFE algorithm that stops at finite time:

$$\widetilde{\texttt{ALG}}(\mathbb{P}, k)[\cdot] := \begin{cases} \texttt{ALG}(\mathbb{P}, k)[\cdot], & 1 \leq k \leq K, \\ \texttt{PF-UCB}(\mathbb{P}, k - K)[\cdot], & K < k \leq 2K, \\ \text{output random policy for any preference}, & k > 2K. \end{cases}$$

To justify the correctness of this augmented algorithm, consider two error random variables:

$$\texttt{error}(\mathbb{P}, k, \widetilde{\texttt{ALG}})[w] := \mathbb{E}[V_1^*(w; \mathbb{P}) - V_1^{\pi}(w; \mathbb{P}) \mid \mathbb{P}, \mathcal{H}^{k-1}], \quad \text{for } \pi = \widetilde{\texttt{ALG}}(\mathbb{P}, k)[w],$$

$$\texttt{error}(\mathbb{P}, k, \widetilde{\texttt{ALG}}) := \max_w \texttt{error}(\mathbb{P}, k, \widetilde{\texttt{ALG}})[w],$$

and a stopping time

$$\mathring{K} := (2K) \wedge \min\left\{ k : \texttt{error}(\mathbb{P}, k, \widetilde{\texttt{ALG}}) \leq \epsilon \right\}.$$

Clearly $\mathring{K} \leq 2K$ is bounded.

The augmented PFE algorithm $\widetilde{\mathtt{ALG}}(\mathbb{P}, \mathring{K})[\cdot]$ *is correct.* By construction, we have that for any MDP with fixed transition kernel, if $\widetilde{\mathtt{ALG}}$ is run for $2K$ episodes, its output becomes the same output of $\mathtt{PF\text{-}UCB}$ that runs for K episodes, which is $(\epsilon, \delta \leq 0.9)$-correct for any preference according to Theorem 8. Mathematically, for any $\mathbb{P}$,

$$\boldsymbol{P}\left\{\mathtt{error}(\mathbb{P}, 2K, \widetilde{\mathtt{ALG}}) > \epsilon \mid \mathbb{P}\right\} < \delta \leq 0.9,$$

then over the randomness of choosing a transition kernel $\mathbb{P} \sim \boldsymbol{P}_0$, we have that

$$\boldsymbol{P}\left\{\mathtt{error}(\mathbb{P}, 2K, \widetilde{\mathtt{ALG}}) > \epsilon\right\} = \mathbb{E}_{\mathbb{P}\sim\boldsymbol{P}_0}\left[\boldsymbol{P}\left\{\mathtt{error}(\mathbb{P}, 2K, \widetilde{\mathtt{ALG}}) > \epsilon \mid \mathbb{P}\right\}\right] < 0.9.$$

By discussing whether $\mathring{K} < 2K$ (where the error is smaller than ϵ with probability 1) or $\mathring{K} = 2K$ (where we apply the above inequality), we have that

$$\boldsymbol{P}\left\{\mathtt{error}(\mathbb{P}, \mathring{K}, \widetilde{\mathtt{ALG}}) > \epsilon\right\} < 0.9.$$

This implies that $\widetilde{\mathtt{ALG}}$ that runs for $\mathring{K}$ episodes is $(\epsilon, 0.9)$-correct for the set of MDP described in the proof of Theorem 9. Then by Theorem 9, we have that

$$\mathbb{E}[\mathring{K}] \geq \frac{1}{c} \cdot \frac{\min\{d, S\} H^2 SA}{\epsilon^2} = \frac{1}{c} \cdot K,$$

for some absolute constant $c > 1$. On the other hand,

$$\frac{K}{c_1} \leq \mathbb{E}[\mathring{K}] = \mathbb{E}\left[\mathbb{1}\left[\mathring{K} > \frac{K}{2c}\right] \cdot \mathring{K}\right] + \mathbb{E}\left[\mathbb{1}\left[\mathring{K} \leq \frac{K}{2c}\right] \cdot \mathring{K}\right] \leq 2K \cdot \boldsymbol{P}\left\{\mathring{K} > \frac{K}{2c}\right\} + \frac{K}{2c},$$

which implies that

$$\boldsymbol{P}\left\{\mathring{K} > \frac{K}{2c}\right\} \geq \frac{1}{4c},$$

then by the definition of $\mathring{K}$, we have that

$$\boldsymbol{P}\left\{\text{for each } 1 \leq k \leq \frac{K}{2c}, \ \mathtt{error}(\mathbb{P}, k, \widetilde{\mathtt{ALG}}) > \epsilon\right\} \geq \boldsymbol{P}\left\{\mathring{K} > \frac{K}{2c}\right\} \geq \frac{1}{4c},$$

which further yields that for each $1 \leq k \leq \frac{K}{2c_1}$,

$$\mathbb{E}\left[\mathtt{error}(\mathbb{P}, k, \widetilde{\mathtt{ALG}})\right] \geq \frac{1}{4c} \cdot \epsilon, \tag{16}$$

and

$$\mathbb{E}\left[\sum_{k=1}^{K/(2c)} \mathtt{error}(\mathbb{P}, k, \widetilde{\mathtt{ALG}})\right] \geq \frac{1}{4c} \cdot \epsilon \cdot \frac{K}{2c} = \frac{1}{8c^2} \cdot \sqrt{\min\{d, S\} H^2 SAK}. \tag{17}$$

Regret lower bound. Recall the definition of $\mathtt{error}(\mathbb{P}, k, \widetilde{\mathtt{ALG}})$ and note that for $k \leq K/(2c) \leq K$, $\widetilde{\mathtt{ALG}}[\mathbb{P}, k]$ is the same as $\mathtt{ALG}(\mathbb{P}, k)$, which plans in the same way as $\mathtt{alg}$ in the k-th episode.

We note that (17) holds for any adversary that chooses its preferences inductively based on the online MORL algorithm. Now we specify the adversary $\mathtt{adv}$ as follows: for $k \leq K/(2c)$, $\mathtt{adv}$ specifies a preference $w^{\mathtt{adv}}$ based on the rule of $\mathtt{alg}$, the dataset $\mathcal{H}^{k-1}$, and the transition kernel $\mathbb{P}$, such that

$$(w^{\mathtt{adv}})^k := \arg\max_w \mathbb{E}[V_1^*(w; \mathbb{P}) - V_1^\pi(w; \mathbb{P}) \mid \mathbb{P}, \mathcal{H}^{k-1}], \quad \text{for } \pi = \widetilde{\mathtt{ALG}}(\mathbb{P}, k)[w],$$

then

$$\mathtt{error}(\mathbb{P}, k, \widetilde{\mathtt{ALG}})[(w^{\mathtt{adv}})^k] = \mathtt{error}(\mathbb{P}, k, \widetilde{\mathtt{ALG}}).$$

As a consequence, we have that

$$\mathbb{E}[\mathtt{regret}(K)] \geq \mathbb{E}[\mathtt{regret}(K/(2c))] = \mathbb{E}\left[\sum_{k=1}^{K/(2c)} \mathtt{error}(\mathbb{P}, k, \widetilde{\mathtt{ALG}})[(w^{\mathtt{adv}})^k]\right]$$

$$= \mathbb{E}\left[\sum_{k=1}^{K/(2c)} \mathtt{error}(\mathbb{P}, k, \widetilde{\mathtt{ALG}})\right] \geq \frac{1}{8c^2} \cdot \sqrt{\min\{d, S\} H^2 SAK},$$

and a rescaling of the constant completes our proof.

$\square$

F   Discussion on [Zhang et al., 2020a]

There is a technical error in the proof of Lemma 2 in [Zhang et al., 2020a]. In particular, the sequence considered in Page 11, between equations (16) and (17),

$$\left\{ \mathbb{1}\left[k_i \le K\right] \cdot \left[\left[\widehat{\mathbb{P}}^{k_1} - \mathbb{P}_h\right]\left[\overline{V}^{\pi_k}_{h+1} - V^{\pi_k}_{h+1}\right](s,a) + \left(r_h^{k_1} - \mathbb{E}[r_h](s_h^{k_i}, a_h^{k_i})\right)\right]\right\}_{i=1}^{\tau},$$

is not a martingale difference sequence, since π_k depends on the randomness upto episode k, however $k_1, \ldots, k_\tau$ are all no larger than k. Thus Azuma-Hoeffding's inequality cannot be applied for this sequence.

To fix this error, one may consider applying a covering argument and union bound over the value functions, but the obtained bound in their Theorem 1 should be revised to $\widetilde{\mathcal{O}}\left(\log(N)H^5S^2A/\epsilon^2\right)$.