

An Information-Theoretic Analysis of Bayesian Reinforcement Learning

Amaury Gouverneur, Borja Rodríguez-Gálvez, Tobias J. Oechtering, and Mikael Skoglund

Division of Information Science and Engineering (ISE)

KTH Royal Institute of Technology

{amauryg, borjarg, oech, skoglund}@kth.se

Abstract—Building on the framework introduced by Xu and Raginsky [1] for supervised learning problems, we study the best achievable performance for model-based Bayesian reinforcement learning problems. With this purpose, we define minimum Bayesian regret (MBR) as the difference between the maximum expected cumulative reward obtainable either by learning from the collected data or by knowing the environment and its dynamics. We specialize this definition to reinforcement learning problems modeled as Markov decision processes (MDPs) whose kernel parameters are unknown to the agent and whose uncertainty is expressed by a prior distribution. One method for deriving upper bounds on the MBR is presented and specific bounds based on the relative entropy and the Wasserstein distance are given. We then focus on two particular cases of MDPs, the multi-armed bandit problem (MAB) and the online optimization with partial feedback problem. For the latter problem, we show that our bounds can recover from below the current information-theoretic bounds by Russo and Van Roy [2].

Index Terms—information-theoretic bounds, Markov decision process, multi-armed bandit problem, reinforcement learning, Bayesian regret, mutual information, Wasserstein distance

I. INTRODUCTION

In model-based reinforcement learning problems [3], [4], an agent interacts sequentially with a dynamic environment by taking actions in order to maximize its long-term performance.

This paper, as most related work in this field, focuses on systems and control objectives that are modeled as finite time horizon *Markov decision processes* (MDPs). At each time $t = 1, \dots, T$, the agent observes the environment state S_t and takes an action A_t following a decision policy φ_t . Independently of the action, the environment produces a random outcome Y_t . The reward is obtained as a deterministic function of the system's outcome and the chosen action, $R_t = r(Y_t, A_t)$. The data is collected in a history $H^{t+1} = (S_1, A_1, R_1, \dots, S_t, A_t, R_t)$ and the system evolves to a state S_{t+1} . The procedure then repeats until the end of the time horizon, $t = T$.

In the Bayesian setting, the MDP model Φ is treated as a random element of some parametric model family, which is drawn according to a prior distribution of the environment parameters Θ . The goal of the agent is to identify a policy that yields the highest expected cumulative reward $\mathbb{E}[\sum_{t=1}^T r(Y_t, A_t)]$ under the uncertainty of these parameters.

This work was supported in part by the Knut and Alice Wallenberg Foundation, the Swedish Foundation for Strategic Research, and the Swedish Research Council under contract 2019-03606.

The decision-making process in Bayesian reinforcement learning is typically more computationally demanding than the frequentist approach, however this setting presents various advantages as it facilitates regularization, handles parameter uncertainty naturally, and provides ways to solve the exploration-exploitation trade-off [5].

Following the work from Xu and Raginsky on Bayesian supervised learning [1], we put aside the computational aspect to study the best achievable performance for model-based Bayesian reinforcement learning. We define the *minimum Bayesian regret* as the difference between the *Bayesian cumulative reward* $R_\phi(\kappa_H)$, defined as the maximum expected cumulative reward attainable by learning from the sequentially collected data, and $R_\phi(\kappa_\Theta)$, the maximum expected cumulative reward that could be reached if the environment parameters were known. We develop information-theoretic upper bounds on the minimum Bayesian regret under various assumptions for the reward function using the relative entropy and the Wasserstein distance.

Structure of the paper: Section II summarizes the contributions of this paper. The notations are introduced in Section III. Section IV presents the different models of decision processes studied and gives the definition of the Bayesian cumulative reward and the minimum Bayesian regret. Section V and VI are devoted to information-theoretic upper bounds on the MBR. Finally, conclusions are presented in Section VII.

II. CONTRIBUTIONS

In this work, inspired by Xu and Raginsky's [1] framework on the study of the best achievable performance of supervised learning problems, we propose an analogous framework for the study of model-based reinforcement learning problems. Our contributions in this regard can be summarized as:

- 1) Developing a theoretical framework of model-based Bayesian MDPs suited for information-theoretic studies.
- 2) Proposing a definition of the minimum Bayesian regret (MBR) for reinforcement learning problems modeled as Markov decision processes.
- 3) Presenting a data processing inequality for the Bayesian cumulative reward in Lemma 1.
- 4) Formulating upper bounds on the MBR for general MDPs based on the relative entropy (Proposition 1) and the Wasserstein distance (Proposition 2). We present particular cases of these bounds for the case of bounded reward

functions in Corollaries 1 and 2, and the tightness of these results are compared in Remark 3.

- 5) Deriving MBR bounds for the multi-armed bandit and for the online optimization with partial feedback problems. In this last setting, we show how our bound recovers *from below* results from Russo and Van Roy [2].

III. NOTATIONS AND PRELIMINARIES

Throughout the paper, random variables X are written in capital letters, their realizations x in lower-case letters, and their set of outcomes \mathcal{X} in calligraphic letters. The probability distributions of a random variable X is denoted as \mathbb{P}_X . When more than one random variable is considered, e.g., X and Y , we use $\mathbb{P}_{X,Y}$ to denote their joint distribution and $\mathbb{P}_X\mathbb{P}_Y$ for their product distribution¹. We write the conditional probability distribution of Y given X as $\mathbb{P}_{Y|X}$, defining a probability distribution $\mathbb{P}_{Y|X=x}$ over \mathcal{Y} for each element $x \in \mathcal{X}$.

We use the underscore notation X_t to represent a random variable at time $t = 1, \dots, T$ and the exponent notation X^t to denote a sequence of random variables $X^t \equiv (X_1, \dots, X_t)$ for $t = 2, \dots, T$. For consistency we let $X^1 \equiv X_1$.

The relative entropy between two probability distributions \mathbb{P} and \mathbb{Q} is defined as $D_{\text{KL}}(\mathbb{P} \parallel \mathbb{Q}) := \int \log\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right)d\mathbb{P}$ if \mathbb{P} is absolutely continuous with respect to \mathbb{Q} and $D_{\text{KL}}(\mathbb{P} \parallel \mathbb{Q}) \rightarrow \infty$ otherwise. The notation $d\mathbb{P}/d\mathbb{Q}$ is the Radon-Nikodym derivative. Similarly, the mutual information between X and Y is defined as $I(X; Y) := D_{\text{KL}}(\mathbb{P}_{X,Y} \parallel \mathbb{P}_X\mathbb{P}_Y)$, and the conditional mutual information between X and Y , given Z , as $I(X; Y|Z) := \mathbb{E}[I(X; Y|Z = z)]$, where $I(X; Y|Z = z) := D_{\text{KL}}(\mathbb{P}_{X,Y|Z=z} \parallel \mathbb{P}_{X|Z=z}\mathbb{P}_{Y|Z=z})$.

Finally, if two probability distributions \mathbb{P} and \mathbb{Q} are defined in a Polish space \mathcal{X} with respect to a metric ρ , then their Wasserstein distance of order $p \geq 1$ is $\mathbb{W}_p(\mathbb{P}, \mathbb{Q}) := \left(\inf_{\mathbb{D} \in \Pi(\mathbb{P}, \mathbb{Q})} \int \rho d\mathbb{D}\right)^{1/p}$, where $\Pi(\mathbb{P}, \mathbb{Q})$ is the set of all couplings of \mathbb{P} and \mathbb{Q} ; i.e., all joint distributions on $\mathcal{X} \times \mathcal{X}$ with marginals \mathbb{P} and \mathbb{Q} . As this work is focused on upper bounds and since by Hölder's inequality $\mathbb{W}_p \leq \mathbb{W}_q$ for all $p \leq q$ [6, Remark 6.6], in what follows, we will only be using the Wasserstein distance of order 1, $\mathbb{W} := \mathbb{W}_1$. For a discrete random variable X , the Shannon entropy is defined as $H(X) := \mathbb{E}[-\log(\mathbb{P}_X(X))]$.

IV. MODEL AND DEFINITIONS

In this section we first introduce formally Markov decision processes. We then present the multi-armed bandit and the online optimization with partial feedback problems two special cases of MDPs. After that, we describe Bayesian cumulative reward and prove that it respects a data-processing inequality. Finally, we define minimum Bayesian regret.

A. Markov Decision Process

In a *Markov decision process* (MDP), at each time step $1, \dots, T$, an agent interacts with the environment by observing

¹Note that this slight abuse of notation does not mean that the product distribution is the product of the distributions.

the system's state $S_t \in \mathcal{S}$ and selecting accordingly an action $A_t \in \mathcal{A}$. The system then produces an outcome $Y_t \in \mathcal{Y}$ which the agent associates with a scalar reward $R_t \in \mathbb{R}$.

In Bayesian reinforcement learning, the environment is completely characterized by a random variable $\Theta \in \mathcal{O}$ with probability distribution \mathbb{P}_Θ . Therefore, an MDP Φ is defined by a transition kernel $\kappa_{\text{trans}} : \mathcal{S} \times (\mathcal{S} \times \mathcal{A} \times \mathcal{O}) \rightarrow [0, 1]$ such that $\mathbb{P}_{S_{t+1}|S_t, A_t, \Theta} = \kappa_{\text{trans}}(\cdot, (S_t, A_t, \Theta))$, an outcome kernel $\kappa_{\text{out}} : \mathcal{Y} \times (\mathcal{S} \times \mathcal{O}) \rightarrow [0, 1]$ such that $\mathbb{P}_{Y_t|S_t, \Theta} = \kappa_{\text{out}}(\cdot, (S_t, \Theta))$, an initial state prior distribution $\mathbb{P}_{S_1|\Theta}$ such that $S_1 \sim \mathbb{P}_{S_1|\Theta}$, and a reward function $r : \mathcal{Y} \times \mathcal{A} \rightarrow \mathbb{R}$. The reward is a deterministic function of the system's outcome and the chosen action, hence there is a reward's kernel $\kappa_{\text{reward}} : \mathcal{B}(\mathbb{R}) \times (\mathcal{S}, \mathcal{A}, \mathcal{O}) \rightarrow [0, 1]$ such that $\mathbb{P}_{R_t|S_t, A_t, \Theta} = \kappa_{\text{reward}}(\cdot, (S_t, A_t, \Theta))$.

The task in Bayesian reinforcement learning is to learn a policy $\varphi = \{\varphi_t : \mathcal{S} \times \mathcal{H}^t \rightarrow \mathcal{A}\}_{t=1}^T$ taking an action A_t based on the current observation S_t and the past collected data H^t , where $H_{t+1} = (S_t, A_t, R_t)$ that maximizes the *cumulative expected reward* $r_\Phi(\varphi) := \mathbb{E}[\sum_{t=1}^T r(Y_t, \varphi_t(S_t, H^t))]$.

B. Static state MDP and multi-armed bandit problem

A *static state* MDP is an MDP whose transition kernel κ_{trans} is such that the system's state remains constant, i.e. $S_t = S$ for all t . We will use the notation Π to refer to such MDP.

A *multi-armed bandit* (MAB) problem can be formalized as a static state MDP whose environment parameters Θ and outcomes Y_t are independent of S . Similarly to an MDP, the task in a MAB problem is to learn a policy $\varphi = \{\varphi_t : \mathcal{S} \times \mathcal{H}^t \rightarrow \mathcal{A}\}_{t=1}^T$ taking an action A_t based on the past collected data H^t , where $H_{t+1} = (A_t, R_t)$ that maximizes the cumulative expected reward $r_\Pi(\varphi) := \mathbb{E}[\sum_{t=1}^T r(Y_t, \varphi_t(S, H^t))]$.

A variant of that problem is the *online optimization problem with partial feedback* studied by Russo and Van Roy [2]. This problem can also be modeled as a static MDP Π with a finite action space \mathcal{A} where at each time $t = 1, \dots, T$, the agent selects an action A_t and observes a "per-action outcome" $Y_{t,A_t} \in \mathcal{Y}'$ giving rise to past collected data H^t with $H_{t+1} = (A_t, Y_{t,A_t})$. The agent associates the "per-action outcome" with a reward $R_t = r'(Y_{t,A_t})$ through a preference function $r' : \mathcal{Y}' \rightarrow \mathbb{R}$. In this setting, the random outcome $Y_t \in \mathcal{Y}$ is the vector formed with all the possible outcomes, $Y_t \equiv \{Y_{t,a}\}_{a \in \mathcal{A}}$ and the reward function $r : \mathcal{Y} \times \mathcal{A} \rightarrow \mathbb{R}$ is a function such that for all $Y_t \in \mathcal{Y}$ and $A_t \in \mathcal{A}$, we have $r(Y_t, A_t) = r'(Y_{t,A_t})$. In this problem, as well, the environment parameters Θ and outcomes Y_t are independent of S .

C. The Bayesian Cumulative Reward

A decision policy that maximizes the expected cumulative reward among all policies is called a *Bayesian decision policy*. The corresponding maximum expected cumulative reward is defined as the *Bayesian cumulative reward*.

Definition 1. *The Bayesian cumulative reward (BCR) of a Markov decision process Φ is defined as $R_\Phi := \sup_\varphi r_\Phi(\varphi)$, where the supremum is taken over the collection φ of all*

decision rules $\varphi_t : \mathcal{S} \times \mathcal{H}^t \rightarrow \mathcal{A}$ such that the expectation is defined.

The notion of Bayesian cumulative reward can be generalized to allow the agent to select an action using some knowledge X^t such that each X_{t+1} is obtained from (S_t, A_t, Y_t, Θ) . In this generalized model, the knowledge X_t is obtained through a knowledge kernel $\kappa_{\text{know}} : \mathcal{X} \times (\mathcal{S} \times \mathcal{A} \times \mathcal{Y} \times \mathcal{O}) \rightarrow [0, 1]$ such that $\mathbb{P}_{X_{t+1}|S_t, A_t, Y_t, \Theta} = \kappa_{\text{know}}(\cdot, (S_t, A_t, Y_t, \Theta))$. Now, let $\varphi = \{\varphi_t : \mathcal{S} \times \mathcal{X}^t \rightarrow \mathcal{A}\}_{t=1}^T$ be a policy in this relaxed setting. Then, the *generalized Bayesian cumulative reward* (also written as BCR when no confusion is possible) of an MDP Φ with knowledge kernel κ_{know} is $R_{\Phi}(\kappa_{\text{know}}) := \sup_{\varphi} r_{\Phi}(\kappa_{\text{know}}, \varphi)$, where

$$r_{\Phi}(\kappa_{\text{know}}, \varphi) := \mathbb{E} \left[\sum_{t=1}^T r(Y_t, \varphi_t(S_t, X^t)) \right]$$

and again the supremum is taken over the collection φ of all decision rules $\varphi_t : \mathcal{S} \times \mathcal{X}^t \rightarrow \mathcal{A}$ such that the expectation above is defined.

Remark 1. Given an MDP Φ , let $\mathcal{X} = \mathcal{S} \times \mathcal{A} \times \mathbb{R}$ and κ_{know} be a kernel such that $X_{t+1} = (S_t, A_t, R_t)$ and denote this kernel κ_{H} . Note that $X_t = H_t$ and $R_{\Phi}(\kappa_{\text{H}}) = R_{\Phi}$.

After defining the generalized Bayesian cumulative reward, one can study the case where the agent has access to some processed information Z_t obtained from the knowledge X^t . Let κ_{process} denote a collection of processing kernels $\{\kappa_{\text{process}, t} : \mathcal{X} \times (\mathcal{X}^t) \rightarrow [0, 1]\}_{t=1}^T$ such that $\mathbb{P}_{Z_t|X^t} = \kappa_{\text{process}, t}(\cdot, (X^t))$ for each $t = 1, \dots, T$. Then the *processed Bayesian cumulative reward* with knowledge kernel κ_{know} and process kernels κ_{process} is $R_{\Phi}(\kappa_{\text{know}}, \kappa_{\text{process}}) := \sup_{\psi} r_{\Phi}(\kappa_{\text{know}}, \kappa_{\text{process}}, \psi)$, where

$$r_{\Phi}(\kappa_{\text{know}}, \kappa_{\text{process}}, \psi) := \mathbb{E} \left[\sum_{t=1}^T r(Y_t, \psi_t(S_t, Z_t)) \right]$$

and the supremum is taken over the collection ψ of all decision rules $\psi = \{\psi_t : \mathcal{S} \times \mathcal{Z} \rightarrow \mathcal{A}\}_{t=1}^T$ such that the expectation above is defined.

D. Data processing inequality for the BCR

An important property of the Bayesian cumulative reward is the data processing inequality (DPI), stating that no amount of processing of the knowledge random variables can increase the cumulative reward. This is formalized in the following lemma.

Lemma 1. Let κ_{U} be a knowledge kernel associated with an MDP Φ and $\kappa_{\text{V}|U}$ a collection of processing kernels. Then, the cumulative Bayesian reward using the knowledge from U is at least as large as the processed Bayesian cumulative reward using the processed knowledge from V . More precisely,

$$R_{\Phi}(\kappa_{\text{U}}) \geq R_{\Phi}(\kappa_{\text{U}}, \kappa_{\text{V}|U})$$

Intuition of the proof. The proof follows by iteratively employing [7, Lemma 3.22] in a similar fashion to [1, Lemma 1]

and taking care that the random objects in the definitions of $R_{\Phi}(\kappa_{\text{U}})$ and $R_{\Phi}(\kappa_{\text{U}}, \kappa_{\text{V}|U})$ follow the distributions described by the dynamics of the MDP Φ and their respective actions φ_t and ψ_t . The complete proof is in appendix A. \square

E. The Minimum Bayesian Regret (MBR)

We define the *fundamental limit of the Bayesian cumulative reward* as the Bayesian cumulative reward for a knowledge kernel such that $X_t = \Theta$, that is when the environment parameters are known to the agent. We denote such a kernel as κ_{Θ} .

Definition 2. The fundamental limit of the Bayesian cumulative reward of a Markov decision process Φ is defined as

$$R_{\Phi}(\kappa_{\Theta}) := \sup_{\{\psi_t\}_{t=1}^T} \mathbb{E} \left[\sum_{t=1}^T r(Y_t, \psi_t(S_t, \Theta)) \right],$$

where the kernel κ_{Θ} is such that $X_t = \Theta$ for all $t = 1, \dots, T$.

Assumption 1. For the rest of the paper, we will assume that the supremum from Definition 2 exists and we will denote by $\psi^* = \{\psi_t^*\}_{t=1}^T$ a policy that achieves it.

We define the gap between this limit and the Bayesian cumulative reward as the *minimum Bayesian regret*.

Definition 3. The minimum Bayesian regret (MBR) of a Markov decision process Φ is defined as

$$\text{MBR}_{\Phi} := R_{\Phi}(\kappa_{\Theta}) - R_{\Phi}(\kappa_{\text{H}}).$$

The MBR characterizes the regret of the optimal decision policy that has access to the collected data, but not to environment parameters, and is therefore an algorithm-independent quantity. It can be interpreted as the inherent difficulty of the reinforcement learning problem resulting from the lack of knowledge about the environment parameters Θ .

V. UPPER BOUNDS ON THE MINIMUM BAYESIAN REGRET

In this section, we start by giving an upper bound of the minimum Bayesian regret in terms of the difference of the fundamental limit of the BCR, $R_{\Phi}(\kappa_{\Theta})$, and the processed BCR with the optimal Bayes parameters' estimator $\mathbb{P}_{\Theta|H^t}$ as the processing kernel and the optimal policy of $R_{\Phi}(\kappa_{\Theta})$. That is, the difference between the best obtainable risk knowing the environment parameters Θ , and the best obtainable risk inferring the parameters with an optimal estimator. This bound, in turn, can be developed into a bound that compares the sum of the individual terms in the optimal trajectory of $R_{\Phi}(\kappa_{\Theta})$ and those obtained with the processing kernels $\mathbb{P}_{\Theta|H^t}$. This way, we can employ similar techniques to those in the literature (e.g., [8], [9]) and bound the MBR in terms of the sum of terms depending on the statistical difference between the distributions of those two trajectories.

A. The Thompson sampling regret

Consider the fundamental limit of BCR, $R_\Phi(\kappa_\Theta)$, and its optimal trajectory ψ^* . A natural algorithm to try to solve an MDP Φ when environment Θ is unknown is to estimate the environment parameters with some processing kernel of the history $\kappa_{\Theta|H}$ and select an optimal action based on such processing. An elegant scenario would be to have the additional information of knowing which is the optimal trajectory ψ^* and to be able to calculate the Bayes optimal estimator $\mathbb{P}_{\Theta|H^t}$ to process the history. In fact, for a static MDP Π , this algorithm is studied in the literature and is known as the Thompson's sampling algorithm [2], [10]–[14]. Therefore, the next lemma shows that the MBR is bounded from above by the difference of $R_\Phi(\kappa_\Theta)$ and the BCR of such an algorithm, $r_\Phi(\kappa_H, \kappa_{\Theta|H}, \psi^*)$.

Lemma 2. *For any MDP Φ , the MBR can be upper bounded as follows,*

$$\text{MBR}_\Phi \leq R_\Phi(\kappa_\Theta) - r_\Phi(\kappa_H, \kappa_{\Theta|H}, \psi^*).$$

Proof. The proof starts by using Lemma 1 to lower bound $R_\Phi(\kappa_H)$ with $R_\Phi(\kappa_H, \kappa_{\Theta|H})$. The last inequality follows from the definition of $R_\Phi(\kappa_H, \kappa_{\Theta|H})$ being the supremum over ψ of $r_\Phi(\kappa_H, \kappa_{\Theta|H}, \psi)$. More precisely,

$$\begin{aligned} \text{MBR}_\Phi &= R_\Phi(\kappa_\Theta) - R_\Phi(\kappa_H) \\ &\leq R_\Phi(\kappa_\Theta) - R_\Phi(\kappa_H, \kappa_{\Theta|H}) \\ &\leq R_\Phi(\kappa_\Theta) - r_\Phi(\kappa_H, \kappa_{\Theta|H}, \psi^*). \end{aligned}$$

□

In what follows, we will use the notations Y_t^* and S_t^* for the outcomes and states obtained from the actions derived from ψ^* , the kernels that describe the MDP Φ , and the knowledge kernel κ_Θ . Similarly we will let \hat{Y}_t , \hat{S}_t and \hat{H}_t be the outcomes, states, and histories obtained from the actions derived from ψ^* , the kernels that describe the MDP Φ with knowledge kernel κ_H , and processing kernels $\kappa_{\Theta|H}$. The following lemma builds on Lemma 2 and shows how the MBR can be written as the sum of the individual differences of the expected rewards obtained following the optimal trajectory $(Y_t^*, S_t^*)_{t=1}^T$ and the trajectory of the aforementioned algorithm $(\hat{Y}_t, \hat{S}_t)_{t=1}^T$ given the history \hat{H}^t .

Unrolling $R_\Phi(\kappa_\Theta)$ and $r_\Phi(\kappa_H, \kappa_{\Theta|H}, \psi^*)$ and using the linearity of the expectation and the law of total expectation reveals that the right-hand side term from Lemma 2 can be written as

$$\sum_{t=1}^T \mathbb{E} \left[\mathbb{E} \left[r(Y_t^*, \psi_t^*(S_t^*, \Theta)) - r(\hat{Y}_t, \psi_t^*(\hat{S}_t, \hat{\Theta}_t)) \middle| \Theta, \hat{\Theta}_t, \hat{H}^t \right] \right]. \quad (1)$$

Remark 2. *The importance of this re-formulation lays in the fact that the first term inside the conditional expectation is distributed according to $\mathbb{P}_{Y^*, S^* | \Theta}$ since (Y^*, S^*) are independent of the history \hat{H}^t when the environment parameters Θ are known. Similarly, the second term is distributed according*

to $\mathbb{P}_{\hat{Y}_t, \hat{S}_t | \hat{H}^t}$ since (\hat{Y}_t, \hat{S}_t) are independent of the sampled parameters $\hat{\Theta}_t$ when the history is known. Both facts follow from the Markov chain $(Y_t^, S_t^*) - \Theta - (\hat{Y}_t, \hat{S}_t) - \hat{H}^t - \hat{\Theta}_t$. Therefore, conditioned on the history \hat{H}^t and the environment parameters $\Theta, \hat{\Theta}_t$, the terms in the sum of (1) are a difference of expectations of random objects which randomness comes from distributions on the same space, which permits us to employ known decoupling techniques to bound these differences in terms of such distributions.*

In the sequel, we use this fact to bound the MBR as the sum of terms depending on the statistical difference between the distributions of the elements from the optimal trajectory $(Y_t^*, S_t^*) | \Theta$ and the trajectory described by the algorithm with the Bayes optimal parameters' estimator $(\hat{Y}_t, \hat{S}_t) | \hat{H}^t$. More precisely, we use the techniques from e.g. [2], [8] when the reward is sub-Gaussian, from e.g. [9], [15] when it is Lipschitz, and from [9] to connect both settings when the reward is bounded.

B. Sub-Gaussian reward functions

We consider arbitrary reward functions $r : \mathcal{Y} \times \mathcal{A} \rightarrow \mathbb{R}$ mapping an outcome and an action to a scalar reward. Under the assumption that the random reward $r(\hat{Y}_t, \psi_t^*(\hat{S}_t, \theta))$ 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 $\hat{h}^t \in \mathcal{H}^t$, the MBR_Φ is bounded by a sum of terms related to the relative entropy between the distribution of the elements of each step of the optimal trajectory, i.e., Y_t^*, S_t^* , and the Thompson's sampled trajectory, i.e., \hat{Y}_t, \hat{S}_t . This is formalized in the following Proposition.

Proposition 1. *If for all $t = 1, \dots, T$, the random reward $r(\hat{Y}_t, \psi_t^*(\hat{S}_t, \theta))$ 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,*

$$\text{MBR}_\Phi \leq \sum_{t=1}^T \mathbb{E} \left[\sqrt{2\sigma_t^2 D_{\text{KL}}(\mathbb{P}_{Y_t^*, S_t^* | \Theta} \| \mathbb{P}_{\hat{Y}_t, \hat{S}_t | \hat{H}^t})} \right].$$

Proof. The proof follows from applying Donsker-Varadhan's inequality [16, Theorem 5.2.1] to (1) using Remark 2 in a similar fashion to [2], [8]. □

C. Lipschitz reward functions

In this subsection, we suppose that the set of outcomes and actions $(\mathcal{Y}, \mathcal{A})$ together with the metric $\rho : (\mathcal{Y} \times \mathcal{A}) \times (\mathcal{Y} \times \mathcal{A}) \rightarrow \mathbb{R}_+$, form a Polish metric space.

Assume that the reward function $r : \mathcal{Y} \times \mathcal{A} \rightarrow \mathbb{R}$ is L -Lipschitz under the metric ρ , that is that $|r(y, a) - r(y', a')| \leq L\rho((y, a), (y', a'))$ for all $y, y' \in \mathcal{Y}$ and $a, a' \in \mathcal{A}$. Under this assumption, the Wasserstein distance can be used to upper bound the minimum Bayesian regret.

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

$$\text{MBR}_\Phi \leq L \sum_{t=1}^T \mathbb{E} \left[\mathbb{W}(\mathbb{P}_{Y_t^*, S_t^* | \Theta}, \mathbb{P}_{\hat{Y}_t, \hat{S}_t | \hat{H}^t}) \right].$$

Proof. The proof follows from applying Kantorovich–Rubinstein duality [6, Remark 6.5] to (1) using Remark 2 analogously to [9], [15]. \square

D. Bounded reward functions

We can obtain upper bounds on the minimum Bayesian regret for bounded reward functions as particular cases of both Proposition 1 and Proposition 2. We will consider without loss of generality reward functions bounded in $[0, 1]$.

First, from Hoeffding’s lemma [17, Theorem 1], we have that if $r : \mathcal{Y} \times \mathcal{A} \rightarrow [0, 1]$ then the reward is 1/4-sub-Gaussian under any distribution of the arguments. This fact and Proposition 1 leads to Corollary 1.

Corollary 1. *If the reward function is bounded in $[0, 1]$, then, for any MDP Φ ,*

$$\text{MBR}_{\Phi} \leq \sum_{t=1}^T \mathbb{E} \left[\sqrt{\frac{1}{2} D_{\text{KL}}(\mathbb{P}_{Y_t^*, S_t^* | \Theta} \parallel \mathbb{P}_{\hat{Y}_t, \hat{S}_t | \hat{H}^t})} \right].$$

Second, we can note that a bounded $[0, 1]$ function is 1-Lipschitz under the discrete metric (or Hamming distortion) $\rho((y, a), (y', a')) := \mathbb{1}_{(y, a) \neq (y', a')}$ where $\mathbb{1}$ is the indicator function. Using this fact, we can obtain Corollary 2 from Proposition 2.

Corollary 2. *If the reward function is bounded in $[0, 1]$, then, for any MDP Φ ,*

$$\text{MBR}_{\Phi} \leq \sum_{t=1}^T \mathbb{E} [\mathbb{W}(\mathbb{P}_{Y_t^*, S_t^* | \Theta}, \mathbb{P}_{\hat{Y}_t, \hat{S}_t | \hat{H}^t})].$$

Remark 3. *Corollary 2 provides a tighter bound than Corollary 1. Indeed, if the geometry is ignored (i.e., the discrete metric is considered), then for all $t = 1, \dots, T$,*

$$\begin{aligned} & \mathbb{E} [\mathbb{W}(\mathbb{P}_{Y_t^*, S_t^* | \Theta}, \mathbb{P}_{\hat{Y}_t, \hat{S}_t | \hat{H}^t})] \\ &= \mathbb{E} [\mathbb{TV}(\mathbb{P}_{Y_t^*, S_t^* | \Theta}, \mathbb{P}_{\hat{Y}_t, \hat{S}_t | \hat{H}^t})] \\ &\leq \mathbb{E} \left[\sqrt{\frac{1}{2} D_{\text{KL}}(\mathbb{P}_{Y_t^*, S_t^* | \Theta} \parallel \mathbb{P}_{\hat{Y}_t, \hat{S}_t | \hat{H}^t})} \right], \end{aligned}$$

where the equality follows from [6, Proof of Theorem 6.15] and inequality follows from Pinsker’s [18, Theorem 6.5] and Bretagnolle–Huber’s result [19, Proof of Lemma 2.1].

VI. UPPER BOUNDS FOR STATIC MDPs

In this section, we leverage the bound from Section V to obtain bounds on the minimum Bayesian regret for static Markov decision processes. We focus here on the case where the reward function is bounded in $[0, 1]$ and leave the sub-Gaussian and Lipschitz cases to the Appendix B, since they are analogous to the previous section. We first present upper bounds on the MBR for the multi-armed bandit problem. We then produce upper bounds to the online optimization with partial feedback problem, and show how they can recover from below the results from Russo and Van Roy [2].

Similarly to Section V, we can apply Lemma 2 to upper bound the MBR for static MDPs. In the case of a static MDP, the right-hand side of that bound can be written as

$$\sum_{t=1}^T \mathbb{E} \left[\mathbb{E} \left[r(Y_t, \psi_t^*(S, \Theta)) - r(Y_t, \psi_t^*(S, \hat{\Theta}_t)) \mid \Theta, \hat{\Theta}_t, \hat{H}^t \right] \right].$$

This rewriting of the bound is obtained the same way as (1): unrolling $R_{\pi}(\kappa_{\Theta})$ and $r_{\Pi}(\kappa_H, \kappa_{H|\Theta}, \psi^*)$, using the linearity of the expectation, the law of total expectation and the fact that the state S does not depend on the time $t = 1, \dots, T$.

In the case where the outcomes $\{Y_t\}_{t=1, \dots, T}$ do not depend on the state S , as in a MAB problem, it is possible to rewrite the actions taken by optimal policy $\psi_t^*(S, \Theta)$ as $\gamma^*(\Theta)$, where the function $\gamma^* : \mathcal{O} \rightarrow \mathcal{A}$ is such that for all $S \in \mathcal{S}$ and all $\Theta \in \mathcal{O}$, it holds that $\psi_t^*(S, \Theta) = \gamma^*(\Theta)$. In that case, it comes that the right-hand side term from Lemma 2 can be written as

$$\sum_{t=1}^T \mathbb{E} \left[\mathbb{E} [r(Y_t, A^*) - r(Y_t, \hat{A}_t)] \mid A^*, \hat{A}_t, \hat{H}^t \right]. \quad (2)$$

Remark 4. *Under this reformulation, the outcome in the first term inside the conditional expectation is distributed according to $\mathbb{P}_{Y_t | A^*, \hat{H}^t}$ and the second term is distributed according to $\mathbb{P}_{Y_t | \hat{H}^t}$. This happens since Y_t is independent of the sampled environment parameters $\hat{\Theta}_t$, and therefore independent of the sampled action \hat{A}_t when the history \hat{H}^t is known.*

A. Multi-armed bandit problem

In this subsection, we propose minimum Bayesian regret bounds for multi-armed bandit problems Π .

The tightest bound we obtain relates the MBR_{Π} to the Wasserstein distance between the conditional probability of the outcome given the optimal action and the history collected following a Thompson sampling policy, and the conditional probability of the outcome given only the history.

Proposition 3. *If the reward function is bounded in $[0, 1]$, then for any static MDP Π ,*

$$\text{MBR}_{\Pi} \leq \sum_{t=1}^T \mathbb{E} [\mathbb{W}(\mathbb{P}_{Y_t | A^*, \hat{H}^t}, \mathbb{P}_{Y_t | \hat{H}^t})].$$

Proof. The proof follows from applying Kantorovich–Rubinstein duality [6, Remark 6.5] to (2) using Remark 4 in the same way as [9], [15]. \square

Using the same arguments as in Remark 3, together with Jensen’s inequality, one can relax the bound from Proposition 3 and relate the MBR_{Π} to the conditional mutual information between the outcome Y_t and the optimal action A^* given the history \hat{H}^t . This is formalized in the following corollary.

Corollary 3. *If the reward function is bounded in $[0, 1]$, then for any static MDP Π ,*

$$\text{MBR}_{\Pi} \leq \sum_{t=1}^T \sqrt{\frac{1}{2} \mathbb{I}(Y_t; A^* | \hat{H}^t)}.$$

This conditional mutual information can be interpreted as the remaining ‘‘amount of surprise about the output Y_t ’’ after observing the history \hat{H}^t that is removed when the optimal action A^* is revealed.

B. Online optimization with partial feedback problem

In the special case of online optimization with partial feedback, the right-hand-side term from Lemma 2 in (2) can be formulated in a compact form using the preference function:

$$\sum_{t=1}^T \mathbb{E} \left[\mathbb{E} [r'(Y_{t,A^*})] - \mathbb{E} [r'(Y_{t,\hat{A}_t})] \mid A^*, \hat{A}_t, \hat{H}^t \right]. \quad (3)$$

Remark 5. *In this last rewriting, the outcome in the first term inside the conditional expectation is distributed according to $\mathbb{P}_{Y_{t,A^*} \mid A^*, \hat{H}^t}$ and the second term is distributed according to $\mathbb{P}_{Y_{t,\hat{A}_t} \mid \hat{H}^t}$. This holds since Y_t is independent of the sampled environment parameters $\hat{\Theta}_t$, and therefore independent of the sampled action \hat{A}_t when the history \hat{H}^t is known.*

As the terms in (3) are a difference of expectations of random objects which randomness comes from distributions on the same space, we can upper bound the minimum Bayesian regret using the Wasserstein distance in terms of such distributions following the techniques from [9], [15]. This is formalized in the following proposition.

Proposition 4. *If the reward function is bounded in $[0, 1]$, then for any online optimization problem with partial feedback Π ,*

$$\text{MBR}_{\Pi} \leq \sum_{t=1}^T \mathbb{E} [\mathbb{W}(\mathbb{P}_{Y_{t,A^*} \mid A^*, \hat{H}^t}, \mathbb{P}_{Y_{t,A^*} \mid \hat{H}^t})].$$

Proof. The proof follows from applying Kantorovich–Rubinstein duality [6, Remark 6.5] to (3) using Remark 5. \square

This bound can also be relaxed following a similar procedure as Remark 3 to relate the MBR_{Π} to the relative entropy between the distribution of the ‘‘per-action outcome’’ Y_{t,A^*} given the optimal action A^* and the history \hat{H}^t and given the history only.

Corollary 4. *If the reward function is bounded in $[0, 1]$, then for any online optimization problem with partial feedback Π ,*

$$\text{MBR}_{\Pi} \leq \sum_{t=1}^T \mathbb{E} \left[\sqrt{\frac{1}{2} D_{\text{KL}}(\mathbb{P}_{Y_{t,A^*} \mid A^*, \hat{H}^t} \parallel \mathbb{P}_{Y_{t,A^*} \mid \hat{H}^t})} \right].$$

As the above stated Proposition 4 is derived using Lemma 2, its bound naturally holds for the regret of the Thompson sampling algorithm, namely:

$$R_{\Pi}(\kappa_{\Theta}) - r_{\Pi}(\kappa_{\text{H}}, \kappa_{\Theta \mid \text{H}}, \psi^*) \leq \sum_{t=1}^T \mathbb{E} [\mathbb{W}(\mathbb{P}_{Y_{t,A^*} \mid A^*, \hat{H}^t}, \mathbb{P}_{Y_{t,A^*} \mid \hat{H}^t})]$$

We can further relax this bound to recover results from Russo and Van Roy [2]. More specifically, we can recover the

general bound combining [2, Propositions 1 and 3], and the specific bound combining [2, Propositions 1 and 4], for which it is assumed that the outcome Y_t is perfectly revealed upon observing $Y_{t,a}$ for any $a \in \mathcal{A}$. These claims are formalized in Corollary 5.

Corollary 5. *If the reward function is bounded in $[0, 1]$, then for any online optimization problem with partial feedback Π , we have the following inequality on the bound from Proposition 4:*

$$\sum_{t=1}^T \mathbb{E} [\mathbb{W}(\mathbb{P}_{Y_{t,A^*} \mid A^*, \hat{H}^t}, \mathbb{P}_{Y_{t,A^*} \mid \hat{H}^t})] \leq \sqrt{\frac{1}{2} |\mathcal{A}| \text{H}(A^*) T}.$$

Under the additional assumption that the outcome Y_t is perfectly revealed upon observing $Y_{t,a}$ for any $a \in \mathcal{A}$, one can obtain a tighter result:

$$\sum_{t=1}^T \mathbb{E} [\mathbb{W}(\mathbb{P}_{Y_{t,A^*} \mid A^*, \hat{H}^t}, \mathbb{P}_{Y_{t,A^*} \mid \hat{H}^t})] \leq \sqrt{\frac{1}{2} \text{H}(A^*) T}.$$

Intuition of the proof. For both results, the proof starts from using the same steps as Remark 3 to relax the bound from Proposition 4. Then, both of the proofs rely on the application of Cauchy-Schwartz’s and Jensen’s inequalities to obtain a bound using the sum of the conditional mutual information between the optimal action A^* and the ‘‘per-action outcome’’ Y_{t,A^*} given the history \hat{H}^t . One can then show that the entropy of the optimal action $\text{H}(A^*)$ upper bounds this sum of conditional mutual information $\text{I}(A^*; Y_{t,A^*} \mid \hat{H}^t)$ to obtain the desired results. The assumption that the outcome Y_t is perfectly revealed upon observing $Y_{t,a}$ for any $a \in \mathcal{A}$ averts an extra use of the Cauchy-Schwartz inequality and thus allows to avoid an explicit dependence on the number of actions through the multiplicative constant $\sqrt{|\mathcal{A}|}$. The full proof can be found in Appendix B-B. \square

VII. CONCLUSION

In this paper, building on the results from [1], we introduce a framework to study the Bayesian cumulative reward and the minimum Bayesian regret for reinforcement learning problems in the form of Markov decision process. The latter, is an algorithm-independent quantity and reflects the difficulty of the reinforcement learning problem. We prove a data processing inequality for the Bayesian cumulative reward and present upper bounds on the minimum Bayesian regret using the Wasserstein distance and the relative entropy. We leverage these results to the particular cases of the multi-armed bandit and the online optimization with partial feedback problems. For this last problem, our bound can be relaxed to recover from below the results presented in [2].

REFERENCES

- [1] A. Xu and M. Ragniksy, ‘‘Minimum excess risk in bayesian learning,’’ *arXiv preprint arXiv:2012.14868*, 2020.
- [2] D. Russo and B. Van Roy, ‘‘An information-theoretic analysis of thompson sampling,’’ *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2442–2471, 2016.

- [3] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," *Advances in neural information processing systems*, vol. 12, 1999.
- [4] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-dynamic programming*. Athena Scientific, 1996.
- [5] M. Ghavamzadeh, S. Mannor, J. Pineau, A. Tamar *et al.*, "Bayesian reinforcement learning: A survey," *Foundations and Trends® in Machine Learning*, vol. 8, no. 5-6, pp. 359–483, 2015.
- [6] C. Villani, *Optimal transport: old and new*. Springer, 2009, vol. 338.
- [7] O. Kallenberg, *Probabilistic symmetries and invariance principles*, 2nd ed. Springer, 2002.
- [8] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," *arXiv preprint arXiv:1705.07809*, 2017.
- [9] B. Rodríguez Gálvez, G. Bassi, R. Thobaben, and M. Skoglund, "Tighter expected generalization error bounds via wasserstein distance," *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [10] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3-4, pp. 285–294, 1933.
- [11] S. L. Scott, "A modern bayesian look at the multi-armed bandit," *Applied Stochastic Models in Business and Industry*, vol. 26, no. 6, pp. 639–658, 2010.
- [12] O. Chapelle and L. Li, "An empirical evaluation of thompson sampling," *Advances in neural information processing systems*, vol. 24, 2011.
- [13] B. C. May, N. Korda, A. Lee, and D. S. Leslie, "Optimistic bayesian sampling in contextual-bandit problems," *Journal of Machine Learning Research*, vol. 13, pp. 2069–2106, 2012.
- [14] I. Osband, D. Russo, and B. Van Roy, "(more) efficient reinforcement learning via posterior sampling," *Advances in Neural Information Processing Systems*, vol. 26, 2013.
- [15] H. Wang, M. Diaz, J. C. S. Santos Filho, and F. P. Calmon, "An information-theoretic view of generalization via wasserstein distance," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 577–581.
- [16] R. M. Gray, *Entropy and information theory*. Springer Science & Business Media, 2011.
- [17] W. Hoeffding, "Probability inequalities for sums of bounded random variables," in *The collected works of Wassily Hoeffding*. Springer, 1994, pp. 409–426.
- [18] Y. Polyanskiy and Y. Wu, "Lecture notes on Information Theory," *MIT (6.441), UIUC (ECE 563), Yale (STAT 664)*, 2017.
- [19] J. Bretagnolle and C. Huber, "Estimation des densités: risque minimax," in *Séminaire de Probabilités XII*. Springer, 1978, pp. 342–363.

APPENDIX A
LEMMAS AND EXTRA REMARKS

Lemma 1. *Let κ_U be a knowledge kernel associated with an MDP Φ and $\kappa_{V|U}$ a collection of processing kernels. Then, the cumulative Bayesian reward using the knowledge from U is at least as large as the processed Bayesian cumulative reward using the processed knowledge from V . More precisely,*

$$R_\Phi(\kappa_U) \geq R_\Phi(\kappa_U, \kappa_{V|U})$$

Proof. The proof starts by writing explicitly the inequality to be proven, namely

$$R_\Phi(\kappa_U) = \sup_{\{\varphi_t\}_{t=1}^T} \mathbb{E} \left[\sum_{t=1}^T r(Y_t, \varphi_t(S_t, U^t)) \right] \geq \sup_{\{\psi_t\}_{t=1}^T} \mathbb{E} \left[\sum_{t=1}^T r(Y'_t, \psi_t(S'_t, V_t)) \right] = R_\Phi(\kappa_U, \kappa_{V|U}),$$

where Y_t and S_t are the outcomes and states obtained from the actions derived from φ_t , the kernels that describe the MDP Φ , and κ_U ; where Y'_t , S'_t , and V_t are the outcomes, the states, and processed knowledge obtained from the actions derived from ψ_t and the kernels that describe the MDP Φ , and κ_U and $\kappa_{V|U}$.

Now, the proof follows by iteratively employing [7, Lemma 3.22] in a similar fashion to [1, Lemma 1]. To start, consider the kernel $\kappa_{V|U,1}$ from \mathcal{U} to \mathcal{V} , which are assumed to be Borel spaces. Then, there exists a measurable function $f_1 : \mathcal{U} \times [0, 1] \rightarrow \mathcal{V}$ such that if $\Xi \sim \text{Uniform}[0, 1]$ then $f(u, \Xi) \sim \kappa_{V|U,1}(\cdot, u)$ for all $u \in \mathcal{U}$ [7, Lemma 3.22]. Then,

$$R_\Phi(\kappa_U) = \sup_{\{\varphi_t\}_{t=1}^T} \mathbb{E} \left[r(Y'_1, \varphi_1(S'_1, U^1)) + \sum_{t=2}^T r(Y_t, \varphi_t(S_t, U^t)) \right] \quad (4)$$

$$\geq \sup_{\psi_1, \{\varphi_t\}_{t=2}^T} \sup_{\xi_1 \in [0, 1]} \mathbb{E} \left[r(Y'_1, \psi_1(S'_1, f_1(U^1, \xi))) + \sum_{t=2}^T r(\bar{Y}_t, \varphi_t(\bar{S}_t, \bar{U}^t)) \right] \quad (5)$$

$$\geq \sup_{\psi_1, \{\varphi_t\}_{t=2}^T} \mathbb{E} \left[r(Y'_1, \psi_1(S'_1, f_1(U^1, \Xi))) + \sum_{t=2}^T r(\tilde{Y}_t, \varphi_t(\tilde{S}_t, \tilde{U}^t)) \right] \quad (6)$$

$$= \sup_{\psi_1, \{\varphi_t\}_{t=2}^T} \mathbb{E} \left[r(Y'_1, \psi_1(S'_1, V_1)) + r(Y'_2, \varphi_2(S'_2, U'^2)) + \sum_{t=3}^T r(\tilde{Y}_t, \varphi_t(\tilde{S}_t, \tilde{U}^t)) \right], \quad (7)$$

where (4) follows since neither S_1 and Y_1 nor S'_1 and Y'_1 depend on the actions derived from φ or ψ , and therefore $S_1, S'_1 \sim \mathbb{P}_{S|\Theta}$ and $Y_1, Y'_1 \sim \mathbb{P}_{Y_1|S_1, \Theta}$. Then, (5) follows since the supremum of functions $\varphi_1 : \mathcal{S} \times \mathcal{U} \rightarrow \mathcal{A}$ is restricted to the functions $\sup_{\psi} \sup_{\xi} \psi(\cdot, f_1(\cdot, \xi))$, and where \bar{Y}_t, \bar{S}_t , and \bar{U} denote the outcomes, states, and knowledge obtained from the actions derived from ψ_1 and φ_t for $t \geq 2$, the kernels that describe the MDP Φ , and κ_U . After that, (6) holds since Ξ is independent of all random objects and $\sup_x f(x) \geq \mathbb{E}[f(X)]$. Here, \tilde{Y}_t, \tilde{S}_t , and \tilde{U}^t denote the outcomes and states obtained from the actions derived from ψ_1 and φ_t for $t \geq 2$, the kernels that describe the MDP Φ , and κ_U . Finally, in (7) it is used that $V_1 = f(U^1, \Xi)$ and therefore that $S'_2 = \tilde{S}_2$ and $Y'_2 = \tilde{Y}_2$. Similarly, U'^2 is the knowledge obtained from ψ_1 , the kernels that describe the MDP Φ , and κ_U and $\kappa_{V|U,1}$.

Repeating the technique above focusing on $r(Y'_2, \varphi_2(S'_2, U'^2))$ and using [7, Lemma 3.22] with the kernel $\kappa_{V|U,2}$ from \mathcal{U}^2 to \mathcal{V} one obtains that

$$R_\Phi(\kappa_U) \geq \sup_{\{\psi_t\}_{t=1}^2, \{\varphi_t\}_{t=3}^T} \mathbb{E} \left[\sum_{t=1}^2 r(Y'_t, \psi_t(S'_t, V_t)) + r(Y'_3, \varphi_3(S'_3, U'^3)) + \sum_{t=4}^T r(\tilde{Y}_t, \varphi_t(\tilde{S}_t, \tilde{U}^t)) \right], \quad (8)$$

where the notation is abused and \tilde{Y}_t and \tilde{S}_t denote the outcomes and states obtained from the actions derived from ψ_1, ψ_2 , and φ_t for $t \geq 3$, the kernels that describe the MDP Φ , and κ_U and $\kappa_{V|U,1}$ for $t < 3$. Also as before, U'^3 is the knowledge obtained from ψ_t for $t < 3$, the kernels that describe the MDP Φ , and κ_U and $\kappa_{V|U,t}$ for $t < 3$.

Finally, iterating this technique results in

$$R_\Phi(\kappa_U) \geq \sup_{\{\psi_t\}_{t=1}^T} \mathbb{E} \left[\sum_{t=1}^T r(Y'_t, \psi_t(S'_t, V_t)) \right] = R_\Phi(\kappa_U, \kappa_{V|U}) \quad (9)$$

and completes the proof. \square

Remark 6. *In the proof, ψ_1 may be different in both (7) and (8), since the supremum may vary. However, Y'_2, S'_2 , and U'^2 still represent the outcome, the state, and the knowledge obtained from the action derived from ψ_1 , the kernels that describe the MDP, and κ_U and $\kappa_{V|U,1}$. The same is true for all Y'_t, S'_t , and U'^t along the proof, ensuring that the random objects in (9) are distributed as in the definition of $R_\Phi(\kappa_U, \kappa_{V|U})$.*

APPENDIX B
UPPER BOUNDS FOR STATIC MDPs FOR SUB-GAUSSIAN AND LIPSCHITZ LOSSES

A. Multi-armed bandit problems

Proposition 5. *If for all $t = 1, \dots, T$, the random reward $r(Y, A^*)$ is σ_t^2 -sub-Gaussian under $\mathbb{P}_{Y_t|\hat{H}^t=\hat{h}^t}$ for all $\theta \in \mathcal{O}$ and all $\hat{h}^t \in \mathcal{H}^t$, then for any static MDP Π ,*

$$\text{MBR}_\Pi \leq \sum_{t=1}^T \sqrt{2\sigma_t^2 \mathbb{I}(Y_t; A^* | \hat{H}^t)}.$$

Proof. The proof starts from applying Donsker-Varadhan's inequality [16, Theorem 5.2.1] to (2) using Remark 4 in the same way as [2], [8]. The last inequality is obtained using Jensen's inequality and identifying the conditional mutual information between the outcome Y_t and the optimal action A^* given the history \hat{H}^t . Namely,

$$\text{MBR}_\Pi \leq \sum_{t=1}^T \mathbb{E} \left[\mathbb{E} [r(Y_t, A^*) - r(Y_t, \hat{A}_t) | A^*, \hat{A}_t, \hat{H}^t] \right] \leq \sum_{t=1}^T \mathbb{E} \left[\sqrt{2\sigma_t^2 D_{\text{KL}}(\mathbb{P}_{Y_t|A^*, \hat{H}^t} \| \mathbb{P}_{Y_t|\hat{H}^t})} \right] \leq \sum_{t=1}^T \sqrt{2\sigma_t^2 \mathbb{I}(Y_t; A^* | \hat{H}^t)}.$$

□

Proposition 6. *Suppose that \mathcal{Y} is a metric space with metric ρ . If the reward function $r : \mathcal{Y} \times \mathcal{A} \rightarrow \mathbb{R}$ is L -Lipschitz in \mathcal{Y} under the metric ρ , then*

$$\text{MBR}_\Pi \leq L \sum_{t=1}^T \mathbb{E} [\mathbb{W}(\mathbb{P}_{Y_t|A^*}, \mathbb{P}_{Y_t|\hat{H}^t})].$$

Proof. The proof follows from applying Kantorovich–Rubinstein duality [6, Remark 6.5] to (2) using Remark 4 analogously to [9], [15]. □

B. Online optimization with partial feedback problems

Proposition 7. *If for all $t = 1, \dots, T$, the random reward $r(Y, a^*)$ is σ_t^2 -sub-Gaussian under $\mathbb{P}_{Y|\hat{H}^t=\hat{h}^t}$ for all $a^* \in \mathcal{A}$ and all $\hat{h}^t \in \mathcal{H}^t$, then for any online optimization problem with partial feedback Π ,*

$$\text{MBR}_\Pi \leq \sum_{t=1}^T \mathbb{E} \left[\sqrt{2\sigma_t^2 D_{\text{KL}}(\mathbb{P}_{Y_t, A^* | \Theta} \| \mathbb{P}_{Y_t, A^* | \hat{H}^t})} \right]$$

Proof. The proof follows from applying Donsker-Varadhan's inequality [16, Theorem 5.2.1] to (3) using Remark 5 in a similar fashion to [2], [8]. □

Proposition 8. *Suppose that \mathcal{Y} is a metric space with metric ρ . If the reward function $r : \mathcal{Y} \times \mathcal{A} \rightarrow \mathbb{R}$ is L -Lipschitz in \mathcal{Y} under the metric ρ , then for any online optimization problem with partial feedback Π*

$$\text{MBR}_\Pi \leq L \sum_{t=1}^T \mathbb{E} [\mathbb{W}(\mathbb{P}_{Y_t, A^* | A^*}, \mathbb{P}_{Y_t, A^* | \hat{H}^t})].$$

Proof. The proof follows from applying Kantorovich–Rubinstein duality [6, Remark 6.5] to (3) using Remark 5 analogously to [9], [15]. □

Remark 7. *One can show that the entropy of the optimal action $\mathbb{H}(A^*)$ upper bounds the sum of conditional mutual information between the optimal action A^* and the “per-action outcome” Y_{t, A_t} given the history \hat{H}^t . This result is obtained in the same way as in [2] through the following chain of inequalities,*

$$\sum_{t=1}^T \mathbb{I}(A^*; Y_{t, A_t} | \hat{H}^t) \stackrel{(a)}{\leq} \sum_{t=1}^T \mathbb{I}(A^*; (Y_{t, A_t}, A_t) | \hat{H}^t) \stackrel{(b)}{=} \mathbb{I}(A^*; \{Y_{t, A_t}, A_t\}_{t=1}^T) \stackrel{(c)}{\leq} \mathbb{H}(A^*)$$

where (a) follows from [18, Theorem 2.3.5], equality (b) is given by the chain rule and (c) is obtained from [18, Theorem 2.4.4].

Corollary 5. *If the reward function is bounded in $[0, 1]$, then for any online optimization problem with partial feedback Π , we have the following inequality on the bound from Proposition 4:*

$$\sum_{t=1}^T \mathbb{E} [\mathbb{W}(\mathbb{P}_{Y_t, A^* | A^*, \hat{H}^t}, \mathbb{P}_{Y_t, A^* | \hat{H}^t})] \leq \sqrt{\frac{1}{2} |\mathcal{A}| \mathbb{H}(A^*) T}.$$

Under the additional assumption that the outcome Y_t is perfectly revealed upon observing $Y_{t,a}$ for any $a \in \mathcal{A}$, one can obtain a tighter result:

$$\sum_{t=1}^T \mathbb{E}[\mathbb{W}(\mathbb{P}_{Y_{t,A^*}|A^*,\hat{H}^t}, \mathbb{P}_{Y_{t,A^*}|\hat{H}^t})] \leq \sqrt{\frac{1}{2}\mathbf{H}(A^*)T}.$$

Proof. Under the assumption that the outcome Y_t is perfectly revealed upon observing $Y_{t,a}$ for any $a \in \mathcal{A}$, one can show the following chain of inequalities:

$$\sum_{t=1}^T \mathbb{E}[\mathbb{W}(\mathbb{P}_{Y_{t,A^*}|A^*,\hat{H}^t}, \mathbb{P}_{Y_{t,A^*}|\hat{H}^t})] \leq \sum_{t=1}^T \mathbb{E}\left[\sqrt{\frac{1}{2}D_{\text{KL}}(\mathbb{P}_{Y_{t,A_t}|A^*,\hat{H}^t} \parallel \mathbb{P}_{Y_{t,A_t}|\hat{H}^t})}\right] \quad (10)$$

$$\leq \sum_{t=1}^T \sqrt{\frac{1}{2}\mathbf{I}(A^*; Y_{t,A_t}|\hat{H}^t)} \quad (11)$$

$$\leq \sqrt{\frac{1}{2}T \sum_{t=1}^T \mathbf{I}(A^*; Y_{t,A_t}|\hat{H}^t)} \quad (12)$$

$$\leq \sqrt{\frac{1}{2}T\mathbf{H}(A^*)}$$

where (10) is obtained using the same arguments as in Remark 3 and, as Y_t is perfectly revealed from observing $Y_{t,a}$ for any $a \in \mathcal{A}$, we have that $D_{\text{KL}}(\mathbb{P}_{Y_{t,a}|A^*,\hat{H}^t} \parallel \mathbb{P}_{Y_{t,a}|\hat{H}^t}) = D_{\text{KL}}(\mathbb{P}_{Y_t|A^*,\hat{H}^t} \parallel \mathbb{P}_{Y_t|\hat{H}^t})$. Then Jensen's inequality leads to (11) and Cauchy-Schwartz inequality to (12). Finally, applying Remark 7 yields the desired result.

When no information structure is assumed among the outcome $Y_t \equiv \{Y_{t,a}\}$ for all $a \in \mathcal{A}$, inspired by the arguments used to prove [2, Proposition 3], one can show a looser inequality, through a chain of inequalities :

$$\sum_{t=1}^T \mathbb{E}[\mathbb{W}(\mathbb{P}_{Y_{t,A^*}|A^*,\hat{H}^t}, \mathbb{P}_{Y_{t,A^*}|\hat{H}^t})] \leq \sum_{t=1}^T \mathbb{E}\left[\sum_{a \in \mathcal{A}} \mathbb{P}(A^* = a|\hat{H}^t) \sqrt{\frac{1}{2}D_{\text{KL}}(\mathbb{P}_{Y_{t,a}|A^*=a,\hat{H}^t} \parallel \mathbb{P}_{Y_{t,a}|\hat{H}^t})}\right] \quad (13)$$

$$\leq \sum_{t=1}^T \mathbb{E}\left[\sqrt{\frac{1}{2}|\mathcal{A}| \sum_{a \in \mathcal{A}} \mathbb{P}(A^* = a|\hat{H}^t)^2 D_{\text{KL}}(\mathbb{P}_{Y_{t,a}|A^*=a,\hat{H}^t} \parallel \mathbb{P}_{Y_{t,a}|\hat{H}^t})}\right] \quad (14)$$

$$\leq \sum_{t=1}^T \mathbb{E}\left[\sqrt{\frac{1}{2}|\mathcal{A}| \sum_{a,b \in \mathcal{A}} \mathbb{P}(A^* = a|\hat{H}^t)\mathbb{P}(A^* = b|\hat{H}^t) D_{\text{KL}}(\mathbb{P}_{Y_{t,a}|A^*=b,\hat{H}^t} \parallel \mathbb{P}_{Y_{t,a}|\hat{H}^t})}\right] \quad (15)$$

$$\leq \sum_{t=1}^T \sqrt{\frac{1}{2}|\mathcal{A}|\mathbf{I}(A^*; Y_{t,A_t}|\hat{H}^t)} \quad (16)$$

$$\leq \sqrt{\frac{1}{2}|\mathcal{A}|T \sum_{t=1}^T \mathbf{I}(A^*; Y_{t,A_t}|\hat{H}^t)} \quad (17)$$

$$\leq \sqrt{\frac{1}{2}|\mathcal{A}|T\mathbf{H}(A^*)}$$

where (13) follows from Remark 3, (14) is obtained using Cauchy-Schwartz inequality, and (15) by adding the non-negative extra terms $\frac{1}{2}|\mathcal{A}| \sum_{a \in \mathcal{A}} \mathbb{P}(A^* = a|\hat{H}^t) \sum_{b \in \mathcal{A} \setminus a} \mathbb{P}(A^* = b|\hat{H}^t) D_{\text{KL}}(\mathbb{P}_{Y_{t,a}|A^*=b,\hat{H}^t} \parallel \mathbb{P}_{Y_{t,a}|\hat{H}^t})$ in the square root in (14). Then, (16) follows from using Jensen's inequality and identifying the conditional mutual information between the optimal action A^* and the ‘‘per-action outcome’’ Y_{t,A_t} given the history \hat{H}^t . Lastly, Cauchy-Schwartz inequality leads to (17) and Remark 7 gives the claimed result. \square