Information-Theoretic Minimax Regret Bounds for Reinforcement Learning based on Duality

Raghav Bongole, Amaury Gouverneur, Borja Rodríguez-Gálvez, Tobias J. Oechtering, and Mikael Skoglund
Division of Information Science and Engineering (ISE)
KTH Royal Institute of Technology
{bongole, amauryg, borjarg, oech, skoglund}@kth.se

Abstract—We study agents acting in an unknown environment where the agent's goal is to find a robust policy. We consider robust policies as policies that achieve high cumulative rewards for all possible environments. To this end, we consider agents minimizing the maximum regret over different environment parameters, leading to the study of minimax regret. This research focuses on deriving information-theoretic bounds for minimax regret in Markov Decision Processes (MDPs) with a finite time horizon. Building on concepts from supervised learning, such as minimum excess risk (MER) and minimax excess risk, we use recent bounds on the Bayesian regret to derive minimax regret bounds. Specifically, we establish minimax theorems and use bounds on the Bayesian regret to perform minimax regret analysis using these minimax theorems. Our contributions include defining a suitable minimax regret in the context of MDPs, finding information-theoretic bounds for it, and applying these bounds in various scenarios.

Index Terms—information theory, reinforcement learning, minimax theorems.

I. INTRODUCTION

The study of reinforcement learning (RL) in adversarial environments has received significant attention in recent years [1]. Traditional RL approaches assume a stochastic environment where the transition probabilities and rewards are fixed but unknown. A common goal in traditional RL is to find a policy that minimizes regret, which is the difference between the optimal expected cumulative reward that could be achieved if the environment was known and the expected cumulative reward achieved by the agent.

In many scenarios, we are interested in finding a robust policy that performs well for all possible environments. We consider an agent that wants to find a robust policy when the environment acts as an adversary, selecting the worst-case environment parameters. For a given class of Markov Decision Processes, the goal of the agent is to choose a (possibly randomized) policy that tries to minimize the regret for the worst case. This regret is referred to as minimax regret. While theoretical bounds for the minimax regret in some settings are established [2], [3], its information-theoretic nature has remained largely unexplored. This research establishes information-theoretic bounds for minimax regret.

In supervised learning, a notable concept is the minimum excess risk (MER) [4], [5]. MER is an algorithm-independent quantity that measures, for a supervised learning problem,

the gap between its fundamental limit and the best possible algorithm that is uncertain about the environment but has access to the data. MER quantifies the gap between the best possible loss incurred when the algorithm has knowledge of the environment parameters and when it only has access to the data. The work by Xu and Raginsky [6] establishes information-theoretic bounds for the MER in various settings.

Another studied notion in supervised learning is the minimax excess risk, as discussed by [7]. In [7], the authors employ minimax theorems to identify conditions under which the minimax duality holds, that is when $\max_x \min_y f(x, y)$ equals $\min_y \max_x f(x, y)$ for a given risk function $f(x, y)$. Then, the authors use bounds established by [6] for MER to bound the minimax excess risk.

The concept of minimax regret has been investigated in various settings within the literature. Lattimore and Szepesvári explore this topic in the context of partial monitoring, deriving information-theoretic bounds in the Bayesian setting and establishing minimax theorems [8]. Then they use these bounds on the Bayesian regret to subsequently bound the minimax regret in finite-action partial monitoring settings. In the context of Markov Decision Processes (MDPs), Buening et al. [9] establish a minimax theorem for finite state and action spaces, assuming that the reward is a known deterministic function of the state and action. In this paper, we extend their results and prove a minimax theorem for continuous states and action spaces given additional conditions such as continuity of regret. Gouverneur et al. [10] establish bounds for the minimum Bayesian regret (MBR) in reinforcement learning. Like the MER, the MBR is an algorithm-independent quantity that measures the gap between the fundamental limit of an RL problem and the best possible cumulative reward that an agent can achieve.

Our research builds on these foundations by focusing on the RL problem in the context of MDPs with finite time horizon. Inspired by [6]–[10], we derive minimax theorems and use the bounds on the minimum Bayesian regret to establish information-theoretic minimax regret bounds. More specifically, we make the following contributions. (i) We define a notion of minimax regret for RL that is suitable for information-theoretic analysis. (ii) Using duality principles, we then establish connections between the minimax regret and the minimum Bayesian regret. (iii) We then derive bounds on minimax regret, demonstrating their applicability in various

scenarios. (iv) Finally, we derive explicit minimax regret bounds for various settings, such as multi-armed bandits, linear bandits, and contextual bandits.

II. NOTATION AND PRELIMINARIES

Sets are denoted by calligraphic letters (e.g., state space $\mathcal{S}$), various notions of regret by fraktur font (e.g., $\mathfrak{R}$) and σ-algebras are denoted by script letters (e.g., $\mathcal{S}$).

Random variables are denoted by capital letters (e.g., Θ), and their realizations by lowercase letters (e.g., θ). The expectation of X is denoted by $\mathbb{E}_X[X]$ (or $\mathbb{E}[X]$ if it is clear from context), and conditional expectation of X given y is given by $\mathbb{E}^y[X]$. $\Delta(\mathcal{X})$ denotes the set of all probability measures on $\mathcal{X}$. Entropy $H(X)$, KL-divergence $D_{KL}(P \parallel Q)$, and mutual information $I(X;Y)$ are defined in standard terms. Consider a Polish space $\mathcal{X}$ equipped with a metric ρ. For probability distributions P and Q on $\mathcal{X}$, the Wasserstein distance is defined as $W(P,Q) := \inf_{D \in \Gamma(P,Q)} \int \rho \, dD$, where $\Gamma(P,Q)$ is the set of joint distributions with marginals P and Q.

III. MODEL AND DEFINITIONS

In the context of reinforcement learning, we consider an agent navigating through an uncertain environment, often modeled as a Markov Decision Process (MDP) [10] with a finite time horizon $T \in \mathbb{N}$ time steps. At each time step $t \in \{1, \ldots, T\}$, the environment is characterized by a state $S_t \in \mathcal{S}$ and the agent selects an action $A_t \in \mathcal{A}$. The state then transitions to S_{t+1}, and the environment produces an outcome $Y_t \in \mathcal{Y}$ that the agent associates with a reward R_t.

More formally, we define a class of MDPs $\mathcal{M}$, parameterized by a random variable $\Theta \in \mathcal{O}$. This class is defined by a state space $\mathcal{S}$, an action space $\mathcal{A}$, a transition kernel $p : \mathcal{S} \times (\mathcal{S} \times \mathcal{A} \times \mathcal{O}) \to [0,1]$ such that $\mathbb{P}_{S_{t+1}|S_t,A_t,\Theta} = p(\cdot, (S_t, A_t, \Theta))$, and an outcome kernel $y : \mathcal{Y} \times (\mathcal{S} \times \mathcal{O}) \to [0,1]$ such that $\mathbb{P}_{Y_t|S_t,\Theta} = y(\cdot, (S_t, \Theta))$, and an initial state prior distribution $\mathbb{P}_{S_1|\Theta}$ such that $S_1 \sim \mathbb{P}_{S_1|\Theta}$. The reward received by the agent is modeled as a deterministic function of the outcome Y_t and the chosen action A_t, denoted by $r(Y_t, A_t)$. We can write the random variable R_t as $R_t = r(Y_t, A_t)$. Thus the class of MDPs $\mathcal{M}$ can be seen as a 6-tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, p, y, r, T)$.

For a fixed θ, we define $\mathcal{M}_\theta$ as the MDP corresponding to θ. The value of θ is unknown to the agent, however, $\mathcal{S}, \mathcal{A}, p, y, r$ and T are known. Learning involves characterizing the environment parameter θ. In the Bayesian setting, the agent is assumed to have access to a known prior $\mathbb{P}_\Theta$. However, our primary interest lies in the frequentist setting, where the environment parameter θ is unknown, and the goal is to characterize bounds for the worst-case parameter θ. We leverage Bayesian reinforcement learning bounds to derive these frequentist bounds.

Let the history up to time t be a sequence of random variables denoted by $H^t \in \mathcal{H}^t$, where $H^t = (H_1, \ldots, H_t)$ and $H_{t+1} = (S_t, A_t, R_t)$. The agent selects actions based on its current state and history, formalized through a policy function $\pi := \{\pi_t : \mathcal{S} \times \mathcal{H}^t \to \mathcal{A}\}_{t=1}^T$. Further, let $\mathcal{P}$ be the set of all policies.

We define the utility, quantifying the cumulative reward the agent expects to achieve by following a given policy π for a given MDP $\mathcal{M}_\theta$.

Definition 1 (Utility). *The utility of an agent in an MDP $\mathcal{M}_\theta$ following a policy π is given by $U_\mathcal{M}(\pi_t, \theta) := \mathbb{E}^\theta \left[\sum_{t=1}^T r(Y_t, \pi(S_t, H^t)) \right]$, where the expectation is taken over the random variables Y_t, S_t and H^t for $t \in \{1, ..T\}$ drawn following policy π.*

To evaluate the efficiency of a policy, we can compare it with the optimal utility, which represents the best possible outcome if the agent has perfect knowledge of the environment parameter θ.

Definition 2 (Optimal Utility). *The optimal utility of an agent in an MDP $\mathcal{M}_\theta$ is the maximum utility that can be attained when the policy function has access to the environment parameters. This quantity is algorithm-independent and is given by $U_\mathcal{M}^*(\theta) := \sup_{\{f_{\theta,t}:\mathcal{S}\to\mathcal{A}\}_{t=1}^T} \mathbb{E}^\theta \left[\sum_{t=1}^T r(Y_t, f_{\theta,t}(S_t)) \right]$.*

The agent's performance is assessed by the regret, which is defined as the difference between the optimal utility and the utility achieved by the agent.

Definition 3 (Regret). *The regret of the agent is the difference between the optimal utility of an MDP and the utility under a particular policy π. It is given by $\mathfrak{R}_\mathcal{M}(\pi, \theta) := U_\mathcal{M}^*(\theta) - U_\mathcal{M}(\pi, \theta)$.*

Our goal is to characterize the optimal worst-case regret, defined as the minimax regret. It evaluates the policy's performance under the most challenging environment parameter.

Definition 4 (Minimax Regret). *The minimax regret, $\mathfrak{M}_\mathcal{M}$, is the regret obtained by an agent following the optimal policy for the worst-case environment parameters. The minimax regret is given by $\mathfrak{M}_\mathcal{M} := \inf_{\mathbb{P}_\Pi \in \Delta(\mathcal{P})} \sup_{\theta \in \mathcal{O}} \mathbb{E}_\Pi [\mathfrak{R}_\mathcal{M}(\Pi, \theta)]$.*

In contrast to minimax regret, Bayesian regret considers a prior distribution over environment parameters, providing an average-case analysis.

Definition 5 (Bayesian Regret). *The Bayesian regret of an MDP following a policy π, denoted by $\mathfrak{BR}_\mathcal{M}(\pi, \mathbb{P}_\Theta)$ is defined as the average regret with respect to a prior distribution of $\mathbb{P}_\Theta$ of random variable Θ. The Bayesian Regret is given by $\mathfrak{BR}_\mathcal{M}(\pi, \mathbb{P}_\Theta) := \mathbb{E}_{\Theta \sim \mathbb{P}_\Theta}[\mathfrak{R}_\mathcal{M}(\pi, \Theta)]$.*

Proposition 1. *Let $\mathbb{P}_\Theta$ be absolutely continuous with respect to μ with density p_Θ. Then, the Bayesian regret can be expressed as*

$$\mathfrak{BR}_\mathcal{M}(\pi, \mathbb{P}_\Theta) = \sup \mathbb{E} \left[\sum_{t=1}^T r(Y_t, f_t(S_t, \Theta)) \right] - \mathbb{E}_\Theta [U_\mathcal{M}(\pi, \Theta)]$$

where the supremum is taken over the function space $\{f_t : \mathcal{S} \times \mathcal{O} \to \mathcal{A}\}_{t=1}^T$.

Proof. By using the linearity of expectation, we can interchange the supremum and the expectation, leading to the following equalities for the Bayesian regret:

$$\mathfrak{BR}_{\mathcal{M}}(\pi, \mathbb{P}_\Theta) = \mathbb{E}_{\Theta \sim \mathbb{P}_\Theta}[\mathfrak{R}_{\mathcal{M}}(\pi, \Theta)] = \int_\theta \mathfrak{R}(\pi, \theta) p_\Theta(\theta) d\mu$$

$$= \int_\theta \sup_{\{f_t: \mathcal{S} \times \mathcal{O} \to \mathcal{A}\}_{t=1}^T} V(f, \theta) p_\Theta(\theta) d\mu - \int_\theta U_{\mathcal{M}}(\pi, \theta) p_\Theta(\theta) d\mu$$

$$= \sup_{\{f_t: \mathcal{S} \times \mathcal{O} \to \mathcal{A}\}_{t=1}^T} \mathbb{E}\left[\sum_{t=1}^T r(Y_t, f_t(S_t, \Theta))\right] - \mathbb{E}_\Theta[U_{\mathcal{M}}(\pi, \Theta)].$$

where $V(f, \theta) := \mathbb{E}^\theta\left[\sum_{t=1}^T r(Y_t, f_t(S_t, \theta))\right]$ and $f = \{f_t\}_{t=1}^T$. $\square$

Further, we define the minimum Bayesian regret as the lowest achievable average regret under any policy, given the prior distribution over the environment parameters.

Definition 6 (Minimum Bayesian Regret). *Let us consider* $\pi := \{\pi_t : \mathcal{S} \times \mathcal{H}^t \to \mathcal{A}\}_{t=1}^T$. *The minimum Bayesian regret (MBR) under a prior* $\mathbb{P}_\Theta$ *is denoted by* $\mathfrak{F}_{\mathcal{M}}(\mathbb{P}_\Theta)$ *and is given by* $\mathfrak{F}_{\mathcal{M}}(\mathbb{P}_\Theta) := \inf_\pi \mathfrak{BR}_{\mathcal{M}}(\pi, \mathbb{P}_\Theta)$.

Finally, we introduce the worst-case minimum Bayesian regret, corresponding to the maximum MBR achievable under any prior. This concept is critical in game-theoretic scenarios, where the environment selects the most challenging prior, and the agent optimizes its policy in response.

Definition 7 (Worst-case MBR). *The Worst-case MBR is the maximum MBR that can be achieved by any prior. It can also be interpreted in a game-theoretic fashion as the Bayesian regret when the environment plays first, choosing the worst prior, and the agent plays next, selecting the best possible policy for that prior. Hence, the worst-case MBR is given by* $\mathfrak{F}_{\mathcal{M}}^* := \sup_{\mathbb{P}_\Theta \in \Delta(\mathcal{O})} \mathfrak{F}_{\mathcal{M}}(\mathbb{P}_\Theta)$.

IV. MINIMAX THEOREMS

In this section, we will state and prove minimax theorems that provide conditions where the worst-case minimum Bayesian regret and the minimax regret are equal.

A. Minimax theorem with continuity conditions

We will adapt the minimax theorem from Cesa-Bianchi et al. [11, Theorem 7.1]. To prove the minimax theorem for regret under these conditions, we need the following to hold:

(i) $\mathbb{E}_\Pi \mathbb{E}_\Theta[\mathfrak{R}_{\mathcal{M}}(\Pi, \Theta)]$ is bounded and real valued.
(ii) $\Delta(\mathcal{P})$ and $\Delta(\mathcal{O})$ are convex sets.
(iii) $\Delta(\mathcal{O})$ is in addition, a compact set.
(iv) $\mathbb{E}_\Pi \mathbb{E}_\Theta[\mathfrak{R}_{\mathcal{M}}(\Pi, .)]$ is continuous with respect to $\mathbb{P}_\Theta$ for a fixed $\mathbb{P}_\Pi$.
(v) $\mathbb{E}_\Pi \mathbb{E}_\Theta[\mathfrak{R}_{\mathcal{M}}(\Pi, \Theta)]$ is convex for fixed $\mathbb{P}_\Pi$ and is concave for fixed $\mathbb{P}_\Theta$.

Theorem 1. *Let* $(\mathcal{O}, \delta_{\mathcal{O}})$ *be a compact metric space and* $(\pi, \delta_{\mathcal{P}})$ *be a metric space of policies. Let* $\Delta(\mathcal{O})$ *and* $\Delta(\mathcal{P})$ *be the set of all Borel probability measures on* $(\mathcal{O}, \delta_{\mathcal{O}})$ *and* $(\pi, \delta_{\mathcal{P}})$ *respectively. Under the conditions* $\mathfrak{R}_{\mathcal{M}}(\pi, \theta)$ *is*

bounded for all $\pi \in \mathcal{P}$ *and* $\theta \in \mathcal{O}$, *and if* $\mathfrak{R}_{\mathcal{M}}(\pi, .)$ *is a continuous function of* θ *for each* π *we have,* $\mathfrak{M}_{\mathcal{M}} = \mathfrak{F}_{\mathcal{M}}^*$.

Proof. The proof follows from Theorem 2 in [7], where they establish minimax duality conditions for excess risk in supervised learning. To prove the minimax theorem in our setting, we verify conditions (i) through (v). Condition (i) holds by assumption. Condition (ii) is satisfied because the set of all Borel probability measures on a metric space is known to form a convex set. For condition (iii), we observe that $\Delta(\mathcal{O})$ is compact with respect to the Prokhorov metric δ_m [12, Proposition 5.3]. Condition (iv) is verified by demonstrating sequential continuity, which is equivalent to continuity in a metric space [13]. Let us define $g(\theta)$ for any fixed θ as $g(\theta) = \mathfrak{R}_{\mathcal{M}}(\pi, \theta)$. We know $g(\theta)$ is both bounded and continuous. If $\mathbb{P}_{\Theta_n}$ converges to $\mathbb{P}_{\Theta_0}$ under the Prokhorov metric δ_m, then $\mathbb{P}_{\Theta_n}$ converges to $\mathbb{P}_{\Theta_0}$ weakly in measure, as $(\mathcal{O}, \delta_{\mathcal{O}})$ is a compact and separable metric space [12, Theorem 4.2]. We define $h(\mathbb{P}_\Theta)$ as the expected value of $g(\Theta)$ under the distribution $\mathbb{P}_\Theta$, i.e., $h(\mathbb{P}_\Theta) = \mathbb{E}_\Theta[g(\Theta)] = \int g(\theta) d\mathbb{P}_\Theta$. Furthermore, let us define the sequence $\{h_n\}_{n \in \mathbb{N}}$, where $h_n = h(\mathbb{P}_{\Theta_n})$. Convergence in measure, combined with the fact that $g(\theta)$ is bounded and continuous, implies that $\int g(\theta) d\mathbb{P}_{\Theta_n}$ converges to $\int g(\theta) d\mathbb{P}_{\Theta_0}$ pointwise. Hence, we conclude that h_n converges to h_0 [12, Theorem 3.2]). The Dominated Convergence Theorem, along with the boundedness and pointwise convergence, ensures that $\lim_{n \to \infty} \mathbb{E}_\Pi[h_n] = \mathbb{E}_\Pi[h_0]$ [14], confirming the continuity of $\mathbb{E}_\Pi \mathbb{E}_\Theta[\mathfrak{R}_{\mathcal{M}}(\Pi, \cdot)]$ with respect to $\mathbb{P}_\Theta$. Condition (v) is satisfied because the expectation is linear with respect to the distribution. $\square$

Theorem 1 establishes that the minimax duality holds even in the case of continuous state and action spaces provided some additional conditions on regret are met. Further, using Theorem 1, we demonstrate that the minimax theorem holds even in the case of stochastic rewards with finite states and actions and with a finite class of MDPs. Also, notably, $\mathcal{Y}$ and $\mathcal{H}$ need not be finite.

Corollary 1. *If* $\mathfrak{R}_{\mathcal{M}}(\pi, \theta)$ *is bounded for all* $\pi \in \mathcal{P}$ *and* $\theta \in \mathcal{O}$, *and if* $\mathcal{S}$, $\mathcal{A}$, *and* $\mathcal{O}$ *are finite, then* $\mathfrak{M}_{\mathcal{M}} = \mathfrak{F}_{\mathcal{M}}^*$.

Proof. Let us equip $\mathcal{P}$ and $\mathcal{O}$ with the discrete metrics $\delta_{\mathcal{P}}$ and $\delta_{\mathcal{O}}$, respectively. Since $\mathcal{O}$ is finite, the space $(\mathcal{O}, \delta_{\mathcal{O}})$ is compact under the discrete metric $\delta_{\mathcal{O}}$. Additionally, $\mathfrak{R}_{\mathcal{M}}(\pi, \cdot)$ is continuous with respect to θ for a fixed π. This follows from the fact that, under the discrete metric, any convergent sequence $\{\theta_n\}_{n \in \mathbb{N}} \subset \mathcal{O}$ eventually becomes constant, implying that $\mathfrak{R}_{\mathcal{M}}(\pi, \theta_n)$ converges. Therefore, all the assumptions of Theorem 1 are satisfied, and we conclude that $\mathfrak{M}_{\mathcal{M}} = \mathfrak{F}_{\mathcal{M}}^*$. $\square$

V. UPPER BOUNDS ON MINIMAX REGRET

Under certain conditions, the minimum Bayesian regret can be bounded above by a quantity dependent on the prior [10]. Specifically, we have the following inequalities: $\mathfrak{F}_{\mathcal{M}}(\mathbb{P}_\Theta) = \inf_{\pi \in \mathcal{P}} \mathbb{E}_\Theta[\mathfrak{R}_{\mathcal{M}}(\pi, \Theta)] = \inf_{\mathbb{P}_\Pi \in \Delta(\mathcal{P})} \mathbb{E}_\Pi \mathbb{E}_\Theta[\mathfrak{R}_{\mathcal{M}}(\Pi, \Theta)] \leq$

$K_1(\mathbb{P}_\Theta)$, where $K_1(\mathbb{P}_\Theta)$ is a quantity determined by the prior $\mathbb{P}_\Theta$. Hence, we can bound the worst-case MBR as follows:

$$\mathfrak{F}_\mathcal{M}^\star = \sup_{\mathbb{P}_\Theta \in \Delta(\mathcal{O})} \inf_{\mathbb{P}_\Pi \in \Delta(\mathcal{P})} \mathbb{E}_\Theta \mathbb{E}_\Pi [\mathfrak{R}_\mathcal{M}(\Pi, \Theta)] \le K_2,$$

where K_2 is a quantity independent of the prior. Furthermore, when the conditions of the minimax theorem are satisfied, the minimax regret itself can be similarly bounded:

$$\begin{aligned}
\mathfrak{M}_\mathcal{M} &= \inf_{\mathbb{P}_\Pi \in \Delta(\mathcal{P})} \sup_{\theta \in \mathcal{O}} \mathbb{E}_\Pi [\mathfrak{R}_\mathcal{M}(\Pi, \theta)] \\
&= \inf_{\mathbb{P}_\Pi \in \Delta(\mathcal{P})} \sup_{\mathbb{P}_\Theta \in \Delta(\mathcal{O})} \mathbb{E}_\Theta \mathbb{E}_\Pi [\mathfrak{R}_\mathcal{M}(\Pi, \Theta)] \\
&= \sup_{\mathbb{P}_\Theta \in \Delta(\mathcal{O})} \inf_{\mathbb{P}_\Pi \in \Delta(\mathcal{P})} \mathbb{E}_\Theta \mathbb{E}_\Pi [\mathfrak{R}_\mathcal{M}(\Pi, \Theta)] \le K_2.
\end{aligned}$$

This shows that, under minimax duality conditions, we can apply bounds on MBR to control the minimax regret. The theorem presented below provides information-theoretic upper bounds on the minimax regret by using the minimum Bayesian regret bounds outlined in [10, Section V] under diverse conditions of the reward function and the probability distributions of states and observations. These bounds are derived using terms associated with a natural Bayesian reinforcement learning approach, specifically the Thompson sampling algorithm.

Theorem 2. *Consider the function $f^\star$ that maximizes the expected utility, defined as*

$$f^\star = \arg\sup_{f = \{f_t : \mathcal{S} \times \mathcal{O} \to \mathcal{A}\}_{t=1}^T} \mathbb{E}\left[\sum_{t=1}^T r(Y_t, f_t(S_t, \Theta))\right].$$

Let $S_t^\star$ and $Y_t^\star$ denote the states and observations at time t when the function $f^\star$ is followed. Additionally, let $\hat{S}_t$, $\hat{Y}_t$, and $\hat{H}_t$ represent the states, observations, and history at time t obtained using the Thompson sampling algorithm [15], [16]. Then, we obtain the following upper bounds for the minimax regret:

1) *If for all $t = 1, \ldots, T$, the random reward obtained by following $f^\star$, $r(\hat{Y}_t, f_t^\star(\hat{S}_t, \theta))$ is σ_t^2-sub-Gaussian under $\mathbb{P}_{\hat{Y}_t, \hat{S}_t | \hat{H}_t = \hat{h}_t}$ for all $\theta \in \mathcal{O}$ and all $\hat{h}_t \in \mathcal{H}_t$, then*

$$\mathfrak{M}_\mathcal{M} \le \sup_{\mathbb{P}_\Theta} \sum_{t=1}^T \mathbb{E}\left[\sqrt{2\sigma_t^2 D_{KL}(\mathbb{P}_{Y_t^\star, S_t^\star | \Theta} \,||\, \mathbb{P}_{\hat{Y}_t, \hat{S}_t | \hat{H}_t})}\right].$$

2) *Suppose that $(\mathcal{Y} \times \mathcal{A})$ is a metric space with metric ρ. If the reward function $r : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$ is L-Lipschitz under the metric ρ, then*

$$\mathfrak{M}_\mathcal{M} \le \sup_{\mathbb{P}_\Theta} L \sum_{t=1}^T \mathbb{E}\left[W(\mathbb{P}_{Y_t^\star, S_t^\star | \Theta}, \mathbb{P}_{\hat{Y}_t, \hat{S}_t | \hat{H}_t})\right].$$

where $W(\cdot, \cdot)$ represents the Wasserstein distance.

Thus, we can find Furthermore, the minimax regret for specific problems can be upper bounded using specialized bounds as derived in [10], [17], and [18].

A. Finite Multi-arm Bandit Problem with bounded reward

The finite Multi-Armed Bandit (MAB) problem with bounded rewards is a specific class of MDPs. Let $\mathcal{A}$ be finite. Formally, the finite MAB problem is defined by the class $\mathcal{B} = (\mathcal{S}, \mathcal{A}, p, y, r, T)$, where $\mathcal{S} = \{s\}$ and Y_t is independent of S_t given Θ, for all t. Moreover, let $\mathcal{O}$ be a

finite set. Let us introduce a prior distribution $\mathbb{P}_\Theta$. We can then apply Corollary 1 as we have a finite set of MDPs with finite states and action spaces.

The bound derived in [10, Corollary 5] can be used to establish an upper bound on the minimum Bayesian regret: $\mathfrak{F}_\mathcal{B}(\mathbb{P}_\Theta) \le \sup_{\mathbb{P}_\Theta \in \Delta(\mathcal{O})} \sqrt{\frac{1}{2}|A|H(A^\star)T)}$, and hence obtain the minimax regret bound: $\mathfrak{M}_\mathcal{B} \le O(\sqrt{|\mathcal{A}| \log |\mathcal{A}|T})$, which matches the upper bound obtained in [19] and exceeds the lower bound only by a $\sqrt{\log |\mathcal{A}|}$ factor [20].

B. Linear Bandits

Consider a linear bandit problem defined by $\mathcal{L} = (\mathcal{S}, \mathcal{A}, p, y, r, T)$, where $S = \{s\}$ and Y_t is independent of S given Θ. The actions are represented as a d-dimensional vector, i.e., $\mathcal{A} \subset \mathbb{R}^d$, and the reward from an action $a \in \mathcal{A}$ satisfies $\mathbb{E}^\theta[r(Y_t, a)] = a^T \theta$. Furthermore, we assume the conditions from Theorem 1 hold. In addition, consider a ball-structured action space and the parameter space $\mathcal{A}, \mathcal{O} \subseteq \mathbb{B}^d(0, 1)$, where $\mathbb{B}^d(0, 1)$ denotes the d-dimensional closed Euclidean unit ball. Under these conditions, an algorithm $\hat{\pi}$-specifically the Thompson sampling algorithm can achieve a regret bound [18] given by: $\mathfrak{BR}_\mathcal{L}(\hat{\pi}, \mathbb{P}_\Theta) \le O(d\sqrt{T \log T})$. This further implies that the MBR can be bounded above, i.e., $\mathfrak{F}_\mathcal{L}(\mathbb{P}_\Theta) \le O(d\sqrt{T \log T})$. Consequently, under the conditions of the minimax theorem, the minimax regret is bounded by $\mathfrak{M}_\mathcal{L} \le O(d\sqrt{T \log T})$, which exceeds the lower bound for this problem [19] by only a $\sqrt{\log T}$ factor.

C. Contextual Bandits

Consider a contextual bandit problem defined by $\mathcal{C} = (\mathcal{S}, \mathcal{A}, p, y, r, T)$. The transition kernel p is such that the next state at time $t + 1$ is independent of both the previous state and the action taken at time t. We also assume the conditions from Corollary 1 hold. Assuming the rewards are bounded in $[0, 1]$, for any contextual bandit problem $\mathcal{C}$, there exists an algorithm $\hat{\pi}$ (specifically, the Thompson sampling algorithm) such that the Bayesian regret after T rounds is bounded as follows [17]: $\mathfrak{BR}_\mathcal{C}(\hat{\pi}, \mathbb{P}_\Theta) \le \sqrt{\frac{|\mathcal{A}|TH(\Theta)}{2}}$. This further implies that the MBR can be bounded above, i.e., $\mathfrak{F}_\mathcal{C}(\mathbb{P}_\Theta) \le \sqrt{\frac{|\mathcal{A}|TH(\Theta)}{2}}$. Therefore, under the conditions of the minimax theorem, the minimax regret is bounded by $\mathfrak{M}_\mathcal{C} \le O\left(\sqrt{|\mathcal{A}|T \log |\mathcal{O}|}\right)$, which matches the optimal rate [21, Section 1.2].

VI. CONCLUSION

This work demonstrates how Bayesian regret bounds can be applied to establish information-theoretic minimax regret bounds for RL problems in the form of Markov decision processes. We extend the minimax theorem to more general spaces beyond finite state and action spaces, deriving minimax regret bounds across various problem settings, including bandit, contextual bandit, and reinforcement learning problems. Our analysis recovers upper bounds for specific problem classes. Future research directions include exploring additional conditions for duality and investigating other suitable regret definitions, including risk functionals.

References

[1] J. Moos, K. Hansel, H. Abdulsamad, S. Stark, D. Clever, and J. Peters, "Robust reinforcement learning: A review of foundations and recent advances," *Machine Learning and Knowledge Extraction*, vol. 4, no. 1, pp. 276–315, 2022.

[2] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.

[3] M. G. Azar, I. Osband, and R. Munos, "Minimax regret bounds for reinforcement learning," in *International conference on machine learning*. PMLR, 2017, pp. 263–272.

[4] L. Györfi, T. Linder, and H. Walk, "Lossless transformations and excess risk bounds in statistical inference," *Entropy*, vol. 25, no. 10, p. 1394, 2023.

[5] H. Hafez-Kolahi, B. Moniri, S. Kasaei, and M. S. Baghshah, "Rate-distortion analysis of minimum excess risk in bayesian learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 3998–4007.

[6] A. Xu and M. Raginsky, "Minimum excess risk in Bayesian learning," *IEEE Transactions on Information Theory*, vol. 68, no. 12, pp. 7935–7955, 2022.

[7] H. Hafez-Kolahi, B. Moniri, and S. Kasaei, "Information-theoretic analysis of minimax excess risk," *IEEE Transactions on Information Theory*, 2023.

[8] T. Lattimore and C. Szepesvári, "An information-theoretic approach to minimax regret in partial monitoring," in *Conference on Learning Theory*. PMLR, 2019, pp. 2111–2139.

[9] T. K. Buening, C. Dimitrakakis, H. Eriksson, D. Grover, and E. Jorge, "Minimax-bayes reinforcement learning," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2023, pp. 7511–7527.

[10] A. Gouverneur, B. Rodríguez-Gálvez, T. J. Oechtering, and M. Skoglund, "An information-theoretic analysis of Bayesian reinforcement learning," in *2022 58th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2022, pp. 1–7.

[11] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*. Cambridge university press, 2006.

[12] O. van Gaans, "Probability measures on metric spaces," *Lecture notes*, 2003.

[13] V. Orevkov, "Equivalence of two definitions of continuity," *Journal of Soviet Mathematics*, vol. 1, no. 1, pp. 92–99, 1973.

[14] R. M. Gray, *Probability, random processes, and ergodic properties*. Springer Science & Business Media, 2009.

[15] D. Russo and B. Van Roy, "An information-theoretic analysis of Thompson sampling," *Journal of Machine Learning Research*, vol. 17, no. 68, pp. 1–30, 2016.

[16] O. Chapelle and L. Li, "An empirical evaluation of Thompson sampling," *Advances in neural information processing systems*, vol. 24, 2011.

[17] A. Gouverneur, B. Rodríguez-Gálvez, T. J. Oechtering, and M. Skoglund, "Thompson sampling regret bounds for contextual bandits with sub-gaussian rewards," in *2023 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2023, pp. 1306–1311.

[18] S. Dong and B. Van Roy, "An information-theoretic analysis for Thompson sampling with many actions," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[19] D. J. Foster, S. M. Kakade, J. Qian, and A. Rakhlin, "The statistical complexity of interactive decision making," *arXiv preprint arXiv:2112.13487*, 2021.

[20] J.-Y. Audibert and S. Bubeck, "Minimax policies for adversarial and stochastic bandits," in *COLT*, 2009, pp. 217–226.

[21] D. Foster and A. Rakhlin, "Beyond UCB: Optimal and efficient contextual bandits with regression oracles," in *International Conference on Machine Learning*. PMLR, 2020, pp. 3199–3210.