Deciding What to Model: Value-Equivalent Sampling for Reinforcement Learning

Dilip Arumugam
Department of Computer Science
Stanford University
dilip@cs.stanford.edu

Benjamin Van Roy
Department of Electrical Engineering
Department of Management Science & Engineering
Stanford University
bvr@stanford.edu

Abstract

The quintessential model-based reinforcement-learning agent iteratively refines its estimates or prior beliefs about the true underlying model of the environment. Recent empirical successes in model-based reinforcement learning with function approximation, however, eschew the true model in favor of a surrogate that, while ignoring various facets of the environment, still facilitates effective planning over behaviors. Recently formalized as the value equivalence principle, this algorithmic technique is perhaps unavoidable as real-world reinforcement learning demands consideration of a simple, computationally-bounded agent interacting with an overwhelmingly complex environment, whose underlying dynamics likely exceed the agent's capacity for representation. In this work, we consider the scenario where agent limitations may entirely preclude identifying an exactly value-equivalent model, immediately giving rise to a trade-off between identifying a model that is simple enough to learn while only incurring bounded sub-optimality. To address this problem, we introduce an algorithm that, using rate-distortion theory, iteratively computes an approximately-value-equivalent, lossy compression of the environment which an agent may feasibly target in lieu of the true model. We prove an information-theoretic, Bayesian regret bound for our algorithm that holds for any finite-horizon, episodic sequential decision-making problem. Crucially, our regret bound can be expressed in one of two possible forms, providing a performance guarantee for finding either the simplest model that achieves a desired sub-optimality gap or, alternatively, the best model given a limit on agent capacity.

1 Introduction

A central challenge of the reinforcement-learning problem [154, 87] is exploration, where a sequential decision-making agent must judiciously balance exploitation of knowledge accumulated thus far against the need to further acquire information for optimal long-term performance. Historically, provably-efficient reinforcement-learning algorithms [91, 35, 90, 20, 23, 151, 81, 121, 45, 120, 22, 46, 10, 85, 168, 56, 105] have often relied upon one of two possible mechanisms for addressing the exploration challenge in a principled manner: optimism in the face of uncertainty or posterior sampling. Briefly, methods in the former category begin with optimistically-biased value estimates for all state-action pairs; an agent acting greedily with respect to these estimates will be incentivized to visit all state-action pairs a sufficient number of times until this bias dissipates and the agent is left with an accurate estimate of the value function for deriving optimal behavior. In contrast, posterior-sampling methods primarily operate based on Thompson sampling [156, 141] whereby the agent begins with a prior belief over the Markov Decision Process (MDP) with which it is interacting and acts optimally with respect to a single sample drawn from these beliefs. The resulting experience sampled from the true environment allows the agent to derive a corresponding posterior distribution

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

and this Posterior Sampling for Reinforcement Learning (PSRL) [152] algorithm proceeds iteratively in this manner, eventually arriving at a posterior sharply concentrated around the true environment MDP. While both paradigms have laid down solid theoretical foundations for provably-efficient reinforcement learning, a line of work has demonstrated how posterior-sampling methods can be more favorable both in theory and in practice [121–123, 120, 114, 125, 63].

While existing analyses of reinforcement-learning algorithms have largely focused on providing guarantees for learning optimal solutions, real-world reinforcement learning demands consideration for a computationally-bounded agent interacting with an overwhelmingly complex environment [105]. A simplified view of this notion can be succinctly depicted in the multi-armed bandit setting [97, 38, 100]; as the number of arms increases, a Thompson sampling agent's relentless pursuit of the optimal arm will lead to large regret [138, 139]. On the other hand, one might simply settle for the first ε-optimal arm found, for some $\varepsilon > 0$, which may be identified in far fewer time periods. The goal of this work is to augment PSRL so as to accommodate these satisficing solutions in addition to optimal ones, paralleling existing work for satisficing in multi-armed bandit problems [140, 139, 15, 16]. To help elucidate the utility of satisficing solutions in the reinforcement-learning setting, we offer the following illustrative example:

Example 1 (A Multi-Resolution MDP)**.** *For a large but finite $N \in \mathbb{N}$, consider a sequence of MDPs, $\{\mathcal{M}_n\}_{n\in[N]}$, which all share a common action space $\mathcal{A}$ but vary in state space $\mathcal{S}_n$, reward function, and transition function. Moreover, for each $n \in [N]$, the rewards of the nth MDP are bounded in the interval $[0, \frac{1}{n}]$. An agent is confronted with the resulting product MDP, $\mathcal{M}$, defined on the state space $\mathcal{S}_1 \times \ldots \times \mathcal{S}_N$ with action space $\mathcal{A}$ and rewards summed across the N constituent reward functions. The transition function is defined such that each action $a \in \mathcal{A}$ is executed across all N MDPs simultaneously and the resulting individual transitions are combined into a transition of $\mathcal{M}$.*

Example 1 presents a simple scenario where, as $N \uparrow \infty$, a complex environment retains a wealth of information and yet, due to the scale of N and the boundedness of rewards for each constituent MDP $\mathcal{M}_n$, only a subset of that information is within the agent's reach or even necessary for producing reasonably competent behavior. Despite this fact, PSRL will persistently act to fully identify the transition and reward structure of all $\{\mathcal{M}_n\}_{n\in[N]}$, for any value of N. Without knowing which MDPs are more important *a priori* and even as data accumulates during learning, PSRL is unable to forego learning granular components of $\mathcal{M}$, eventually accumulating optimal reward at the cost of more time. Intuitively, however, one might anticipate that there exists a value $M \ll N$ such that learning the subsequence of MDPs $\{\mathcal{M}_n\}_{n\in[M]}$ in fewer time periods is sufficient for achieving a desired degree of sub-optimality, since the rewards of the remaining MDPs $\{\mathcal{M}_n\}_{n>M}$ make suitably negligible contributions to the overall rewards of $\mathcal{M}$. Alternatively, for a computationally-bounded decision maker, the agent's resource limitations ought to translate into a value $C \ll N$ such that $\{\mathcal{M}_n\}_{n\in[C]}$ is feasible and learning this subsequence is the best possible outcome under the agent capacity constraints. In this work, we introduce an algorithm that, in a purely data-driven and automated fashion, implicitly identifies such a value M or C to facilitate tractable, near-optimal learning in what may otherwise be an intractable problem. Following Arumugam and Van Roy [15], a key tool for defining a notion of satisficing in reinforcement learning will be rate-distortion theory [146, 25].

The paper proceeds as follows: we introduce our problem formulation in Section 3, present our generalization of PSRL in Section 4, and provide a complementary regret analysis in Section 5. Due to space constraints, technical proofs, an overview of related work, and discussion of our results in a broader context are relegated to the appendix.

2 Preliminaries

In this section, we provide brief background on information theory and details on our notation. All random variables are defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For any random variable $X : \Omega \to \mathcal{X}$ taking values on the measurable space $(\mathcal{X}, \mathbb{X})$, we use $\sigma(X) \triangleq \{X^{-1}(A) \mid A \in \mathbb{X}\} \subseteq \mathcal{F}$ to denote the σ-algebra generated by X. For any natural number $N \in \mathbb{N}$, we denote the index set as $[N] \triangleq \{1, 2, \ldots, N\}$. For any arbitrary set $\mathcal{X}$, $\Delta(\mathcal{X})$ denotes the set of all probability distributions with support on $\mathcal{X}$. For any two arbitrary sets $\mathcal{X}$ and $\mathcal{Y}$, we denote the class of all (measurable) functions mapping from $\mathcal{X}$ to $\mathcal{Y}$ as $\{\mathcal{X} \to \mathcal{Y}\} \triangleq \{f \mid f : \mathcal{X} \to \mathcal{Y}\}$. While our exposition throughout the paper will consistently refer to bits of information, it will be useful for the purposes of analysis that all logarithms be in base e.

2.1 Information Theory

Here we introduce various concepts in probability theory and information theory used throughout this paper. We encourage readers to consult [41, 71, 129, 59] for more background.

We define the mutual information between any two random variables X, Y through the Kullback-Leibler (KL) divergence:

$$\mathbb{I}(X;Y) = D_{\mathrm{KL}}(\mathbb{P}((X,Y) \in \cdot) \,||\, \mathbb{P}(X \in \cdot) \times \mathbb{P}(Y \in \cdot)) \qquad D_{\mathrm{KL}}(P \,||\, Q) = \begin{cases} \int \log\left(\frac{dP}{dQ}\right) dP & P \ll Q \\ +\infty & P \not\ll Q \end{cases},$$

where P and Q are both probability measures on the same measurable space and $\frac{dP}{dQ}$ denotes the Radon-Nikodym derivative of P with respect to Q. An analogous definition of conditional mutual information holds through the expected KL-divergence for any three random variables X, Y, Z:

$$\mathbb{I}(X;Y \mid Z) = \mathbb{E}\left[D_{\mathrm{KL}}(\mathbb{P}((X,Y) \in \cdot \mid Z) \,||\, \mathbb{P}(X \in \cdot \mid Z) \times \mathbb{P}(Y \in \cdot \mid Z))\right].$$

With these definitions in hand, we may define the entropy and conditional entropy for any two random variables X, Y as

$$\mathbb{H}(X) = \mathbb{I}(X;X) \qquad \mathbb{H}(Y \mid X) = \mathbb{H}(Y) - \mathbb{I}(X;Y).$$

This yields the following identities for mutual information and conditional mutual information for any three arbitrary random variables X, Y, and Z:

$$\mathbb{I}(X;Y) = \mathbb{H}(X) - \mathbb{H}(X \mid Y) = \mathbb{H}(Y) - \mathbb{H}(Y|X), \qquad \mathbb{I}(X;Y|Z) = \mathbb{H}(X|Z) - \mathbb{H}(X \mid Y,Z) = \mathbb{H}(Y|Z) - \mathbb{H}(Y|X,Z).$$

Through the chain rule of the KL-divergence and the fact that $D_{\mathrm{KL}}(P \,||\, P) = 0$ for any probability measure P, we obtain another equivalent definition of mutual information,

$$\mathbb{I}(X;Y) = \mathbb{E}\left[D_{\mathrm{KL}}(\mathbb{P}(Y \in \cdot \mid X) \,||\, \mathbb{P}(Y \in \cdot))\right],$$

as well as the chain rule of mutual information: $\mathbb{I}(X;Y_1,\ldots,Y_n) = \sum_{i=1}^{n} \mathbb{I}(X;Y_i \mid Y_1,\ldots,Y_{i-1})$. Finally, for any three random variables X, Y, and Z which form the Markov chain $X \to Y \to Z$, we have the following data-processing inequality: $\mathbb{I}(X;Z) \leq \mathbb{I}(X;Y)$.

3 Problem Formulation

We formulate a sequential decision-making problem as a finite-horizon, episodic Markov Decision Process (MDP) [24, 130] defined by $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \beta, H \rangle$. Here $\mathcal{S}$ denotes a set of states, $\mathcal{A}$ is a set of actions, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to [0,1]$ is a deterministic reward function providing evaluative feedback signals (in the unit interval) to the agent, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is a transition function prescribing distributions over next states, $\beta \in \Delta(\mathcal{S})$ is an initial state distribution, and $H \in \mathbb{N}$ is the maximum episode length or horizon.

As is standard in Bayesian reinforcement learning [70], neither the transition function nor the reward function are known to the agent and, consequently, both are treated as random variables. Since all other components of the MDP are thought of as known a priori, the randomness in the model $(\mathcal{R}, \mathcal{T})$ fully accounts for the randomness in $\mathcal{M}$, which is also a random variable. We denote by $\mathcal{M}^\star$ the true MDP with model $(\mathcal{R}^\star, \mathcal{T}^\star)$ that the agent interacts with and attempts to solve over the course of K episodes. Within each episode, the agent acts for exactly H steps beginning with an initial state $s_1 \sim \beta$. For each $h \in [H]$, the agent observes the current state $s_h \in \mathcal{S}$, selects action $a_h \sim \pi_h(\cdot \mid s_h) \in \mathcal{A}$, enjoys a reward $r_h = \mathcal{R}(s_h, a_h) \in [0,1]$, and transitions to the next state $s_{h+1} \sim \mathcal{T}(\cdot \mid s_h, a_h) \in \mathcal{S}$.

A stationary, stochastic policy for timestep $h \in [H]$, $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$, encodes a pattern of behavior mapping individual states to distributions over possible actions. Letting $\{\mathcal{S} \to \Delta(\mathcal{A})\}$ denote the class of all stationary, stochastic policies, a non-stationary policy $\pi = (\pi_1, \ldots, \pi_H) \in \{\mathcal{S} \to \Delta(\mathcal{A})\}^H$ is a collection of exactly H stationary, stochastic policies whose overall performance in any MDP $\mathcal{M}$ at timestep $h \in [H]$ when starting at state $s \in \mathcal{S}$ and taking action $a \in \mathcal{A}$ is assessed by its associated action-value function $Q^\pi_{\mathcal{M},h}(s,a) = \mathbb{E}\left[\sum_{h'=h}^{H} \mathcal{R}(s_{h'}, a_{h'}) \;\middle|\; s_h = s, a_h = a\right]$, where the

expectation integrates over randomness in the action selections and transition dynamics. Taking the corresponding value function as $V^{\pi}_{\mathcal{M},h}(s) = \mathbb{E}_{a\sim\pi_h(\cdot|s)}\left[Q^{\pi}_{\mathcal{M},h}(s,a)\right]$, we define the optimal policy $\pi^\star = (\pi^\star_1, \pi^\star_2, \ldots, \pi^\star_H)$ as achieving supremal value $V^\star_{\mathcal{M},h}(s) = \sup_{\pi\in\{\mathcal{S}\to\Delta(\mathcal{A})\}^H} V^{\pi}_{\mathcal{M},h}(s)$ for all $s \in \mathcal{S}$, $h \in [H]$. For brevity, we will write any value function $V \in \{S \to \mathbb{R}\}$ without its argument to implicitly integrate over randomness in the initial state: $V = \mathbb{E}_{s_1\sim\beta(\cdot)}[V(s_1)]$. We let $\tau_k = (s^{(k)}_1, a^{(k)}_1, r^{(k)}_1, \ldots, s^{(k)}_H, a^{(k)}_H, r^{(k)}_H, s^{(k)}_{H+1})$ be the random variable denoting the trajectory experienced by the agent in the kth episode. Meanwhile, $H_k = \{\tau_1, \tau_2, \ldots, \tau_{k-1}\} \in \mathcal{H}_k$ is the random variable representing the entire history of the agent's interaction within the environment at the start of the kth episode; the sequence of history random variables $\{H_k\}_{k\in[K]}$ induce and, by definition, are adapted to the filtration $\{\sigma(H_k)\}_{k\in[K]}$ of $(\Omega, \mathcal{F})$. We call attention to the fact that we have yet to make any further restrictions on the state-action space $\mathcal{S} \times \mathcal{A}$, such as finiteness; notably, the main results of this paper are not limited to tabular MDPs. As mentioned by Lattimore and Szepesvári [100] (also as Proposition 7.28 of Bertsekas and Shreve [28]), the Ionescu-Tulcea Theorem [80] ensures the existence of a probability space upon which τ_k and H_k are well-defined random variables for all episodes $k \in [K]$.

Abstractly, a reinforcement-learning algorithm is a sequence of non-stationary policies $(\pi^{(1)}, \ldots, \pi^{(K)})$ where for each episode $k \in [K]$, $\pi^{(k)} : \mathcal{H}_k \to \{\mathcal{S} \to \Delta(\mathcal{A})\}^H$ is a function of the current history H_k. We define the regret of a reinforcement-learning algorithm over K episodes as

$$\textsc{Regret}(K, \pi^{(1)}, \ldots, \pi^{(K)}, \mathcal{M}^\star) = \sum_{k=1}^{K} \Delta_k \qquad \Delta_k \triangleq V^\star_{\mathcal{M}^\star,1} - V^{\pi^{(k)}}_{\mathcal{M}^\star,1},$$

where Δ_k denotes the episodic regret or regret incurred during the kth episode with respect to the true MDP $\mathcal{M}^\star$. An agent's initial uncertainty in the (unknown) true MDP $\mathcal{M}^\star$ is reflected by an arbitrary prior distribution $\mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_1)$. Since the regret is a random variable due to our uncertainty in $\mathcal{M}^\star$, we integrate over this randomness to arrive at the Bayesian regret:

$$\textsc{BayesRegret}(K, \pi^{(1)}, \ldots, \pi^{(K)}) = \mathbb{E}\left[\textsc{Regret}(K, \pi^{(1)}, \ldots, \pi^{(K)}, \mathcal{M}^\star)\right].$$

Broadly speaking, our goal is to design a provably-efficient reinforcement-learning algorithm that incurs bounded Bayesian regret.

Throughout the paper, we will denote the entropy and conditional entropy conditioned upon a specific realization of an agent's history H_k, for some episode $k \in [K]$, as $\mathbb{H}_k(X) \triangleq \mathbb{H}(X \mid H_k = H_k)$ and $\mathbb{H}_k(X \mid Y) \triangleq \mathbb{H}_k(X \mid Y, H_k = H_k)$, for two arbitrary random variables X and Y. This notation will also apply analogously to the mutual information $\mathbb{I}_k(X;Y) \triangleq \mathbb{I}(X;Y \mid H_k = H_k) = \mathbb{H}_k(X) - \mathbb{H}_k(X \mid Y) = \mathbb{H}_k(Y) - \mathbb{H}_k(Y \mid X)$, as well as the conditional mutual information $\mathbb{I}_k(X;Y \mid Z) \triangleq \mathbb{I}(X;Y \mid H_k = H_k, Z)$, given an arbitrary third random variable, Z. Note that their dependence on the realization of random history H_k makes both $\mathbb{I}_k(X;Y)$ and $\mathbb{I}_k(X;Y \mid Z)$ random variables themselves. The traditional notion of conditional mutual information given the random variable H_k arises by integrating over this randomness:

$$\mathbb{E}\left[\mathbb{I}_k(X;Y)\right] = \mathbb{I}(X;Y \mid H_k) \qquad \mathbb{E}\left[\mathbb{I}_k(X;Y \mid Z)\right] = \mathbb{I}(X;Y \mid H_k, Z).$$

Additionally, we will also adopt a similar notation to express a conditional expectation given the random history H_k: $\mathbb{E}_k[X] \triangleq \mathbb{E}[X|H_k]$.

4 Satisficing Through Posterior Sampling

4.1 Rate-Distortion Theory

We begin with a brief, high-level overview of rate-distortion theory [146, 25] and encourage readers to consult [41] for more details and [26] for a survey of advances in rate-distortion theory towards solving the lossy source coding problem in information theory. A lossy compression problem consumes as input a fixed information source $\mathbb{P}(X \in \cdot)$ and a measurable distortion function $d : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$ which quantifies the loss of fidelity by using Z in place of X. Then, for any

$D \in \mathbb{R}_{\geq 0}$, the rate-distortion function quantifies the fundamental limit of lossy compression as

$$\mathcal{R}(D) = \inf_{Z \in \Lambda(D)} \mathbb{I}(X; Z) \qquad \Lambda(D) \triangleq \{Z : \Omega \to \mathcal{Z} \mid \mathbb{E}\left[d(X, Z)\right] \leq D\},$$

where the infimum is taken over all random variables Z that incur bounded expected distortion, $\mathbb{E}\left[d(X, Z)\right] \leq D$. Naturally, $\mathcal{R}(D)$ represents the minimum number of bits of information that must be retained from X in order to achieve this bounded expected loss of fidelity[1]. Throughout the paper, various facts of the rate-distortion function will be referenced as needed. For now, we simply note that, in keeping with the problem formulation of the previous section which does not automatically assume discrete random variables, the rate-distortion function is well-defined for abstract information source and channel output random variables [43].

Just as in past work that studies satisficing in multi-armed bandit problems [138, 139, 15], we will use rate-distortion theory to formalize and identify the best simplified MDP $\widetilde{\mathcal{M}}_k$ that the agent will attempt to learn over the course of each episode $k \in [K]$. The dependence on the particular episode comes from the fact that this lossy compression mechanism or channel will treat the agent's current beliefs over the true MDP $\mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_k)$ as the information source to be compressed.

4.2 The Value Equivalence Principle

As outlined in the previous section, the second input for a well-specified lossy-compression problem is a distortion function prescribing non-negative real values to realizations of the information source and channel output random variables $(\mathcal{M}^\star, \widetilde{\mathcal{M}})$ that quantify the loss of fidelity incurred by using $\widetilde{\mathcal{M}}$ in lieu of $\mathcal{M}^\star$. To define this function, we will leverage an approximate notion of value equivalence [72, 73]. For any arbitrary MDP $\mathcal{M}$ with model $(\mathcal{R}, \mathcal{T})$ and any stationary, stochastic policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, define the Bellman operator $\mathcal{B}^\pi_{\mathcal{M}} : \{\mathcal{S} \to \mathbb{R}\} \to \{\mathcal{S} \to \mathbb{R}\}$ as follows:

$$\mathcal{B}^\pi_{\mathcal{M}} V(s) \triangleq \mathbb{E}_{a \sim \pi(\cdot \mid s)} \left[\mathcal{R}(s, a) + \mathbb{E}_{s' \sim \mathcal{T}(\cdot \mid s, a)} \left[V(s')\right]\right], \qquad \forall s \in \mathcal{S}.$$

The Bellman operator is a foundational tool in dynamic-programming approaches to reinforcement learning [29] and gives rise to the classic Bellman equation: for any MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \beta, H \rangle$ and any non-stationary policy $\pi = (\pi_1, \ldots, \pi_H)$, the value functions induced by π satisfy $V^\pi_{\mathcal{M},h}(s) = \mathcal{B}^{\pi_h}_{\mathcal{M}} V^\pi_{\mathcal{M},h+1}(s)$, for all $h \in [H]$ and with $V^\pi_{\mathcal{M},H+1}(s) = 0$, $\forall s \in \mathcal{S}$. For any two MDPs $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \beta, H \rangle$ and $\widehat{\mathcal{M}} = \langle \mathcal{S}, \mathcal{A}, \widehat{\mathcal{R}}, \widehat{\mathcal{T}}, \beta, H \rangle$, Grimm et al. [72] define a notion of equivalence between them despite their differing models. For any policy class $\Pi \subseteq \{\mathcal{S} \to \Delta(\mathcal{A})\}$ and value function class $\mathcal{V} \subseteq \{\mathcal{S} \to \mathbb{R}\}$, $\mathcal{M}$ and $\widehat{\mathcal{M}}$ are value equivalent with respect to Π and $\mathcal{V}$ if and only if $\mathcal{B}^\pi_{\mathcal{M}} V = \mathcal{B}^\pi_{\widehat{\mathcal{M}}} V$, $\forall \pi \in \Pi, V \in \mathcal{V}$. In words, two different models are deemed value equivalent if they induce identical Bellman updates under any pair of policy and value function from $\Pi \times \mathcal{V}$. Grimm et al. [72] prove that when $\Pi = \{\mathcal{S} \to \Delta(\mathcal{A})\}$ and $\mathcal{V} = \{\mathcal{S} \to \mathbb{R}\}$, the set of all exactly value-equivalent models is a singleton set containing only the true model of the environment. The key insight behind value equivalence, however, is that practical model-based reinforcement-learning algorithms need not be concerned with modeling every granular detail of the underlying environment and may, in fact, stand to benefit by optimizing an alternative criterion besides the traditional maximum-likelihood objective [147, 66, 116, 17, 65, 57, 1, 44, 21, 144, 110, 113, 161]. Indeed, by restricting focus to decreasing subsets of policies $\Pi \subset \{\mathcal{S} \to \Delta(\mathcal{A})\}$ and value functions $\mathcal{V} \subset \{\mathcal{S} \to \mathbb{R}\}$, the space of exactly value-equivalent models is monotonically increasing.

For brevity, let $\mathfrak{R} \triangleq \{\mathcal{S} \times \mathcal{A} \to [0, 1]\}$ and $\mathfrak{T} \triangleq \{\mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})\}$ denote the classes of all reward functions and transition functions, respectively. Recall that, with $\langle \mathcal{S}, \mathcal{A}, \beta, H \rangle$ all known, the uncertainty in a random MDP $\mathcal{M}$ is entirely driven by its model $(\mathcal{R}, \mathcal{T})$ such that we may think of the support of $\mathcal{M}^\star$ as $\text{supp}(\mathcal{M}^\star) = \mathfrak{M} \triangleq \mathfrak{R} \times \mathfrak{T}$. We define a distortion function on pairs of MDPs $d : \mathfrak{M} \times \mathfrak{M} \to \mathbb{R}_{\geq 0}$ for any $\Pi \subseteq \{\mathcal{S} \to \Delta(\mathcal{A})\}$, $\mathcal{V} \subseteq \{\mathcal{S} \to \mathbb{R}\}$ as

$$d_{\Pi, \mathcal{V}}(\mathcal{M}, \widehat{\mathcal{M}}) = \sup_{\substack{\pi \in \Pi \\ V \in \mathcal{V}}} ||\mathcal{B}^\pi_{\mathcal{M}} V - \mathcal{B}^\pi_{\widehat{\mathcal{M}}} V||^2_\infty = \sup_{\substack{\pi \in \Pi \\ V \in \mathcal{V}}} \left(\sup_{s \in \mathcal{S}} |\mathcal{B}^\pi_{\mathcal{M}} V(s) - \mathcal{B}^\pi_{\widehat{\mathcal{M}}} V(s)|\right)^2.$$

In words, $d_{\Pi, \mathcal{V}}$ is the supremal squared Bellman error between MDPs $\mathcal{M}$ and $\widehat{\mathcal{M}}$ across all states $s \in \mathcal{S}$ with respect to the policy class Π and value function class $\mathcal{V}$.

[1] With a slight abuse of notation, we overload $\mathcal{R}$.

4.3 Value-Equivalent Sampling for Reinforcement Learning

By virtue of the previous two sections, we are now in a position to define the lossy compression problem that characterizes a MDP $\widetilde{\mathcal{M}}$ that the agent will aspire to learn in each episode $k \in [K]$ instead of the true MDP $\mathcal{M}^\star$. For any $\Pi \subseteq \{\mathcal{S} \to \Delta(\mathcal{A})\}$; $\mathcal{V} \subseteq \{\mathcal{S} \to \mathbb{R}\}$; $k \in [K]$; and $D \geq 0$, we define the rate-distortion function

$$\mathcal{R}_k^{\Pi,\mathcal{V}}(D) = \inf_{\widetilde{\mathcal{M}} \in \Lambda_k(D)} \mathbb{I}_k(\mathcal{M}^\star; \widetilde{\mathcal{M}}), \ \Lambda_k(D) \triangleq \{\widetilde{\mathcal{M}} : \Omega \to \mathfrak{M} \mid \mathbb{E}_k[d_{\Pi,\mathcal{V}}(\mathcal{M}^\star, \widetilde{\mathcal{M}})] \leq D\}. \quad (1)$$

This rate-distortion function characterizes the fundamental limit of MDP compression under our chosen distortion measure resulting in a channel that retains the minimum amount of information from the true MDP $\mathcal{M}^\star$ while yielding an approximately value-equivalent MDP in expectation. Observe that this distortion constraint is a notion of approximate value equivalence which collapses to the exact value equivalence of Grimm et al. [72] as $D \to 0$. Meanwhile, as $D \to \infty$, we accommodate a more aggressive compression of the true MDP $\mathcal{M}^\star$ resulting in less faithful Bellman updates.

Algorithm 1 Posterior Sampling for Reinforcement Learning (PSRL) [152]

Input: Prior $\mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_1)$
for $k \in [K]$ **do**
  Sample $M_k \sim \mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_k)$
  Get optimal policy $\pi^{(k)} = \pi^\star_{M_k}$
  Execute $\pi^{(k)}$ and get trajectory τ_k
  Update history $H_{k+1} = H_k \cup \tau_k$
  Induce posterior $\mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_{k+1})$
end for

Algorithm 2 Value-equivalent Sampling for Reinforcement Learning (VSRL)

Input: Prior $\mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_1)$, Threshold $D \in \mathbb{R}_{\geq 0}$, Distortion function $d_{\Pi,\mathcal{V}} : \mathfrak{M} \times \mathfrak{M} \to \mathbb{R}_{\geq 0}$
for $k \in [K]$ **do**
  Compute $\widetilde{\mathcal{M}}_k$ achieving $\mathcal{R}_k^{\Pi,\mathcal{V}}(D)$ limit (Equation 1)
  Sample MDP $M^\star \sim \mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_k)$
  Sample compression $M_k \sim \mathbb{P}(\widetilde{\mathcal{M}}_k \in \cdot \mid \mathcal{M}^\star = M^\star)$
  Compute optimal policy $\pi^{(k)} = \pi^\star_{M_k}$
  Execute $\pi^{(k)}$ and observe trajectory τ_k
  Update history $H_{k+1} = H_k \cup \tau_k$
  Induce posterior $\mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_{k+1})$
end for

A standard algorithm for our problem setting is widely known as Posterior Sampling for Reinforcement Learning (PSRL) [152, 120], which we present as Algorithm 1, while our Value-equivalent Sampling for Reinforcement Learning (VSRL) is given as Algorithm 2. The key distinction between them is that, at each episode $k \in [K]$, the latter takes the posterior sample $M^\star \sim \mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_k)$ and passes it through the channel that achieves the rate-distortion limit (Equation 1) at this episode to get the M_k whose optimal policy is executed in the environment.

The core impetus for this work is to recognize that, for complex environments, pursuit of the exact MDP $\mathcal{M}^\star$ (as in PSRL) may be an entirely infeasible goal. Consider a MDP that represents control of a real-world, physical system; learning a transition function of the associated environment, at some level, demands that the agent internalize the laws of physics and motion with near-perfect accuracy. More formally, identifying $\mathcal{M}^\star$ demands the agent obtain exactly $\mathbb{H}_1(\mathcal{M}^\star)$ bits of information from the environment which, under an uninformative prior, may either be prohibitively large by far exceeding the agent's capacity constraints or be simply impractical under time and resource constraints.

As a remedy for this problem, we embrace the idea of being "sufficiently satisfying" or *satisficing* [148, 140, 138, 139, 15, 16]; as succinctly stated by Herbert A. Simon during his 1978 Nobel Memorial Lecture, "decision makers can satisfice either by finding optimum solutions for a simplified world, or by finding satisfactory solutions for a more realistic world." Rather than spend an inordinate amount of time trying to recover an optimum solution to the true environment, we will instead design an algorithm that pursues optimum solutions for a sequence of simplified environments. In the next section, our analysis demonstrates that finding such optimum solutions for simplified worlds ultimately acts as a mechanism for achieving a satisfactory solution for the realistic, complex world. Naturally, the loss of fidelity between the simplified and true environments translates into a fixed amount of regret that an agent designer consciously and willingly accepts for two reasons: (1) they expect a reduction in the amount of time, data, and bits of information needed to identify the simplified environment and (2) in tasks where the environment encodes irrelevant information and

exact knowledge is not needed to achieve optimal behavior [66, 72, 73, 161], this worst-case error term may be negligible anyways while still maintaining greater efficiency than traditional PSRL.

Recalling Example 1 that revolves around a particular sequence of MDPs, $\{\mathcal{M}_n\}_{n\in[N]}$, we note that as the distortion threshold D increases, the significance of MDPs in the sequence indexed by larger values of $n \in [N]$ rapidly diminishes. As $D \uparrow \infty$, the lossy compression $\widetilde{\mathcal{M}}_k$ needn't convey information about any of the MDPs in $\{\mathcal{M}_n\}_{n\in[N]}$. Conversely, at $D = 0$, a VSRL agent must necessarily obtain enough information about the entire sequence so as to facilitate planning over Π and $\mathcal{V}$. In between, however, the agent need only concern itself with a particular subsequence of $\{\mathcal{M}_n\}_{n\in[N]}$ while the remaining MDPs can be ignored due to their negligible contribution to overall value and, therefore, expected distortion under $d_{\Pi,\mathcal{V}}$.

5 Regret Analysis

In this section, we offer an information-theoretic analysis of VSRL (Algorithm 2) before refining our regret bounds to the tabular setting. We conclude by highlighting how our performance guarantees can be expressed via a notion of agent capacity that is considerate of real-world reinforcement learning.

5.1 An Information-Theoretic Bayesian Regret Bound

To establish a Bayesian regret bound for VSRL we first require a regret decomposition that acknowledges the agent's new objective of identifying an approximately value-equivalent MDP in each episode, $\widetilde{\mathcal{M}}_k$, rather than the true MDP $\mathcal{M}^\star$. Crucially, this regret decomposition leverages the precise form of our distortion function $d_{\Pi,\mathcal{V}}(\mathcal{M}^\star, \widetilde{\mathcal{M}}_k)$.

Theorem 1. *Take any $\Pi \supseteq \{\mathcal{S} \to \mathcal{A}\}$, any $\mathcal{V} \supseteq \{V^\pi \mid \pi \in \Pi^H\}$, and fix any $D \geq 0$. For each episode $k \in [K]$, let $\widetilde{\mathcal{M}}_k$ be any MDP that achieves the rate-distortion limit of $\mathcal{R}_k^{\Pi,\mathcal{V}}(D)$ with information source $\mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_k)$ and distortion function $d_{\Pi,\mathcal{V}}$. Then,*

$$\textsc{BayesRegret}(K, \pi^{(1)}, \ldots, \pi^{(K)}) \leq \mathbb{E}\left[\sum_{k=1}^{K} \mathbb{E}_k\left[V^\star_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1}\right]\right] + 2KH\sqrt{D}.$$

Theorem 1 shows how the Bayesian regret incurred by VSRL can be separated into an error term the agent must pay for learning a simplified MDP $\widetilde{\mathcal{M}}_k$, rather than $\mathcal{M}^\star$, and the Bayesian regret incurred while trying to learn $\widetilde{\mathcal{M}}_k$. This first term mirrors the satisficing regret of Russo and Van Roy [138, 139] for multi-armed bandits where the performance of the agent in the kth episode is being measured with respect to a compressed MDP $\widetilde{\mathcal{M}}_k$, rather than the true MDP $\mathcal{M}^\star$. While further discussion on the choices of Π and $\mathcal{V}$ is provided later in this section, we simply note that the conditions placed upon them in Theorem 1 are an artifact of VSRL only executing optimal policies in each time period $h \in [H]$ which, under the assumptions of our problem formulation, are deterministic.

The remainder of this section is devoted to an analysis for establishing an information-theoretic bound on the satisficing regret term of Theorem 1. A central tool of our analysis will be the information ratio [136, 137] at the kth episode:

$$\Gamma_k \triangleq \frac{\mathbb{E}_k\left[V^\star_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1}\right]^2}{\mathbb{I}_k(\widetilde{\mathcal{M}}_k; \tau_k, M_k)} \qquad \forall k \in [K].$$

In words, the information ratio is the ratio between squared expected regret in the kth episode with respect to $\widetilde{\mathcal{M}}_k$ and the information gained about $\widetilde{\mathcal{M}}_k$ in the kth episode by sampling MDP M_k and observing trajectory τ_k, given the current history H_k. Numerous prior works have leveraged similar or generalized types of information ratios for analyzing multi-armed bandit problems [135–139, 54, 99, 169, 36, 15, 98] as well as reinforcement-learning problems [104]; in comparison to the latter, we simply note that our analysis bears stronger resemblance to those in multi-armed bandits by not constructing confidence sets over MDPs [121, 120, 104], avoiding a restricted focus to tabular problems. That said, our results are contingent upon the existence of a uniform upper bound to the information ratios across all episodes, a non-trivial result [78] that we leave to future work.

Through our information-ratio analysis, we obtain the following information-theoretic bound on satisficing Bayesian regret:

Theorem 2. *If $\Gamma_k \leq \overline{\Gamma}$, for all $k \in [K]$, then* $\mathbb{E}\left[\sum_{k=1}^{K} \mathbb{E}_k\left[V^{\star}_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1}\right]\right] \leq \sqrt{\overline{\Gamma} K \mathcal{R}_1^{\Pi,\mathcal{V}}(D)}$.

An immediate consequence of the preceding theorems is the following corollary which establishes our main result, an information-theoretic Bayesian regret bound for VSRL. We omit the proof as it follows directly from applying Theorems 1 and 2 in sequence.

Corollary 1. *Take any $\Pi \supseteq \{\mathcal{S} \to \mathcal{A}\}$, any $\mathcal{V} \supseteq \{V^{\pi} \mid \pi \in \Pi^H\}$, and fix any $D > 0$. For any prior distribution $\mathbb{P}(\mathcal{M}^{\star} \in \cdot \mid H_1)$, if $\Gamma_k \leq \overline{\Gamma}$ for all $k \in [K]$, then VSRL (Algorithm 2) has* $\textsc{BayesRegret}(K, \pi^{(1)}, \ldots, \pi^{(K)}) \leq \sqrt{\overline{\Gamma} K \mathcal{R}_1^{\Pi,\mathcal{V}}(D)} + 2KH\sqrt{D}$.

Once again we recall that, since the rate-distortion function is well-defined for arbitrary source and channel output random variables defined on abstract alphabets [43], the Bayesian regret bound of Corollary 1 holds for any finite-horizon, episodic MDP, extending beyond past analyses of PSRL constrained only to tabular MDPs. We defer a discussion of practical considerations for implementing VSRL to the appendix.

At this point, we call attention to the parameterization of our lossy compression problem by a particular policy class Π and value function class $\mathcal{V}$, whose dependence we inherit from the value equivalence principle [72]. The next result clarifies how the performance of VSRL is affected by fluctuations in these classes via a dominance relationship [150] between the induced distortion functions.

Lemma 1. *For any two Π, Π' and any $\mathcal{V}, \mathcal{V}'$ such that $\Pi' \subseteq \Pi \subseteq \{\mathcal{S} \to \Delta(\mathcal{A})\}$ and $\mathcal{V}' \subseteq \mathcal{V} \subseteq \{\mathcal{S} \to \mathbb{R}\}$, we have $\mathcal{R}_k^{\Pi,\mathcal{V}}(D) \geq \mathcal{R}_k^{\Pi',\mathcal{V}'}(D)$, $\forall k \in [K], D > 0$.*

Property 3 of Grimm et al. [72] highlights how the set of value-equivalent MDPs grows as the policy and value function classes shrink. Lemma 1 provides an intuitive, information-theoretic counterpart to their result where, as the sets of policies and value functions over which models will be assessed diminish, an agent may naturally compress more aggressively and throw away larger quantities of bits from each source distribution over the true MDP $\mathcal{M}^{\star}$.

Since a compressed MDP $\widetilde{\mathcal{M}}_k$ that achieves the rate-distortion limit has *expected* distortion bounded by D, one may wonder how the probability of not recovering an approximately-value-equivalent MDP scales as $D \uparrow \infty$. To that end, we conclude this section with a final result that brings clarity to this via a generalization [60] of Fano's inequality [64]. We leave investigation of other generalizations of Fano's inequality that might yield similarly interesting results to future work [160, 8].

Lemma 2. *Take any $\Pi \subseteq \{\mathcal{S} \to \Delta(\mathcal{A})\}$ and $\mathcal{V} \subseteq \{\mathcal{S} \to \mathbb{R}\}$. For any $D \geq 0$ and any $k \in [K]$, define $\delta = \sup_{\widehat{M} \in \mathfrak{M}} \mathbb{P}(d_{\Pi,\mathcal{V}}(\mathcal{M}^{\star}, \widehat{M}) \leq D \mid H_k)$. Then,*

$$\sup_{\widetilde{\mathcal{M}} \in \Lambda_k(D)} \mathbb{P}(d_{\Pi,\mathcal{V}}(\mathcal{M}^{\star}, \widetilde{\mathcal{M}}) > D \mid H_k) \geq 1 - \frac{\mathcal{R}_k^{\Pi,\mathcal{V}}(D) + \log(2)}{\log\left(\frac{1}{\delta}\right)}.$$

For any episode $k \in [K]$, the left-hand side of the inequality in Lemma 2 denotes the worst-case error probability of sampling a compressed MDP $\widetilde{\mathcal{M}}$ that is not approximately-value-equivalent to $\mathcal{M}^{\star}$. The right-hand side conveys that, in order to avoid such an error with reasonable probability, one requires a setting of $D < \infty$ such that $\mathcal{R}_k^{\Pi,\mathcal{V}}(D) \approx \log\left(\frac{1}{\delta}\right)$.

5.2 Specializing to Tabular MDPs

While the preceding subsection constitutes the main contribution of this paper, the presence of information-theoretic terms makes it difficult to compare our guarantees to those obtained in prior work, which typically focuses on the tabular setting. To help remedy this, we offer the following theorem which restricts focus to the case where the agent pursues an exactly value-equivalent model of the tabular environment. Notably, the results of this section still retain a dependence on a uniform upper bound to the information ratio whose exact form is a result left to future work.

Theorem 3. *Take any $\Pi \supseteq \{\mathcal{S} \to \mathcal{A}\}$, any $\mathcal{V} \supseteq \{V^{\pi} \mid \pi \in \Pi^H\}$, and let $D = 0$. For any prior distribution $\mathbb{P}(\mathcal{M}^{\star} \in \cdot \mid H_1)$ over tabular MDPs, if $\Gamma_k \leq \overline{\Gamma}$ for all $k \in [K]$, then VSRL (Algorithm 2) has* $\textsc{BayesRegret}(K, \pi^{(1)}, \ldots, \pi^{(K)}) \leq \mathcal{O}\left(|\mathcal{S}|\sqrt{\overline{\Gamma}|\mathcal{A}|K}\right)$.

An immediate observation is that the Bayesian regret bound of Theorem 3 matches the dependence on the number of states, $|\mathcal{S}|$, obtained in the first (weaker) guarantee established for PSRL by Osband et al. [121]; we suspect that this guarantee for VSRL is unimprovable without further distributional assumptions [120, 119]. As an alternative, we contemplate how a change in the distortion measure used by VSRL might incur an improved regret bound when specialized to the tabular setting.

Specifically, notice that the only piece of the VSRL analysis tethered to the particular form of the distortion function $d_{\Pi,\mathcal{V}}(\mathcal{M}, \widehat{\mathcal{M}})$ is Theorem 1, while all other components remain agnostic to the precise criterion for assessing the loss of fidelity between original and compressed MDPs. Consequently, there is potential for a modified distortion function to offer an improved regret analysis relative to Theorem 3. Rather than concerning ourselves with planning over multiple behaviors, we consider a distortion function based solely on the optimal action-value functions:

$$d_{Q^\star}(\mathcal{M}, \widehat{\mathcal{M}}) = \sup_{h \in [H]} ||Q^\star_{\mathcal{M},h} - Q^\star_{\widehat{\mathcal{M}},h}||^2_\infty = \sup_{h \in [H]} \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} |Q^\star_{\mathcal{M},h}(s,a) - Q^\star_{\widehat{\mathcal{M}},h}(s,a)|^2.$$

We use $\mathcal{R}_k^{Q^\star}(D)$ to denote the rate-distortion function under this new measure of distortion, $d_{Q^\star}(\mathcal{M}, \widehat{\mathcal{M}})$. In order for this new distortion function to be compatible with VSRL, we require an analogue to the regret decomposition of Theorem 1.

Theorem 4. *Fix any $D \geq 0$ and, for each episode $k \in [K]$, let $\widetilde{\mathcal{M}}_k$ be any MDP that achieves the rate-distortion limit of $\mathcal{R}_k^{Q^\star}(D)$ with information source $\mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_k)$ and distortion function $d_{Q^\star}$. Then,* $\textsc{BayesRegret}(K, \pi^{(1)}, \ldots, \pi^{(K)}) \leq \mathbb{E}\left[\sum_{k=1}^{K} \mathbb{E}_k\left[V^\star_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1}\right]\right] + 2K(H+1)\sqrt{D}$.

With this regret decomposition in hand, we may recover the analogue to Corollary 1, whose proof is immediate and, therefore, omitted.

Corollary 2. *Fix any $D > 0$. For any prior distribution $\mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_1)$, if $\Gamma_k \leq \overline{\Gamma}$ for all $k \in [K]$, then VSRL (Algorithm 2) with distortion function $d_{Q^\star}$ has* $\textsc{BayesRegret}(K, \pi^{(1)}, \ldots, \pi^{(K)}) \leq \sqrt{\overline{\Gamma} K \mathcal{R}_1^{Q^\star}(D)} + 2K(H+1)\sqrt{D}$.

As illustrated by the following lemma, the significance of this change in distortion measure from $d_{\Pi,\mathcal{V}}$ to $d_{Q^\star}$ is that the optimal action-value functions may now act as an information bottleneck [158] between the original MDP $\mathcal{M}^\star$ and compressed MDP $\widetilde{\mathcal{M}}_k$.

Lemma 3. *For each episode $k \in [K]$ and for $D = 0$, let $\widetilde{\mathcal{M}}_k$ be any MDP that achieves the rate-distortion limit of $\mathcal{R}_k^{Q^\star}(D)$ with information source $\mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_k)$ and distortion function $d_{Q^\star}$. Then, we have the Markov chain $\mathcal{M}^\star \to Q^\star_{\mathcal{M}^\star} \to \widetilde{\mathcal{M}}_k$, where $Q^\star_{\mathcal{M}^\star} = \{Q^\star_{\mathcal{M}^\star,h}\}_{h \in [H]}$ is the collection of random variables denoting the optimal action-value functions of $\mathcal{M}^\star$.*

Lemma 3, through the data-processing inequality, immediately leads us to an analogue of Theorem 3 that matches the dependence on $|\mathcal{S}|$ in the best known Bayesian regret bound for PSRL [120].

Theorem 5. *For $D = 0$ and any prior distribution $\mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_1)$ over tabular MDPs, if $\Gamma_k \leq \overline{\Gamma}$ for all $k \in [K]$, then VSRL with distortion function $d_{Q^\star}$ has* $\textsc{BayesRegret}(K, \pi^{(1)}, \ldots, \pi^{(K)}) \leq \widetilde{\mathcal{O}}\left(\sqrt{\overline{\Gamma}|\mathcal{S}||\mathcal{A}|KH}\right)$.

Ultimately, Theorem 5 confirms that while there is great flexibility in the original definition of value equivalence to support planning across multiple policies and value functions, focusing on optimal value functions gives rise to more efficient learning. Moreover, comparing the result with the PSRL regret bound of Osband and Van Roy [120] for tabular MDPs, this suggests an achievable uniform upper bound to the information ratio as $\overline{\Gamma} \lesssim H^2$, where the $\lesssim$ accounts for numerical constants and logarithmic factors.

5.3 Capacity-Sensitive Performance Guarantees

We recognize that the information-theoretic regret bounds of the previous two sections, like many other guarantees for provably-efficient reinforcement learning before them, implicitly and unrealistically

assume that an agent is of unbounded capacity and may pursue any approximately-value-equivalent model under a given distortion threshold D. In the context of real-world reinforcement learning [62, 105], however, fundamental limits on computational resources and time leave an agent designer with a bounded agent to be deployed within an overwhelmingly complex environment. As such, this designer may seldom be in a position to dictate an ideal or desired sub-optimality threshold D, but rather must make do with a known constraint on agent capacity; guarantees on sample-efficient reinforcement learning cognizant of such a fundamental constraint are nascent.

While there are numerous possibilities for how one might choose to formally characterize agent capacity, we here adopt a fundamental perspective that learning is the process of acquiring information and so take this capacity to imply the existence of a non-negative real value $R \in \mathbb{R}_{>0}$ such that the agent may only acquire and retain exactly R bits of information. To help contextualize this notion of agent capacity, we introduce the distortion-rate function [146, 25, 41] which quantifies the fundamental limit of expected distortion under an information constraint:

$$\mathcal{D}_k^{Q^\star}(R) = \inf_{\widetilde{\mathcal{M}} \in \Upsilon_k(R)} \mathbb{E}_k \left[d_{Q^\star}(\mathcal{M}^\star, \widetilde{\mathcal{M}}) \right] \qquad \mathcal{D}_k^{Q^\star}(R) = \inf_{\widetilde{\mathcal{M}} \in \Upsilon_k(R)} \mathbb{E}_k \left[d_{Q^\star}(\mathcal{M}^\star, \widetilde{\mathcal{M}}) \right], \quad (2)$$

where the infimum is taken over all channels with bounded rate, $\Upsilon_k(R) \triangleq \{\widetilde{\mathcal{M}} : \Omega \to \mathfrak{M} \mid \mathbb{I}_k(\mathcal{M}^\star; \widetilde{\mathcal{M}}) \leq R\}$. In words, given the agent's current beliefs over the true MDP $\mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_k)$, the infimum of the distortion-rate function is taken over all potential lossy compressions of the environment that fall within the agent's capacity constraint of R bits and identifies the one that preserves the most useful information, as measured by the distortion function. Conveniently, the rate-distortion function and distortion-rate function are inverses of one another [41] ($\mathcal{R}(\mathcal{D}(R)) = R$) such that we recover the following two capacity-sensitive regret bounds directly from Corollaries 1 and 2 by simply taking the input distortion threshold of VSRL equal to the associated distortion-rate function in the first episode ($D = \mathcal{D}_1^{\Pi,\mathcal{V}}(R)$ and $D = \mathcal{D}_1^{Q^\star}(R)$, respectively).

Corollary 3. *Take any $\Pi \supseteq \{\mathcal{S} \to \mathcal{A}\}$, any $\mathcal{V} \supseteq \{V^\pi \mid \pi \in \Pi^H\}$, and let $R > 0$ be the agent capacity. For any prior distribution $\mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_1)$, if $\Gamma_k \leq \overline{\Gamma}$ for all $k \in [K]$, then VSRL (Algorithm 2) with distortion function $d_{\Pi,\mathcal{V}}$ has* $\textsc{BayesRegret}(K, \pi^{(1)}, \ldots, \pi^{(K)}) \leq \sqrt{\overline{\Gamma} K R} + 2KH\sqrt{\mathcal{D}_1^{\Pi,\mathcal{V}}(R)}$.

Corollary 4. *Let $R > 0$ be the agent capacity. For any prior distribution $\mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_1)$, if $\Gamma_k \leq \overline{\Gamma}$ for all $k \in [K]$, then VSRL (Algorithm 2) with distortion function $d_{Q^\star}$ has* $\textsc{BayesRegret}(K, \pi^{(1)}, \ldots, \pi^{(K)}) \leq \sqrt{\overline{\Gamma} K R} + 2K(H+1)\sqrt{\mathcal{D}_1^{Q^\star}(R)}$.

Turning back to Example 1, note how an agent with significantly limited capacity cannot possibly hope to capture all the granularity contained in the entire MDP sequence $\{\mathcal{M}_n\}_{n \in [N]}$, for large values of N. For a capacity of exactly R bits, Corollaries 3 and 4 immediately translate this fundamental limit into a corresponding performance guarantee, allowing the agent to identify a subsequence $\{\mathcal{M}_n\}_{n \in [C]}$ for some $C \ll N$ which only requires gathering R bits of information from the environment.

6 Conclusion

In this paper, we began with a finite-horizon, episodic MDP and considered the ramifications of a real-world reinforcement-learning scenario wherein the relative complexity of the environment is so immense that an agent may find itself incapable of perfectly recovering optimal behavior. An immediate consequence of this reality is the need to strike an appropriate balance between what is performant and what is achievable. We introduced the VSRL algorithm for incrementally synthesizing *simple* and *useful* approximations of the environment from which an agent might still recover near-optimal behaviors. Recognizing the information-theoretic nature of this lossy MDP compression, we provided an analysis of VSRL whose performance guarantees, by virtue of rate-distortion theory, are twofold. The first set of guarantees ensure VSRL recovers the simplest compression of the environment which still incurs bounded sub-optimality, as specified by the agent designer. Alternatively, the second set of guarantees maintain that VSRL finds the best compression of the environment subject to constraints on agent capacity. Through our general problem formulation and information-theoretic analysis, both regret bounds hold for any finite-horizon, episodic MDP, regardless of whether or not the state-action space is finite. That said, the question of how to practically instantiate VSRL for high-dimensional settings is an open problem left to future work.

Acknowledgements

The authors gratefully acknowledge Christopher Grimm for initial discussions that provided an impetus for this work. The authors also thank Adithya Devraj, Shi Dong, John Duchi, Hong Jun Jeon, Saurabh Kumar, and Xiuyuan (Lucy) Lu for insightful comments on various components of the paper. Financial support from Army Research Office (ARO) grant W911NF2010055 is gratefully acknowledged.

References

[1] Romina Abachi, Mohammad Ghavamzadeh, and Amir-massoud Farahmand. Policy-aware model learning for policy gradient methods. *arXiv preprint arXiv:2003.00030*, 2020. 5

[2] Yasin Abbasi-Yadkori and Csaba Szepesvari. Bayesian optimal control of smoothly parameterized systems: The lazy posterior sampling algorithm. *arXiv preprint arXiv:1406.3926*, 2014. 22

[3] David Abel. *A Theory of Abstraction in Reinforcement Learning*. PhD thesis, Brown University, 2020. 23

[4] David Abel, David Hershkowitz, and Michael Littman. Near optimal behavior via approximate state abstraction. In *International Conference on Machine Learning*, pages 2915–2923. PMLR, 2016. 23, 25

[5] David Abel, Dilip Arumugam, Lucas Lehnert, and Michael Littman. State abstractions for lifelong reinforcement learning. In *International Conference on Machine Learning*, pages 10–19, 2018.

[6] David Abel, Dilip Arumugam, Kavosh Asadi, Yuu Jinnai, Michael L Littman, and Lawson LS Wong. State abstraction as compression in apprenticeship learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3134–3142, 2019. 22, 23, 25

[7] David Abel, Nate Umbanhowar, Khimya Khetarpal, Dilip Arumugam, Doina Precup, and Michael Littman. Value preserving state-action abstractions. In *International Conference on Artificial Intelligence and Statistics*, pages 1639–1650. PMLR, 2020. 23

[8] Shuchin Aeron, Venkatesh Saligrama, and Manqi Zhao. Information theoretic bounds for compressed sensing. *IEEE Transactions on Information Theory*, 56(10):5111–5130, 2010. 8

[9] Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. FLAMBE: Structural complexity and representation learning of low rank MDPs. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20095–20107. Curran Associates, Inc., 2020. 23

[10] Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems*, pages 1184–1194, 2017. 1, 22

[11] Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken ELBO. In *International Conference on Machine Learning*, pages 159–168. PMLR, 2018. 25

[12] Mauricio Araya-López, Vincent Thomas, and Olivier Buffet. Near-optimal BRL using optimistic local transitions. In *Proceedings of the 29th International Conference on Machine Learning*, pages 515–522, 2012. 22

[13] Suguru Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 18(1):14–20, 1972. 25

[14] Dilip Arumugam and Benjamin Van Roy. Randomized value functions via posterior state-abstraction sampling. *arXiv preprint arXiv:2010.02383*, 2020. 23

[15] Dilip Arumugam and Benjamin Van Roy. Deciding what to learn: A rate-distortion approach. In *International Conference on Machine Learning*, pages 373–382. PMLR, 2021. 2, 5, 6, 7, 22, 25, 29

[16] Dilip Arumugam and Benjamin Van Roy. The value of information when deciding what to learn. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 6, 22, 25

[17] Kavosh Asadi, Dipendra Misra, and Michael Littman. Lipschitz continuity in model-based reinforcement learning. In *International Conference on Machine Learning*, pages 264–273. PMLR, 2018. 5

[18] John Asmuth, Lihong Li, Michael L Littman, Ali Nouri, and David Wingate. A Bayesian sampling approach to exploration in reinforcement learning. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 19–26, 2009. 22

[19] Karl Johan Åström. Optimal control of Markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1):174–205, 1965. 23

[20] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 89–96, 2009. 1, 22

[21] Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR, 2020. 5

[22] Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017. 1, 22

[23] Peter L Bartlett and Ambuj Tewari. REGAL: a regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 35–42, 2009. 1, 22

[24] Richard Bellman. A Markovian decision process. *Journal of Mathematics and Mechanics*, pages 679–684, 1957. 3

[25] Toby Berger. *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Prentice-Hall, 1971. 2, 4, 10

[26] Toby Berger and Jerry D Gibson. Lossy source coding. *IEEE Transactions on Information Theory*, 44(6):2693–2723, 1998. 4

[27] Donald A. Berry, Robert W. Chen, Alan Zame, David C. Heath, and Larry A. Shepp. Bandit problems with infinitely many arms. *Ann. Statist.*, 25(5):2103–2116, 10 1997. 22

[28] Dimitri Bertsekas and Steven E Shreve. *Stochastic optimal control: the discrete-time case*, volume 5. Athena Scientific, 1996. 4

[29] Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 1995. 5

[30] Dimitri P. Bertsekas and David A. Castañon. Adaptive aggregation methods for infinite horizon dynamic programming. *IEEE Transactions on Automatic Control*, 34(6):589–598, 1989. 23

[31] Richard Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, 18(4):460–473, 1972. 25

[32] Thomas Bonald and Alexandre Proutiere. Two-target algorithms for infinite-armed bandits with Bernoulli rewards. In *Advances in Neural Information Processing Systems*, pages 2184–2192, 2013. 22

[33] Vivek S Borkar, Sanjoy Mitter, and Sekhar Tatikonda. Markov control problems under communication contraints. *Communications in Information and Systems*, 1(1):15–32, 2001. 22

[34] Pinhas Boukris. An upper bound on the speed of convergence of the Blahut algorithm for computing rate-distortion functions (corresp.). *IEEE Transactions on Information Theory*, 19 (5):708–709, 1973. 25

[35] Ronen I Brafman and Moshe Tennenholtz. R-MAX-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002. 1, 22

[36] Sébastien Bubeck and Mark Sellke. First-order Bayesian regret analysis of Thompson sampling. In *Algorithmic Learning Theory*, pages 196–233. PMLR, 2020. 7

[37] Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12(5), 2011. 22

[38] Sébastien Bubeck, Nicolò Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012. 2, 22

[39] Pablo Samuel Castro and Doina Precup. Using linear programming for Bayesian exploration in Markov decision processes. In *IJCAI*, volume 24372442, 2007. 22

[40] Mung Chiang and Stephen Boyd. Geometric programming duals of channel capacity and rate distortion. *IEEE Transactions on Information Theory*, 50(2):245–258, 2004. 25

[41] Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012. 3, 4, 10, 30, 33, 35

[42] Imre Csiszár. On the computation of rate-distortion functions (corresp.). *IEEE Transactions on Information Theory*, 20(1):122–124, 1974. 25

[43] Imre Csiszár. On an extremum problem of information theory. *Studia Scientiarum Mathematicarum Hungarica*, 9, 1974. 5, 8, 25

[44] Brandon Cui, Yinlam Chow, and Mohammad Ghavamzadeh. Control-aware representations for model-based reinforcement learning. In *International Conference on Learning Representations*, 2020. 5

[45] Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*, pages 2818–2826, 2015. 1, 22

[46] Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying PAC and regret: uniform PAC bounds for episodic reinforcement learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5717–5727, 2017. 1, 22

[47] Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. On oracle-efficient PAC RL with rich observations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 1429–1439, 2018. 22

[48] Justin Dauwels. Numerical computation of the capacity of continuous memoryless channels. In *Proceedings of the 26th Symposium on Information Theory in the BENELUX*, pages 221–228. Citeseer, 2005. 25

[49] Peter Dayan. Improving generalization for temporal difference learning: The successor representation. *Neural computation*, 5(4):613–624, 1993. 26

[50] Thomas Dean and Robert Givan. Model minimization in Markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 106–111. AAAI Press, 1997. 23

[51] Richard Dearden, Nir Friedman, and Stuart Russell. Bayesian Q-learning. In *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence*, pages 761–768, 1998. 22

[52] Richard Dearden, Nir Friedman, and David Andre. Model based Bayesian exploration. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 150–159, 1999. 22

[53] Yash Deshpande and Andrea Montanari. Linear bandits in high dimension and recommendation systems. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1750–1754. IEEE, 2012. 22

[54] Shi Dong and Benjamin Van Roy. An information-theoretic analysis for Thompson sampling with many actions. In *Advances in Neural Information Processing Systems*, pages 4157–4165, 2018. 7, 22

[55] Shi Dong, Benjamin Van Roy, and Zhengyuan Zhou. Provably efficient reinforcement learning with aggregated states. *arXiv preprint arXiv:1912.06366*, 2019. 23

[56] Shi Dong, Benjamin Van Roy, and Zhengyuan Zhou. Simple agent, complex environment: Efficient reinforcement learning with agent state. *arXiv preprint arXiv:2102.05261*, 2021. 1, 22, 23

[57] Pierluca D'Oro, Alberto Maria Metelli, Andrea Tirinzoni, Matteo Papini, and Marcello Restelli. Gradient-aware model-based policy search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3801–3808, 2020. 5

[58] Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient RL with rich observations via latent state decoding. In *International Conference on Machine Learning*, pages 1665–1674. PMLR, 2019. 22, 23

[59] John C. Duchi. *Lecture Notes for Statistics 311/Electrical Engineering 377, Stanford University.* 2021. 3, 33, 36

[60] John C. Duchi and Martin J. Wainwright. Distance-based and continuum Fano inequalities with applications to statistical estimation. *arXiv preprint arXiv:1311.2669*, 2013. 8, 33

[61] Michael O'Gordon Duff. *Optimal Learning: Computational procedures for Bayes-adaptive Markov decision processes*. University of Massachusetts Amherst, 2002. 22

[62] Gabriel Dulac-Arnold, Nir Levine, Daniel J Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Machine Learning*, pages 1–50, 2021. 10

[63] Vikranth Dwaracherla, Xiuyuan Lu, Morteza Ibrahimi, Ian Osband, Zheng Wen, and Benjamin Van Roy. Hypermodels for exploration. In *International Conference on Learning Representations*, 2020. 2, 25

[64] Robert M. Fano. *Class Notes for MIT Course 6.574: Transmission of Information, MIT, Cambridge, MA.* 1952. 8, 33

[65] Amir-massoud Farahmand. Iterative value-aware model learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 9090–9101, 2018. 5

[66] Amir-massoud Farahmand, Andre Barreto, and Daniel Nikovski. Value-aware loss function for model-based reinforcement learning. In *Artificial Intelligence and Statistics*, pages 1486–1494. PMLR, 2017. 5, 7

[67] Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite Markov Decision Processes. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 162–169, 2004. 23

[68] Norman Ferns, Pablo Samuel Castro, Doina Precup, and Prakash Panangaden. Methods for computing state similarity in Markov Decision Processes. *arXiv preprint arXiv:1206.6836*, 2012. 23

[69] Sebastian Flennerhag, Jane X Wang, Pablo Sprechmann, Francesco Visin, Alexandre Galashov, Steven Kapturowski, Diana L Borsa, Nicolas Heess, Andre Barreto, and Razvan Pascanu. Temporal difference uncertainties as a signal for exploration. *arXiv preprint arXiv:2010.02255*, 2020. 26

[70] Mohammad Ghavamzadeh, Shie Mannor, Joelle Pineau, and Aviv Tamar. Bayesian reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, 8(5-6):359–483, 2015. 3, 22

[71] Robert M. Gray. *Entropy and information theory*. Springer Science & Business Media, 2011. 3

[72] Christopher Grimm, Andre Barreto, Satinder Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. *Advances in Neural Information Processing Systems*, 33, 2020. 5, 6, 7, 8

[73] Christopher Grimm, André Barreto, Greg Farquhar, David Silver, and Satinder Singh. Proper value equivalence. *Advances in Neural Information Processing Systems*, 34, 2021. 5, 7, 23

[74] Arthur Guez, David Silver, and Peter Dayan. Efficient Bayes-adaptive reinforcement learning using sample-based search. In *Proceedings of the 25th International Conference on Neural Information Processing Systems-Volume 1*, pages 1025–1033, 2012. 22

[75] Arthur Guez, David Silver, and Peter Dayan. Scalable and efficient Bayes-adaptive reinforcement learning based on Monte-Carlo tree search. *Journal of Artificial Intelligence Research*, 48:841–883, 2013.

[76] Arthur Guez, Nicolas Heess, David Silver, and Peter Dayan. Bayes-adaptive simulation-based search with value function approximation. In *Advances in Neural Information Processing Systems*, pages 451–459, 2014. 22

[77] Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual Markov decision processes. *arXiv preprint arXiv:1502.02259*, 2015. 25

[78] Botao Hao and Tor Lattimore. Regret bounds for information-directed reinforcement learning. *arXiv preprint arXiv:2206.04640*, 2022. 7

[79] Matthew T Harrison and Ioannis Kontoyiannis. Estimation of the rate–distortion function. *IEEE Transactions on Information Theory*, 54(8):3757–3762, 2008. 25

[80] Cassius T. Ionescu-Tulcea. Mesures dans les espaces produits. *Atti Acad. Naz. Lincei Rend. Cl Sci. Fis. Mat. Nat*, 8(7), 1949. 4

[81] Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(4), 2010. 1, 22

[82] David Janz, Jiri Hron, Przemysław Mazur, Katja Hofmann, José Miguel Hernández-Lobato, and Sebastian Tschiatschek. Successor uncertainties: exploration and uncertainty in temporal difference learning. *Advances in Neural Information Processing Systems*, 32, 2019. 26

[83] Nan Jiang, Alex Kulesza, and Satinder Singh. Abstraction selection in model-based reinforcement learning. In *International Conference on Machine Learning*, pages 179–188, 2015. 23

[84] Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low Bellman rank are PAC-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR, 2017. 22

[85] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 4868–4878, 2018. 1, 22

[86] Nicholas K Jong and Peter Stone. State abstraction discovery from irrelevant state variables. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pages 752–757, 2005. 23

[87] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996. 1

[88] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998. 23

[89] Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 267–274, 2002. 37

[90] Sham Machandranath Kakade. *On the Sample Complexity of Reinforcement Learning*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London, 2003. 1, 22

[91] Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49(2-3):209–232, 2002. 1, 22

[92] Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, pages 681–690, 2008. 22

[93] Andrei Nikolaevich Kolmogorov and Vladimir Mikhailovich Tikhomirov. ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk*, 14(2):3–86, 1959. 35

[94] J Zico Kolter and Andrew Y Ng. Near-Bayesian exploration in polynomial time. In *Proceedings of the 26th Annual International Conference on machine Learning*, pages 513–520, 2009. 22

[95] Victoria Kostina and Babak Hassibi. Rate-cost tradeoffs in control. *IEEE Transactions on Automatic Control*, 64(11):4525–4540, 2019. 22

[96] Akshay Krishnamurthy, Alekh Agarwal, and John Langford. PAC reinforcement learning with rich observations. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 1848–1856, 2016. 22

[97] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985. 2

[98] Tor Lattimore and Andras Gyorgy. Mirror descent and the information ratio. In *Conference on Learning Theory*, pages 2965–2992. PMLR, 2021. 7

[99] Tor Lattimore and Csaba Szepesvári. An information-theoretic approach to minimax regret in partial monitoring. In *Conference on Learning Theory*, pages 2111–2139. PMLR, 2019. 7

[100] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020. 2, 4, 22

[101] Lihong Li, Thomas J Walsh, and Michael L Littman. Towards a unified theory of state abstraction for MDPs. *ISAIM*, 4:5, 2006. 23, 25

[102] Evan Z Liu, Aditi Raghunathan, Percy Liang, and Chelsea Finn. Decoupling exploration and exploitation for meta-reinforcement learning without sacrifices. In *International conference on machine learning*, pages 6925–6935. PMLR, 2021. 25

[103] Xiuyuan Lu and Benjamin Van Roy. Ensemble sampling. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 3260–3268, 2017. 25

[104] Xiuyuan Lu and Benjamin Van Roy. Information-theoretic confidence bounds for reinforcement learning. *Advances in Neural Information Processing Systems*, 32:2461–2470, 2019. 7

[105] Xiuyuan Lu, Benjamin Van Roy, Vikranth Dwaracherla, Morteza Ibrahimi, Ian Osband, and Zheng Wen. Reinforcement Learning, Bit by Bit. *arXiv preprint arXiv:2103.04047*, 2021. 1, 2, 10, 22, 23

[106] Gerald Matz and Pierre Duhamel. Information geometric formulation and interpretation of accelerated Blahut-Arimoto-type algorithms. In *Information theory workshop*, pages 66–70. IEEE, 2004. 25

[107] Dipendra Misra, Mikael Henaff, Akshay Krishnamurthy, and John Langford. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *International Conference on Machine Learning*, pages 6961–6971. PMLR, 2020. 23

[108] Sanjoy Mitter and Anant Sahai. Information and control: Witsenhausen revisited. In *Learning, Control and Hybrid Systems*, pages 281–293. Springer, 1999. 22

[109] Sanjoy K Mitter. Control with limited information. *European Journal of Control*, 7(2-3): 122–131, 2001. 22

[110] Suraj Nair, Silvio Savarese, and Chelsea Finn. Goal-aware prediction: Learning to model what matters. In *International Conference on Machine Learning*, pages 7207–7219. PMLR, 2020. 5

[111] Ziad Naja, Florence Alberge, and Pierre Duhamel. Geometrical interpretation and improvements of the Blahut-Arimoto's algorithm. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2505–2508. IEEE, 2009. 25

[112] Urs Niesen, Devavrat Shah, and Gregory Wornell. Adaptive alternating minimization algorithms. In *2007 IEEE International Symposium on Information Theory*, pages 1641–1645. IEEE, 2007. 25

[113] Evgenii Nikishin, Romina Abachi, Rishabh Agarwal, and Pierre-Luc Bacon. Control-oriented model-based reinforcement learning with implicit differentiation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. 5

[114] Brendan O'Donoghue, Ian Osband, Remi Munos, and Volodymyr Mnih. The uncertainty Bellman equation and exploration. In *International Conference on Machine Learning*, pages 3836–3845, 2018. 2, 26

[115] Brendan O'Donoghue, Ian Osband, and Catalin Ionescu. Making sense of reinforcement learning and probabilistic inference. In *International Conference on Learning Representations*, 2020. 25

[116] Junhyuk Oh, Satinder Singh, and Honglak Lee. Value prediction network. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6120–6130, 2017. 5, 23

[117] Ian Osband. *Deep Exploration via Randomized Value Functions*. PhD thesis, Stanford University, 2016. 26

[118] Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the Eluder dimension. *Advances in Neural Information Processing Systems*, 27, 2014. 22

[119] Ian Osband and Benjamin Van Roy. Gaussian-Dirichlet posterior dominance in sequential learning. *arXiv preprint arXiv:1702.04126*, 2017. 9

[120] Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *International Conference on Machine Learning*, pages 2701–2710. PMLR, 2017. 1, 2, 6, 7, 9, 22, 38

[121] Ian Osband, Daniel Russo, and Benjamin Van Roy. (More) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26:3003–3011, 2013. 1, 2, 7, 9, 22, 26, 38

[122] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via Bootstrapped DQN. In *Advances in Neural Information Processing Systems*, pages 4026–4034, 2016. 25, 26

[123] Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pages 2377–2386, 2016. 2

[124] Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 31, 2018. 25, 26

[125] Ian Osband, Benjamin Van Roy, Daniel J Russo, and Zheng Wen. Deep exploration via randomized value functions. *Journal of Machine Learning Research*, 20(124):1–62, 2019. 2, 26

[126] Ian Osband, Zheng Wen, Mohammad Asghari, Morteza Ibrahimi, Xiyuan Lu, and Benjamin Van Roy. Epistemic neural networks. *arXiv preprint arXiv:2107.08924*, 2021. 25, 26

[127] Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Botao Hao, Morteza Ibrahimi, Dieterich Lawson, Xiuyuan Lu, Brendan O'Donoghue, and Benjamin Van Roy. Evaluating predictive distributions: Does Bayesian deep learning work? *arXiv preprint arXiv:2110.04629*, 2021. 25

[128] Hari Palaiyanur and Anant Sahai. On the uniform continuity of the rate-distortion function. In *2008 IEEE International Symposium on Information Theory*, pages 857–861. IEEE, 2008. 25

[129] Yury Polyanskiy and Yihong Wu. Lecture notes on information theory. 2019. 3

[130] Martin L. Puterman. *Markov Decision Processes—Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, 1994. 3

[131] Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *International conference on machine learning*, pages 5331–5340. PMLR, 2019. 25

[132] Kenneth Rose. A mapping approach to rate-distortion computation and analysis. *IEEE Transactions on Information Theory*, 40(6):1939–1952, 1994. 25

[133] Jonathan Rubin, Ohad Shamir, and Naftali Tishby. Trading value and information in MDPs. In *Decision Making with Imperfect Decision Makers*, pages 57–74. Springer, 2012. 22

[134] Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010. 22

[135] Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Advances in Neural Information Processing Systems*, 27:1583–1591, 2014. 7, 22, 38

[136] Daniel Russo and Benjamin Van Roy. An information-theoretic analysis of Thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016. 7

[137] Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252, 2018. 7, 22

[138] Daniel Russo and Benjamin Van Roy. Satisficing in time-sensitive bandit learning. *arXiv preprint arXiv:1803.02855*, 2018. 2, 5, 6, 7, 22

[139] Daniel Russo and Benjamin Van Roy. Satisficing in time-sensitive bandit learning. *Mathematics of Operations Research*, 2022. 2, 5, 6, 7, 22

[140] Daniel Russo, David Tse, and Benjamin Van Roy. Time-sensitive bandit learning and satisficing Thompson sampling. *arXiv preprint arXiv:1704.09028*, 2017. 2, 6, 22

[141] Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on Thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018. 1, 22

[142] Ilya O Ryzhov, Warren B Powell, and Peter I Frazier. The knowledge gradient algorithm for a general class of online learning problems. *Operations Research*, 60(1):180–195, 2012. 22

[143] Jossy Sayir. Iterating the Arimoto-Blahut algorithm for faster convergence. In *2000 IEEE International Symposium on Information Theory (Cat. No. 00CH37060)*, page 235. IEEE, 2000. 25

[144] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering Atari, Go, Chess and Shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020. 5, 23, 25

[145] Ehsan Shafieepoorfard, Maxim Raginsky, and Sean P Meyn. Rationally inattentive control of Markov processes. *SIAM Journal on Control and Optimization*, 54(2):987–1016, 2016. 22, 25

[146] Claude E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec., March 1959*, 4:142–163, 1959. 2, 4, 10, 22

[147] David Silver, Hado Hasselt, Matteo Hessel, Tom Schaul, Arthur Guez, Tim Harley, Gabriel Dulac-Arnold, David Reichert, Neil Rabinowitz, Andre Barreto, et al. The Predictron: End-to-end learning and planning. In *International Conference on Machine Learning*, pages 3191–3199. PMLR, 2017. 5, 23

[148] Herbert A. Simon. Models of bounded rationality. *Economic Analysis and Public Policy, MIT Press, Cambridge, Mass*, 1982. 6

[149] Jonathan Sorg, Satinder Singh, and Richard L Lewis. Variance-based rewards for approximate Bayesian reinforcement learning. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 564–571, 2010. 22

[150] Jan-Erik Stjernvall. Dominance–a relation between distortion measures. *IEEE Transactions on Information Theory*, 29(6):798–807, 1983. 8, 32, 33

[151] Alexander L Strehl, Lihong Li, and Michael L Littman. Reinforcement learning in finite MDPs: PAC analysis. *Journal of Machine Learning Research*, 10(Nov):2413–2444, 2009. 1, 22

[152] Malcolm JA Strens. A Bayesian framework for reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 943–950, 2000. 2, 6, 22

[153] Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based RL in contextual decision processes: PAC bounds and exponential improvements over model-free approaches. In *Conference on Learning Theory*, pages 2898–2933. PMLR, 2019. 22

[154] Richard S Sutton and Andrew G Barto. Introduction to reinforcement learning. 1998. 1

[155] Sekhar Tatikonda and Sanjoy Mitter. Control under communication constraints. *IEEE Transactions on Automatic Control*, 49(7):1056–1068, 2004. 22

[156] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933. 1, 22

[157] Naftali Tishby and Daniel Polani. Information theory of decisions and actions. In *Perception-action cycle*, pages 601–636. Springer, 2011. 22

[158] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000. 9, 25

[159] Benjamin Van Roy. Performance loss bounds for approximate value iteration with state aggregation. *Mathematics of Operations Research*, 31(2):234–244, 2006. 23, 25

[160] Sergio Verdú et al. Generalizing the Fano inequality. *IEEE Transactions on Information Theory*, 40(4):1247–1251, 1994. 8

[161] Claas A Voelcker, Victor Liao, Animesh Garg, and Amir-massoud Farahmand. Value gradient weighted model-based reinforcement learning. In *International Conference on Learning Representations*, 2022. 5, 7

[162] Pascal O Vontobel, Aleksandar Kavcic, Dieter M Arnold, and Hans-Andrea Loeliger. A generalization of the Blahut–Arimoto algorithm to finite-state channels. *IEEE Transactions on Information Theory*, 54(5):1887–1918, 2008. 25

[163] Tao Wang, Daniel Lizotte, Michael Bowling, and Dale Schuurmans. Bayesian sparse sampling for on-line reward optimization. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 956–963, 2005. 22

[164] Yizao Wang, Jean-Yves Audibert, and Rémi Munos. Algorithms for infinitely many-armed bandits. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, pages 1729–1736, 2008. 22

[165] Ward Whitt. Approximations of dynamic programs, I. *Mathematics of Operations Research*, 3 (3):231–243, 1978. 23

[166] Hans S Witsenhausen. Separation of estimation and control for discrete time systems. *Proceedings of the IEEE*, 59(11):1557–1566, 1971. 22

[167] Yaming Yu. Squeezing the Arimoto–Blahut algorithm for faster convergence. *IEEE Transactions on Information Theory*, 56(7):3149–3157, 2010. 25

[168] 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. 1, 22

[169] Julian Zimmert and Tor Lattimore. Connections between mirror descent, Thompson sampling and the information ratio. In *Advances in Neural Information Processing Systems*, pages 11973–11982, 2019. 7

[170] Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. VariBAD: A Very Good Method for Bayes-Adaptive Deep RL via Meta-Learning. In *International Conference on Learning Representations*, 2019. 25

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

A Related Work

This paper follows suit with a long line of work on provably-efficient reinforcement learning [91, 35, 90, 20, 23, 151, 81, 121, 45, 120, 22, 46, 10, 85, 168, 56, 105]. As previously discussed, these methods can be categorized based on their use of optimism in the face of uncertainty or posterior sampling to address the exploration challenge. Notably, methods in the latter category are Bayesian reinforcement-learning algorithms [70] that, through their use of Thompson sampling [156, 141], are exclusively concerned with identifying optimal solutions. The notable exception to this statement is the method of Lu et al. [105], which is based on information-directed sampling [135, 137]; while their analysis does accommodate other learning targets besides the optimal policy, an agent designer is responsible for supplying this target to the agent a priori whereas we adaptively compute an information-theoretically sound target grounded in rate-distortion theory.

In contrast to this class of approaches, optimism-based methods tend to obey PAC-MDP guarantees [90, 151] which, given a fixed parameter $\varepsilon > 0$, offer a high-probability bound on the total number of timesteps for which the agent's behavior is worse than ε-sub-optimal. Through this tolerance parameter ε, an agent designer can express a preference for efficiently identifying a deliberately sub-optimal solution; our work can be seen as providing an analogous knob for Bayesian reinforcement-learning methods that deliberately pursue a satisficing solution while also remaining competitive with regret guarantees for optimism-based methods [45, 46, 85, 168]. In this way, our theoretical guarantees are more general than those for PSRL [121, 118, 2, 120, 10]. Importantly, the nature of our contribution is not to be confused with the PAC-BAMDP framework of Kolter and Ng [94] which characterizes algorithms that adhere to a high-probability bound on the total number of sub-optimal timesteps relative to the Bayes-optimal policy [18, 149]. We refer readers to the work of Ghavamzadeh et al. [70] for a broader survey of Bayesian reinforcement-learning methods, including those which do not employ posterior sampling [152], but instead entertain other approximations [51, 52, 163, 39, 12, 74–76] to tractably solve the resulting Bayes-Adaptive Markov Decision Process (BAMDP) [61], typically while foregoing rigorous theoretical guarantees.

A perhaps third distinct class of provably-efficient reinforcement-learning algorithms [96, 84, 47, 58, 153] proceeds by iteratively selecting an element of a function class (typically denoting a collection of regressors for either a value function or transition model), inducing a policy from the chosen function, and then carefully eliminating all hypotheses of the function class that are inconsistent with the observed data resulting from policy rollouts in the environment. To the extent that one might be willing to characterize this high-level algorithmic template as an iterative, manual compression and refinement of the initial function class, our algorithm can be seen as bringing the appropriate tool of rate-distortion theory to bear on the inherent lossy compression problem and developing the complementary information-theoretic analysis.

The concept of designing algorithms to learn such near-optimal or satisficing solutions has been well-studied in the multi-armed bandit setting [38, 100]. Indeed, the need to forego optimizing for an optimal arm arises naturally in various contexts [37, 92, 134, 142, 53, 27, 164, 32]. A general study of such satisficing solutions through the lens of information theory was first proposed by Russo et al. [140], Russo and Van Roy [138, 139] and later extended to develop practical algorithms by Arumugam and Van Roy [15, 16]. Our work provides the natural, theoretical extension of these ideas to the full reinforcement-learning setting, leaving investigation of practical instantiations to future work (see Section B). The algorithm and regret bound we provide bears some resemblance to the compressed Thompson sampling algorithm of Dong and Van Roy [54] for bandit problems. Crucially, while the compressive statistic of the environment utilized by their algorithm is computed once a priori, our algorithm recomputes its learning target in each episode, refining it as the agent's knowledge of the true environment accumulates. Similar to these prior works, we leverage rate-distortion theory [146] as a principled tool for a mathematically-precise characterization of satisficing solutions. We simply note that our use of rate-distortion theory for reinforcement learning in this work stands in stark contrast to that of prior work which examines state abstraction in reinforcement learning [6] or attempts to control the entropy of the resulting policy [157, 133, 145].

We also recognize the connection between this work and prior work at the intersection of information theory and control theory [166, 108, 109, 33, 155, 95]. These works parallel our setting in their consideration for an agent that must stabilize a system with limited *observational* capacity, augmenting the standard control objective subject to a constraint on the rate of the channel that processes raw observations; this problem formulation more closely aligns with a partially-observable Markov

Decision Process [19, 88] or an agent learning with a state abstraction [101, 4, 159]. In contrast, our work is concerned with an overall limit on the total amount of information an agent may acquire from the environment and, in turn, how that translates into its selection of a feasible learning target. That said, we suspect there could be a strong, subtle synergy between these prior works and the capacity-sensitive performance guarantees for our algorithm (see Section 5.3).

B Discussion

In this section, we outline connections between VSRL and follow-up work to the value equivalence principle [73], explore opportunities for even further compression through state abstraction [101, 4], and contemplate potential avenues for how our theory might inform practice.

B.1 Proper Value Equivalence

While the value equivalence principle examines a single application of each Bellman operator, in follow-up work Grimm et al. [73] introduce the notion of proper value equivalence, which considers the limit of infinitely many applications or, stated more concisely, the fixed points of the associated operators. A model $\widetilde{\mathcal{M}}$ is proper value equivalent if $V^{\pi}_{\mathcal{M}^\star,h} = V^{\pi}_{\widetilde{\mathcal{M}},h}$, $\forall \pi \in \Pi$, $h \in [H]$. This notion allows for a simpler parameterization through the policy class Π alone, without the need for a complementary value function class $\mathcal{V}$. Conveniently, through Proposition 2 of Grimm et al. [73], it follows that to obtain the set of proper value equivalent models with respect to Π, one need only find the set of models that are value equivalent for each $\pi \in \Pi$ and its induced value function, V^π. In our context, we can establish an approximate version of this by using the distortion function $d_{\Pi,\mathcal{V}}$ where $\mathcal{V} = \{V^\pi \mid \pi \in \Pi^H\}$ (recall that previous results obeyed the less stringent condition that $\mathcal{V} \supseteq \{V^\pi \mid \pi \in \Pi^H\}$).

Grimm et al. [73] go on to study proper value equivalence for the set of all deterministic policies, $\Pi = \{\mathcal{S} \to \mathcal{A}\}$ and, through their Corollary 1, show that an optimal policy for any model which is proper value equivalent to Π is also optimal in the original MDP $\mathcal{M}^\star$. Again, we recall that our prior guarantees were made under the less restrictive assumption that $\Pi \supseteq \{\mathcal{S} \to \mathcal{A}\}$. Coupling these insights on proper value equivalence together, we see that when VSRL is run with $\Pi = \{\mathcal{S} \to \mathcal{A}\}$ and $\mathcal{V} = \{V^\pi \mid \pi \in \Pi^H\}$, the agent aims to recover an approximately proper value-equivalent model of the true environment and, when $D = 0$, the optimal policy associated with this compressed MDP will be optimal for $\mathcal{M}^\star$. Finally, through their Proposition 5, Grimm et al. [73] identify the set of all proper value equivalent models with respect to $\{\mathcal{S} \to \mathcal{A}\}$ as the largest possible value equivalence class that is guaranteed to yield optimal performance in the true environment. Meanwhile, our Lemma 1 again establishes the information-theoretic analogue of this claim; namely, that VSRL configured to learn a model from this largest value equivalence class requires the fewest bits of information from the true environment. The importance of proper value equivalence culminates with experiments that highlight how MuZero [144] succeeds by optimizing a proper value-equivalent loss function. We leave to future work the question of how VSRL might pave the way towards more principled exploration strategies for practical algorithms like MuZero.

B.2 Greater Compression via State Abstraction

A core disconnect between VSRL and contemporary deep model-based reinforcement learning approaches is that our lossy compression problem forces VSRL to identify a model defined with respect to the original state space whereas methods in the latter category learn a model with respect to a state abstraction. Indeed, algorithms like MuZero and its predecessors [147, 116, 144] never approximate reward functions and transition models with respect to the raw image observations generated by the environment, but instead incrementally learn some latent representation of state upon which a corresponding model is approximated for planning. This philosophy is born out of several years of work that elucidate the important of state abstraction as a key tool for avoiding the irrelevant information encoded in environment states and addressing the challenge of generalization for sample-efficient reinforcement learning large-scale environments [165, 30, 50, 67, 86, 101, 159, 68, 83, 4–6, 55, 58, 14, 107, 9, 7, 3, 56]. In this section, we briefly introduce a small extension of VSRL that builds on these insights to accommodate lossy MDP compressions defined on a simpler, abstract state space (also referred to as aleatoric or situational state by Lu et al. [105], Dong et al. [56]).

Let $\Phi \subseteq \{\mathcal{S} \to [Z]\}$ denote a class of state abstractions or quantizers which map environment states to some discrete, finite abstract state space containing a known, fixed number of abstract states $Z \in \mathbb{N}$. For any abstract action-value function $Q_\phi \in \{[Z] \times \mathcal{A} \to \mathbb{R}\}$ and any state abstraction $\phi \in \Phi$, we denote by $Q_\phi \circ \phi \in \{\mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$ the composition of the state abstraction and abstract value function such that $Q_\phi \circ \phi$ is a value function for the original MDP. We adopt a similar convention for any policy $\pi_\phi \in \{[Z] \to \Delta(\mathcal{A})\}$ such that $\pi_\phi \circ \phi \in \{\mathcal{S} \to \Delta(\mathcal{A})\}$. We now consider carrying out the rate-distortion optimization of VSRL in each episode over abstract MDPs such that $\widetilde{\mathcal{M}}_k \in \mathfrak{M}_\phi \triangleq \{[Z] \times \mathcal{A} \to [0,1]\} \times \{[Z] \times \mathcal{A} \to \Delta([Z])\}$. Just as before, we take the input information source to our lossy compression problem in each episode $k \in [K]$ as the agent's current beliefs over the true MDP, $\mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_k)$. Unlike the preceding sections, our distortion function $d : \mathfrak{M} \times \mathfrak{M}_\phi \to \mathbb{R}_{\geq 0}$ must now quantify the loss of fidelity incurred by using a compressed abstract MDP in lieu of the true environment MDP. Consequently, we define a new distortion function

$$d_\Phi(\mathcal{M}, \widehat{\mathcal{M}}) = \sup_{\phi \in \Phi} \sup_{h \in [H]} ||Q^\star_{\mathcal{M},h} - Q^\star_{\widehat{\mathcal{M}},h} \circ \phi||^2_\infty = \sup_{\phi \in \Phi} \sup_{h \in [H]} \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} |Q^\star_{\mathcal{M},h}(s,a) - Q^\star_{\widehat{\mathcal{M}},h}(\phi(s),a)|^2,$$

whose corresponding rate-distortion function is given by

$$\mathcal{R}^\Phi_k(D) = \inf_{\widetilde{\mathcal{M}} \in \Lambda_k(D)} \mathbb{I}_k(\mathcal{M}^\star; \widetilde{\mathcal{M}}) \qquad \Lambda_k(D) \triangleq \{\widetilde{\mathcal{M}} : \Omega \to \mathfrak{M} \mid \mathbb{E}_k\left[d_\Phi(\mathcal{M}^\star, \widetilde{\mathcal{M}})\right] \leq D\}.$$

Unlike when performing a lossy compression where $\widetilde{\mathcal{M}}_k \in \mathfrak{M}$, the channel that represents the identity mapping is no longer a viable option as we must now generate an abstract MDP that resides in $\mathfrak{M}_\phi$. Consequently, we require the following assumption on Φ to ensure that the set of channels over which we compute the infimum of $\mathcal{R}^\Phi_k(D)$ is non-empty.

Assumption 1. *For each $k \in [K]$, we have that $\Lambda_k(D) \neq \emptyset$.*

Algorithm 3 Compressed Value-equivalent Sampling for Reinforcement Learning (Compressed-VSRL)

Input: Prior distribution $\mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_1)$, Distortion threshold $D \in \mathbb{R}_{\geq 0}$, State abstraction class Φ, Distortion function $d_\Phi : \mathfrak{M} \times \mathfrak{M}_\phi \to \mathbb{R}_{\geq 0}$,
for $k \in [K]$ **do**
  Compute channel $\mathbb{P}(\widetilde{\mathcal{M}}_k \in \cdot | \mathcal{M}^\star)$ achieving $\mathcal{R}^\Phi_k(D)$ limit
  Sample MDP $M^\star \sim \mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_k)$
  Sample compressed MDP $M_k \sim \mathbb{P}(\widetilde{\mathcal{M}}_k \in \cdot \mid \mathcal{M}^\star = M^\star)$
  Set state abstraction ϕ_k to achieve the infimum: $\inf_{\phi \in \Phi} \sup_{h \in [H]} ||Q^\star_{M^\star,h} - Q^\star_{M_k,h} \circ \phi||^2_\infty$
  Compute optimal policy $\pi^\star_{M_k}$ and set $\pi^{(k)} = \pi^\star_{M_k} \circ \phi_k$
  Execute $\pi^{(k)}$ and observe trajectory τ_k
  Update history $H_{k+1} = H_k \cup \tau_k$
  Induce posterior $\mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_{k+1})$
end for

We present our Compressed-VSRL extension as Algorithm 3 which incorporates an additional step beyond VSRL to govern the choice of state abstraction utilized in conjunction with the sampled compressed MDP in each episode.

We strongly suspect that an analysis paralleling that of Corollaries 1 and 2, with an appropriately defined information ratio, can be carried out for Compressed-VSRL as well. However, for the sake of brevity and since the result is neither immediate nor trivial, we leave the information-theoretic regret bound stated as a conjecture.

Conjecture 1. *Fix any $D > 0$. For any prior distribution $\mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_1)$, if $\Gamma_k \leq \overline{\Gamma}$ for all $k \in [K]$, then CVSRL (Algorithm 3) with distortion function d_Φ has*

$$\textsc{BayesRegret}(K, \pi^{(1)}, \ldots, \pi^{(K)}) \leq \sqrt{\overline{\Gamma} K \mathcal{R}^\Phi_1(D)} + 2K(H+1)\sqrt{D}.$$

The significance of Conjecture 1 for allowing a simple, bounded agent to contend with a complex environment manifests when considering analogues to Theorems 3 and 5. Specifically, for any finite-horizon, episodic MDP with a finite action space ($|\mathcal{A}| < \infty$), one may upper bound the rate-distortion function via the entropy in the abstract model $\mathbb{H}_1(\mathcal{R}_\phi, \mathcal{T}_\phi)$. Using the same proof technique as in the preceding results, this facilitates an upper bound $\mathcal{R}_1^\Phi(D) \leq \widetilde{\mathcal{O}}\left(Z^2|\mathcal{A}|\right)$ which lacks any dependence on the complexity of the (potentially infinite) environment state space, $\mathcal{S}$.

B.3 From Theory to Practice

While the performance guarantees of VSRL hold for any finite-horizon, episodic MDP, it is important to reconcile that generality with the practicality of the instantiating the algorithm. The three key barriers to practical, scalable implementations of VSRL applied to complex tasks of interest are the representation of epistemic uncertainty, the computation of the rate-distortion function, and the synthesis of optimal policies for sampled MDPs. The first point is a fundamental obstacle to Bayesian reinforcement-learning algorithms and recent work in deep reinforcement learning has found success with simple, albeit computationally-inefficient, ensembles of networks [122, 103, 124] or even hypermodels [63]. As progress is made towards more computationally-efficient models for representing and resolving epistemic uncertainty through Bayesian deep learning [126, 127], there will be greater potential for a practical implementation of VSRL.

For addressing the second issue, a classic option for computing the channel that achieves the rate-distortion limit is the Blahut-Arimoto algorithm [31, 13] which, in theory, is a well-defined procedure even for random variables defined on abstract alphabets [43, 42]. In practice, however, computing such a channel for continuous outputs remains an open challenge [48]; still, several analyses and refinements have been made to the algorithm so far [34, 132, 143, 106, 40, 112, 162, 111, 167], and the reinforcement-learning community stands to greatly benefit from further improvements. Continuous information sources, however, are less problematic as one may draw a sufficiently large number of i.i.d. samples and substitute this empirical distribution for the source, leading to the so-called plug-in estimator of the rate-distortion function for which consistency and sample-complexity guarantees are known [79, 128]. Moreover, empirical successes for such estimators have already been demonstrated in the multi-armed bandit setting [15, 16].

The last issue touches upon the fact that while tabular problems admit several planning algorithms for recovering the optimal policy associated with the sampled MDP in each episode, the same cannot be said for arbitrary state-action spaces. At best, one might hope for simply recovering an approximation to this policy through some high-dimensional model-based planning algorithm. We leave the questions of how to practically implement such a procedure and understand its impact on our theory to future work.

Of course, all of the aforementioned issues arise when trying to directly implement VSRL roughly as described by Algorithm 2. An alternative, however, is to ask how one might take existing practical algorithms already operating at scale (such as MuZero [144]) and bring those methods closer to the spirit of VSRL? Since these practical model-based reinforcement-learning algorithms are already engaging with some form of state abstraction [101, 4, 159], this might entail further consideration for information-theoretic approaches to guiding representation learning [6, 145] as a proxy to engaging with a rate-distortion trade-off. Additionally, one of the core insights developed in this work is the formalization of model simplifications arising out of the value equivalence principle as a form of lossy compression. Curiously, recent meta reinforcement-learning approaches applied to multi-task or contextual MDPs [77] arrive at a similar need of obtaining compressed representations of the underlying MDP model during a meta-exploration phase in order to facilitate few-shot learning [131, 170, 102]. Given such approaches for learning latent/contextual variational encoders, we strongly suspect that the known connections between rate-distortion theory and information bottlenecks [158, 11] represent a viable path to bridging the ideas between our theory and value equivalence in practice. Notably, while these methods adopt a probabilistic inference perspective and do articulate the encoders as posterior distributions over the underlying task, they lack any representation of epistemic uncertainty [115], similar to MuZero; consequently, this still leaves open the earlier obstacle of how best to represent and maintain notions of epistemic uncertainty in large-scale agents.

A third competing perspective is to recognize that recent empirical successes in Bayesian reinforcement learning often avoid representing uncertainty over the model of the environment in favor of the underlying optimal action-value functions, $\{Q^\star_{\mathcal{M}^\star,h}\}_{h\in[H]}$. Such approaches apply an algorithm

known as Randomized Value Functions (RVF), rather than PSRL [122–124, 114, 125]. Naturally, the optimal action-value functions of the true MDP $\mathcal{M}^\star$ and its model are related and, in fact, there is an equivalence between RVF and PSRL [117]. By maintaining epistemic uncertainty over the action-value functions $\mathbb{P}(\{Q^\star_{\mathcal{M}^\star,h}\}_{h\in[H]} \in \cdot \mid H_k)$ rather than the underlying model, RVF methods are amenable to practical instantiation with deep neural networks [122, 124, 114, 126]. Moreover, extensions of this line of work have gone on to consider other scalable avenues for leveraging such principled, practical solutions to the exploration challenge through successor features [49, 82] and temporal-difference errors [69]. Overall, we strongly suspect that VSRL can also play an analogous role to PSRL in this sense, providing a sound theoretical foundation that gives rise to subsequent practical algorithms of a slightly different flavor.

C Proof of Theorem 1

Before we can prove Theorem 1, we require the following lemma whose proof we adapt from Osband et al. [121]:

Lemma 4. *Let $\mathcal{M}, \widehat{\mathcal{M}}$ be two arbitrary finite-horizon, episodic MDPs with models $(\mathcal{R}, \mathcal{T})$ and $(\widehat{\mathcal{R}}, \widehat{\mathcal{T}})$, respectively. Then, for any non-stationary policy $\pi = (\pi_1, \ldots, \pi_H) \in \{\mathcal{S} \to \Delta(\mathcal{A})\}^H$,*

$$V^\pi_{\mathcal{M},1} - V^\pi_{\widehat{\mathcal{M}},1} = \sum_{h=1}^{H} \mathbb{E}\left[\mathcal{B}^{\pi_h}_{\mathcal{M}} V^\pi_{\mathcal{M},h+1}(s_h) - \mathcal{B}^{\pi_h}_{\widehat{\mathcal{M}}} V^\pi_{\mathcal{M},h+1}(s_h)\right] = \sum_{h=1}^{H} \mathbb{E}\left[\mathcal{B}^{\pi_h}_{\mathcal{M}} V^\pi_{\widehat{\mathcal{M}},h+1}(s_h) - \mathcal{B}^{\pi_h}_{\widehat{\mathcal{M}}} V^\pi_{\widehat{\mathcal{M}},h+1}(s_h)\right].$$

Proof. By simply applying the Bellman equations, we have

$$\begin{aligned}
V^\pi_{\mathcal{M},1} - V^\pi_{\widehat{\mathcal{M}},1} &= \mathbb{E}\left[V^\pi_{\mathcal{M},1}(s_1) - V^\pi_{\widehat{\mathcal{M}},1}(s_1)\right] \\
&= \mathbb{E}\left[\mathcal{B}^{\pi_1}_{\mathcal{M}} V^\pi_{\mathcal{M},2}(s_1) - \mathcal{B}^{\pi_1}_{\widehat{\mathcal{M}}} V^\pi_{\widehat{\mathcal{M}},2}(s_1)\right] \\
&= \mathbb{E}\left[\mathcal{B}^{\pi_1}_{\mathcal{M}} V^\pi_{\mathcal{M},2}(s_1) - \mathcal{B}^{\pi_1}_{\widehat{\mathcal{M}}} V^\pi_{\mathcal{M},2}(s_1) + \mathcal{B}^{\pi_1}_{\widehat{\mathcal{M}}} V^\pi_{\mathcal{M},2}(s_1) - \mathcal{B}^{\pi_1}_{\widehat{\mathcal{M}}} V^\pi_{\widehat{\mathcal{M}},2}(s_1)\right] \\
&= \mathbb{E}\left[\mathcal{B}^{\pi_1}_{\mathcal{M}} V^\pi_{\mathcal{M},2}(s_1) - \mathcal{B}^{\pi_1}_{\widehat{\mathcal{M}}} V^\pi_{\mathcal{M},2}(s_1) + \mathbb{E}_{s_2\sim\widehat{\mathcal{T}}(\cdot|s_1,a_1)}\left[V^\pi_{\mathcal{M},2}(s_2) - V^\pi_{\widehat{\mathcal{M}},2}(s_2)\right]\right] \\
&= \sum_{h=1}^{2} \mathbb{E}\left[\mathcal{B}^{\pi_h}_{\mathcal{M}} V^\pi_{\mathcal{M},h+1}(s_h) - \mathcal{B}^{\pi_h}_{\widehat{\mathcal{M}}} V^\pi_{\mathcal{M},h+1}(s_h)\right] + \mathbb{E}\left[V^\pi_{\mathcal{M},3}(s_3) - V^\pi_{\widehat{\mathcal{M}},3}(s_3)\right] \\
&= \ldots \\
&= \sum_{h=1}^{H} \mathbb{E}\left[\mathcal{B}^{\pi_h}_{\mathcal{M}} V^\pi_{\mathcal{M},h+1}(s_h) - \mathcal{B}^{\pi_h}_{\widehat{\mathcal{M}}} V^\pi_{\mathcal{M},h+1}(s_h)\right] + \underbrace{\mathbb{E}\left[V^\pi_{\mathcal{M},H+1}(s_{H+1}) - V^\pi_{\widehat{\mathcal{M}},H+1}(s_{H+1})\right]}_{=0} \\
&= \sum_{h=1}^{H} \mathbb{E}\left[\mathcal{B}^{\pi_h}_{\mathcal{M}} V^\pi_{\mathcal{M},h+1}(s_h) - \mathcal{B}^{\pi_h}_{\widehat{\mathcal{M}}} V^\pi_{\mathcal{M},h+1}(s_h)\right].
\end{aligned}$$

For the second identity, we have nearly identical steps:

$$
\begin{aligned}
V^{\pi}_{\mathcal{M},1} - V^{\pi}_{\widehat{\mathcal{M}},1} &= \mathbb{E}\left[V^{\pi}_{\mathcal{M},1}(s_1) - V^{\pi}_{\widehat{\mathcal{M}},1}(s_1)\right] \\
&= \mathbb{E}\left[\mathcal{B}^{\pi_1}_{\mathcal{M}} V^{\pi}_{\mathcal{M},2}(s_1) - \mathcal{B}^{\pi_1}_{\widehat{\mathcal{M}}} V^{\pi}_{\widehat{\mathcal{M}},2}(s_1)\right] \\
&= \mathbb{E}\left[\mathcal{B}^{\pi_1}_{\mathcal{M}} V^{\pi}_{\mathcal{M},2}(s_1) - \mathcal{B}^{\pi_1}_{\mathcal{M}} V^{\pi}_{\widehat{\mathcal{M}},2}(s_1) + \mathcal{B}^{\pi_1}_{\mathcal{M}} V^{\pi}_{\widehat{\mathcal{M}},2}(s_1) - \mathcal{B}^{\pi_1}_{\widehat{\mathcal{M}}} V^{\pi}_{\widehat{\mathcal{M}},2}(s_1)\right] \\
&= \mathbb{E}\left[\mathcal{B}^{\pi_1}_{\mathcal{M}} V^{\pi}_{\widehat{\mathcal{M}},2}(s_1) - \mathcal{B}^{\pi_1}_{\widehat{\mathcal{M}}} V^{\pi}_{\widehat{\mathcal{M}},2}(s_1) + \mathbb{E}_{s_2 \sim \mathcal{T}(\cdot|s_1,a_1)}\left[V^{\pi}_{\mathcal{M},2}(s_2) - V^{\pi}_{\widehat{\mathcal{M}},2}(s_2)\right]\right] \\
&= \sum_{h=1}^{2} \mathbb{E}\left[\mathcal{B}^{\pi_h}_{\mathcal{M}} V^{\pi}_{\widehat{\mathcal{M}},h+1}(s_h) - \mathcal{B}^{\pi_h}_{\widehat{\mathcal{M}}} V^{\pi}_{\widehat{\mathcal{M}},h+1}(s_h)\right] + \mathbb{E}\left[V^{\pi}_{\mathcal{M},3}(s_3) - V^{\pi}_{\widehat{\mathcal{M}},3}(s_3)\right] \\
&= \dots \\
&= \sum_{h=1}^{H} \mathbb{E}\left[\mathcal{B}^{\pi_h}_{\mathcal{M}} V^{\pi}_{\widehat{\mathcal{M}},h+1}(s_h) - \mathcal{B}^{\pi_h}_{\widehat{\mathcal{M}}} V^{\pi}_{\widehat{\mathcal{M}},h+1}(s_h)\right] + \underbrace{\mathbb{E}\left[V^{\pi}_{\mathcal{M},H+1}(s_{H+1}) - V^{\pi}_{\widehat{\mathcal{M}},H+1}(s_{H+1})\right]}_{=0} \\
&= \sum_{h=1}^{H} \mathbb{E}\left[\mathcal{B}^{\pi_h}_{\mathcal{M}} V^{\pi}_{\widehat{\mathcal{M}},h+1}(s_h) - \mathcal{B}^{\pi_h}_{\widehat{\mathcal{M}}} V^{\pi}_{\widehat{\mathcal{M}},h+1}(s_h)\right].
\end{aligned}
$$

□

Theorem 6. *Take any $\Pi \supseteq \{\mathcal{S} \to \mathcal{A}\}$, any $\mathcal{V} \supseteq \{V^{\pi} \mid \pi \in \Pi^H\}$, and fix any $D \geq 0$. For each episode $k \in [K]$, let $\widetilde{\mathcal{M}}_k$ be any MDP that achieves the rate-distortion limit of $\mathcal{R}_k^{\Pi,\mathcal{V}}(D)$ with information source $\mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_k)$ and distortion function $d_{\Pi,\mathcal{V}}$. Then,*

$$
\textsc{BayesRegret}(K, \pi^{(1)}, \dots, \pi^{(K)}) \leq \mathbb{E}\left[\sum_{k=1}^{K} \mathbb{E}_k\left[V^{\star}_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1}\right]\right] + 2KH\sqrt{D}.
$$

Proof. By applying definitions from Section 3 and applying the tower property of expectation, we have that

$$
\textsc{BayesRegret}(K, \pi^{(1)}, \dots, \pi^{(K)}) = \mathbb{E}\left[\sum_{k=1}^{K} \mathbb{E}_k\left[\Delta_k\right]\right].
$$

Examining the kth episode in isolation and applying the definition of episodic regret, we have

$$
\begin{aligned}
\mathbb{E}_k\left[\Delta_k\right] &= \mathbb{E}_k\left[V^{\star}_{\mathcal{M}^\star,1} - V^{\pi^{(k)}}_{\mathcal{M}^\star,1}\right] \\
&= \mathbb{E}_k\left[V^{\star}_{\mathcal{M}^\star,1} - V^{\star}_{\widetilde{\mathcal{M}}_k,1} + V^{\star}_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1} + V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\mathcal{M}^\star,1}\right] \\
&= \mathbb{E}_k\left[V^{\star}_{\mathcal{M}^\star,1} - V^{\pi^{\star}_{\mathcal{M}^\star}}_{\widetilde{\mathcal{M}}_k,1} + \underbrace{V^{\pi^{\star}_{\mathcal{M}^\star}}_{\widetilde{\mathcal{M}}_k,1} - V^{\star}_{\widetilde{\mathcal{M}}_k,1}}_{\leq 0} + V^{\star}_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1} + V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\mathcal{M}^\star,1}\right] \\
&\leq \mathbb{E}_k\left[V^{\star}_{\mathcal{M}^\star,1} - V^{\pi^{\star}_{\mathcal{M}^\star}}_{\widetilde{\mathcal{M}}_k,1} + V^{\star}_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1} + V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\mathcal{M}^\star,1}\right]
\end{aligned}
$$

For brevity, we let $\pi_h^\star \triangleq \pi^\star_{\mathcal{M}^\star,h}$ and observe that an application of Lemma 4 yields

$$
\begin{aligned}
\mathbb{E}_k\left[V^\star_{\mathcal{M}^\star,1} - V^{\pi^\star_{\mathcal{M}^\star}}_{\widetilde{\mathcal{M}}_k,1}\right] &= \sum_{h=1}^{H} \mathbb{E}_k\left[\mathcal{B}^{\pi_h^\star}_{\mathcal{M}^\star} V^\star_{\mathcal{M}^\star,h+1}(s_h) - \mathcal{B}^{\pi_h^\star}_{\widetilde{\mathcal{M}}_k} V^\star_{\mathcal{M}^\star,h+1}(s_h)\right] \\
&\leq \sum_{h=1}^{H} \mathbb{E}_k\left[\left|\mathcal{B}^{\pi_h^\star}_{\mathcal{M}^\star} V^\star_{\mathcal{M}^\star,h+1}(s_h) - \mathcal{B}^{\pi_h^\star}_{\widetilde{\mathcal{M}}_k} V^\star_{\mathcal{M}^\star,h+1}(s_h)\right|\right] \\
&= \sum_{h=1}^{H} \mathbb{E}_k\left[\sqrt{\left(\mathcal{B}^{\pi_h^\star}_{\mathcal{M}^\star} V^\star_{\mathcal{M}^\star,h+1}(s_h) - \mathcal{B}^{\pi_h^\star}_{\widetilde{\mathcal{M}}_k} V^\star_{\mathcal{M}^\star,h+1}(s_h)\right)^2}\right] \\
&\leq \sum_{h=1}^{H} \mathbb{E}_k\left[\sqrt{||\mathcal{B}^{\pi_h^\star}_{\mathcal{M}^\star} V^\star_{\mathcal{M}^\star,h+1} - \mathcal{B}^{\pi_h^\star}_{\widetilde{\mathcal{M}}_k} V^\star_{\mathcal{M}^\star,h+1}||_\infty^2}\right] \\
&\leq \sum_{h=1}^{H} \sqrt{\mathbb{E}_k\left[||\mathcal{B}^{\pi_h^\star}_{\mathcal{M}^\star} V^\star_{\mathcal{M}^\star,h+1} - \mathcal{B}^{\pi_h^\star}_{\widetilde{\mathcal{M}}_k} V^\star_{\mathcal{M}^\star,h+1}||_\infty^2\right]} \\
&\leq \sum_{h=1}^{H} \sqrt{\mathbb{E}_k\left[\sup_{\substack{\pi\in\Pi \\ V\in\mathcal{V}}} ||\mathcal{B}^{\pi}_{\mathcal{M}^\star} V - \mathcal{B}^{\pi}_{\widetilde{\mathcal{M}}_k} V||_\infty^2\right]} \\
&= \sum_{h=1}^{H} \sqrt{\mathbb{E}_k\left[d_{\Pi,\mathcal{V}}(\mathcal{M}^\star, \widetilde{\mathcal{M}}_k)\right]} \\
&\leq H\sqrt{D},
\end{aligned}
$$

where the third inequality invokes Jensen's inequality, the fourth inequality holds as $\Pi \supseteq \{\mathcal{S} \to \mathcal{A}\}$ and $\mathcal{V} \supseteq \{V^\pi \mid \pi \in \Pi^H\}$ ensures that $V^\star_{\mathcal{M}^\star,h} \in \mathcal{V}$, $\forall h \in [H]$, and the final inequality holds since $\widetilde{\mathcal{M}}_k$ achieves the rate-distortion limit in the kth episode, by assumption.

We follow the same sequence of steps to obtain

$$
\begin{aligned}
\mathbb{E}_k\left[V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\mathcal{M}^\star,1}\right] &= \sum_{h=1}^{H} \mathbb{E}_k\left[\mathcal{B}^{\pi_h^{(k)}}_{\mathcal{M}^\star} V^{\pi^{(k)}}_{\mathcal{M}^\star,h+1}(s_h) - \mathcal{B}^{\pi_h^{(k)}}_{\widetilde{\mathcal{M}}_k} V^{\pi^{(k)}}_{\mathcal{M}^\star,h+1}(s_h)\right] \\
&\leq \sum_{h=1}^{H} \mathbb{E}_k\left[\left|\mathcal{B}^{\pi_h^{(k)}}_{\mathcal{M}^\star} V^{\pi^{(k)}}_{\mathcal{M}^\star,h+1}(s_h) - \mathcal{B}^{\pi_h^{(k)}}_{\widetilde{\mathcal{M}}_k} V^{\pi^{(k)}}_{\mathcal{M}^\star,h+1}(s_h)\right|\right] \\
&= \sum_{h=1}^{H} \mathbb{E}_k\left[\sqrt{\left(\mathcal{B}^{\pi_h^{(k)}}_{\mathcal{M}^\star} V^{\pi^{(k)}}_{\mathcal{M}^\star,h+1}(s_h) - \mathcal{B}^{\pi_h^{(k)}}_{\widetilde{\mathcal{M}}_k} V^{\pi^{(k)}}_{\mathcal{M}^\star,h+1}(s_h)\right)^2}\right] \\
&\leq \sum_{h=1}^{H} \mathbb{E}_k\left[\sqrt{||\mathcal{B}^{\pi_h^{(k)}}_{\mathcal{M}^\star} V^{\pi^{(k)}}_{\mathcal{M}^\star,h+1} - \mathcal{B}^{\pi_h^{(k)}}_{\widetilde{\mathcal{M}}_k} V^{\pi^{(k)}}_{\mathcal{M}^\star,h+1}||_\infty^2}\right] \\
&\leq \sum_{h=1}^{H} \sqrt{\mathbb{E}_k\left[||\mathcal{B}^{\pi_h^{(k)}}_{\mathcal{M}^\star} V^{\pi^{(k)}}_{\mathcal{M}^\star,h+1} - \mathcal{B}^{\pi_h^{(k)}}_{\widetilde{\mathcal{M}}_k} V^{\pi^{(k)}}_{\mathcal{M}^\star,h+1}||_\infty^2\right]} \\
&\leq \sum_{h=1}^{H} \sqrt{\mathbb{E}_k\left[\sup_{\substack{\pi\in\Pi \\ V\in\mathcal{V}}} ||\mathcal{B}^{\pi}_{\mathcal{M}^\star} V - \mathcal{B}^{\pi}_{\widetilde{\mathcal{M}}_k} V||_\infty^2\right]} \\
&= \sum_{h=1}^{H} \sqrt{\mathbb{E}_k\left[d_{\Pi,\mathcal{V}}(\mathcal{M}^\star, \widetilde{\mathcal{M}}_k)\right]} \\
&\leq H\sqrt{D}.
\end{aligned}
$$

Substituting back into our original expression, we have

$$\begin{aligned}\mathbb{E}_k\left[\Delta_k\right] &= \mathbb{E}_k\left[V^{\star}_{\mathcal{M}^\star,1} - V^{\pi^{(k)}}_{\mathcal{M}^\star,1}\right] \\ &\leq \mathbb{E}_k\left[V^{\star}_{\mathcal{M}^\star,1} - V^{\pi^\star_{\mathcal{M}^\star}}_{\widetilde{\mathcal{M}}_k,1} + V^{\star}_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1} + V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\mathcal{M}^\star,1}\right] \\ &\leq \mathbb{E}_k\left[V^{\star}_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1}\right] + 2H\sqrt{D}.\end{aligned}$$

Applying this upper bound on episodic regret in each episode yields

$$\begin{aligned}\textsc{BayesRegret}(K, \pi^{(1)}, \ldots, \pi^{(K)}) &= \mathbb{E}\left[\sum_{k=1}^{K} \mathbb{E}_k\left[\Delta_k\right]\right] \\ &\leq \mathbb{E}\left[\sum_{k=1}^{K} \mathbb{E}_k\left[V^{\star}_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1}\right]\right] + 2KH\sqrt{D},\end{aligned}$$

as desired. □

D Proof of Lemma 6

In this section, we develop counterparts to the results of Arumugam and Van Roy [15] for the reinforcement-learning setting which relate each rate-distortion function $\mathcal{R}_k^{\Pi,\mathcal{V}}(D)$ to the information accumulated by the agent over the course of learning. Recall that $\tau_k = (s_1^{(k)}, a_1^{(k)}, r_1^{(k)}, \ldots, s_H^{(k)}, a_H^{(k)}, r_H^{(k)}, s_{H+1}^{(k)})$ is a random variable denoting the trajectory experienced by the agent in the kth episode given the history H_k. Let MDP M_k be the MDP sampled in the kth episode.

Lemma 5. *For all $k \in [K]$,*

$$\mathbb{E}_k\left[\mathcal{R}_{k+1}^{\Pi,\mathcal{V}}(D)\right] \leq \mathcal{R}_k^{\Pi,\mathcal{V}}(D) - \mathbb{I}_k(\widetilde{\mathcal{M}}_k; \tau_k \mid M_k).$$

Proof. Recall that, by definition $H_{k+1} = (H_k, \tau_k)$. For all $k \in [K]$, observe that, conditioned on the true MDP $\mathcal{M}^\star$ and sampled MDP M_k which generated the history H_{k+1}, we have that for any compressed MDP $\widetilde{\mathcal{M}}$, $\mathbb{P}(H_{k+1}, \widetilde{\mathcal{M}} \mid \mathcal{M}^\star, M_k) = \mathbb{P}(H_{k+1} \mid \mathcal{M}^\star, M_k)\mathbb{P}(\widetilde{\mathcal{M}} \mid \mathcal{M}^\star, M_k)$. Using this independence $H_{k+1} \perp \widetilde{\mathcal{M}} \mid \mathcal{M}^\star, M_k \; \forall k \in [K]$, we have that

$$0 = \mathbb{I}_k(H_{k+1}; \widetilde{\mathcal{M}} \mid \mathcal{M}^\star, M_k) = \mathbb{I}_k(H_k, \tau_k; \widetilde{\mathcal{M}} \mid \mathcal{M}^\star, M_k) = \mathbb{I}_k(\tau_k; \widetilde{\mathcal{M}} \mid \mathcal{M}^\star, M_k).$$

Moreover, we know that the sampled MDP M_k does not affect our uncertainty in the true MDP $\mathcal{M}^\star$ such that

$$\mathbb{I}_k(\mathcal{M}^\star; \widetilde{\mathcal{M}}) = \mathbb{I}_k(\mathcal{M}^\star; \widetilde{\mathcal{M}} \mid M_k).$$

By the chain rule of mutual information,

$$\mathbb{I}_k(\mathcal{M}^\star; \widetilde{\mathcal{M}}) = \mathbb{I}_k(\mathcal{M}^\star; \widetilde{\mathcal{M}} \mid M_k) = \mathbb{I}_k(\mathcal{M}^\star; \widetilde{\mathcal{M}} \mid M_k) + \mathbb{I}_k(\tau_k; \widetilde{\mathcal{M}} \mid \mathcal{M}^\star, M_k) = \mathbb{I}_k(\mathcal{M}^\star, \tau_k; \widetilde{\mathcal{M}} \mid M_k).$$

Applying the chain rule a second time yields

$$\mathbb{I}_k(\mathcal{M}^\star; \widetilde{\mathcal{M}}) = \mathbb{I}_k(\mathcal{M}^\star, \tau_k; \widetilde{\mathcal{M}} \mid M_k) = \mathbb{I}_k(\widetilde{\mathcal{M}}; \tau_k \mid M_k) + \mathbb{I}_k(\mathcal{M}^\star; \widetilde{\mathcal{M}} \mid \tau_k, M_k).$$

By definition of the rate-distortion function, we have

$$\mathbb{E}_k\left[\mathcal{R}_{k+1}^{\Pi,\mathcal{V}}(D)\right] = \mathbb{E}_k\left[\inf_{\widetilde{\mathcal{M}} \in \Lambda_{k+1}(D)} \mathbb{I}_{k+1}(\mathcal{M}^\star; \widetilde{\mathcal{M}})\right], \qquad \Lambda_{k+1}(D) = \{\widetilde{\mathcal{M}} : \Omega \to \mathfrak{M} \mid \mathbb{E}_{k+1}[d_{\Pi,\mathcal{V}}(\mathcal{M}^\star, \widetilde{\mathcal{M}})] \leq D\}.$$

Recall that, by definition, $\widetilde{\mathcal{M}}_k$ achieves the rate-distortion limit of $\mathcal{R}_k^{\Pi,\mathcal{V}}(D)$, implying that $\mathbb{E}_k[d_{\Pi,\mathcal{V}}(\mathcal{M}^\star, \widetilde{\mathcal{M}}_k)] \leq D$. By the tower property of expectation, we recover that

$$\mathbb{E}_k\left[\mathbb{E}_{k+1}[d_{\Pi,\mathcal{V}}(\mathcal{M}^\star, \widetilde{\mathcal{M}}_k)]\right] = \mathbb{E}_k[d_{\Pi,\mathcal{V}}(\mathcal{M}^\star, \widetilde{\mathcal{M}}_k)] \leq D,$$

and so, in expectation given the current history H_k, $\widetilde{\mathcal{M}}_k \in \Lambda_{k+1}(D)$. Thus, we have that

$$\mathbb{E}_k\left[\mathcal{R}_{k+1}^{\Pi,\mathcal{V}}(D)\right] = \mathbb{E}_k\left[\inf_{\widetilde{\mathcal{M}} \in \Lambda_{k+1}(D)} \mathbb{I}_{k+1}(\mathcal{M}^\star; \widetilde{\mathcal{M}})\right] \leq \mathbb{E}_k\left[\mathbb{I}_{k+1}(\mathcal{M}^\star; \widetilde{\mathcal{M}}_k)\right].$$

Re-arranging terms from our previous chain rule expansions, we may expand the integrand as

$$\begin{aligned}
\mathbb{E}_k\left[\mathbb{I}_{k+1}(\mathcal{M}^\star; \widetilde{\mathcal{M}}_k)\right] &= \mathbb{E}_k\left[\mathbb{I}_k(\mathcal{M}^\star; \widetilde{\mathcal{M}}_k \mid \tau_k, M_k)\right] \\
&= \mathbb{E}_k\left[\mathbb{I}_k(\mathcal{M}^\star; \widetilde{\mathcal{M}}_k) - \mathbb{I}_k(\widetilde{\mathcal{M}}_k, \tau_k \mid M_k)\right] \\
&= \mathbb{I}_k(\mathcal{M}^\star; \widetilde{\mathcal{M}}_k) - \mathbb{I}_k(\widetilde{\mathcal{M}}_k, \tau_k \mid M_k) \\
&= \mathcal{R}_k^{\Pi,\mathcal{V}}(D) - \mathbb{I}_k(\widetilde{\mathcal{M}}_k; \tau_k \mid M_k),
\end{aligned}$$

where the penultimate line follows since both mutual information terms are $\sigma(H_k)$-measurable and the final line follows by definition of $\widetilde{\mathcal{M}}_k$. □

At the beginning of each episode, our generalization of PSRL will identify a compressed MDP $\widetilde{\mathcal{M}}_k$ that achieves the rate-distortion limit based on the current history H_k. As data accumulates and the agent's knowledge of the true MDP is refined, this satisficing MDP $\widetilde{\mathcal{M}}_k$ will be recomputed to reflect that updated knowledge. The previous lemma shows that the expected number of bits the agent must identify to learn this new target MDP decreases as this adaptation occurs, highlighting two possible sources of improvement: (1) shifting from a compressed MDP $\widetilde{\mathcal{M}}_k$ to $\widetilde{\mathcal{M}}_{k+1}$ and (2) a decrease of $\mathbb{I}_k(\widetilde{\mathcal{M}}; \tau_k \mid M_k)$ that occurs from observing the trajectory τ_k. The former reflects the agent's improved ability in synthesizing an approximately value-equivalent MDP to pursue instead of $\mathcal{M}^\star$ while the latter captures information gained about the previous target $\widetilde{\mathcal{M}}_k$ from the experienced trajectory τ_k.

Fact 1 ([41]). *For any $\Pi, \mathcal{V}$ and all $k \in [K]$, $\mathcal{R}_k^{\Pi,\mathcal{V}}(D)$ is a non-negative, convex, and monotonically-decreasing function in D.*

Let $\widetilde{\mathcal{M}}$ be a compressed MDP that is exactly value-equivalent to $\mathcal{M}^\star$ which, by definition, implies a distortion of exactly zero. Further recall that $\mathcal{M}^\star$ is itself a MDP that achieves zero distortion, albeit one that has no guarantee of achieving the rate-distortion limit. Fact 1 yields the following chain of inequalities that hold for all $k \in [K]$ and $D \geq 0$:

$$\mathcal{R}_k^{\Pi,\mathcal{V}}(D) \leq \mathbb{I}_k(\mathcal{M}^\star; \widetilde{\mathcal{M}}) \leq \mathbb{I}_k(\mathcal{M}^\star; \mathcal{M}^\star) = \mathbb{H}_k(\mathcal{M}^\star).$$

This chain of inequalities confirms an important goal of satisficing in PSRL; namely, that the compressed MDP an agent attempts to solve in each episode $k \in [K]$, $\widetilde{\mathcal{M}}_k$, requires fewer bits of information than what is needed to fully identify the true MDP $\mathcal{M}^\star$. This gives rise to the following corollary:

Corollary 5. *For any $k \in [K]$,*

$$\mathbb{E}_k\left[\sum_{k'=k}^{K} \mathbb{I}_{k'}(\widetilde{\mathcal{M}}_{k'}; \tau_{k'} \mid M_{k'})\right] \leq \mathbb{H}_k(\mathcal{M}^\star).$$

Instead of proving this corollary, we prove the following lemma which yields the corollary through Fact 1:

Lemma 6. *For any $k \in [K]$,*

$$\mathbb{E}_k\left[\sum_{k'=k}^{K} \mathbb{I}_{k'}(\widetilde{\mathcal{M}}_{k'}; \tau_{k'} \mid M_{k'})\right] \leq \mathcal{R}_k^{\Pi,\mathcal{V}}(D).$$

Proof. Observe that by Lemma 5, for all $k \in [K]$,

$$\mathbb{I}_k(\widetilde{\mathcal{M}}_k; \tau_k \mid M_k) \leq \mathcal{R}_k^{\Pi,\mathcal{V}}(D) - \mathbb{E}_k\left[\mathcal{R}_{k+1}^{\Pi,\mathcal{V}}(D)\right].$$

Directly substituting in, we have

$$\mathbb{E}_k\left[\sum_{k'=k}^{K} \mathbb{I}_{k'}(\widetilde{\mathcal{M}}_{k'}; \tau_{k'} \mid M_{k'})\right] \leq \mathbb{E}_k\left[\sum_{k'=k}^{K}\left(\mathcal{R}_{k'}^{\Pi,\mathcal{V}}(D) - \mathbb{E}_{k'}\left[\mathcal{R}_{k'+1}^{\Pi,\mathcal{V}}(D)\right]\right)\right].$$

Applying linearity of expectation and breaking apart the sum yields

$$\mathbb{E}_k\left[\sum_{k'=k}^{K} \mathbb{I}_{k'}(\widetilde{\mathcal{M}}_{k'}; \tau_{k'} \mid M_{k'})\right] \leq \sum_{k'=k}^{K} \mathbb{E}_k\left[\mathcal{R}_{k'}^{\Pi,\mathcal{V}}(D)\right] - \sum_{k'=k}^{K} \mathbb{E}_k\left[\mathbb{E}_{k'}\left[\mathcal{R}_{k'+1}^{\Pi,\mathcal{V}}(D)\right]\right].$$

Note that the first term can simply be separated into

$$\sum_{k'=k}^{K} \mathbb{E}_k\left[\mathcal{R}_{k'}^{\Pi,\mathcal{V}}(D)\right] = \mathbb{E}_k\left[\mathcal{R}_{k}^{\Pi,\mathcal{V}}(D)\right] + \sum_{k'=k+1}^{K} \mathbb{E}_k\left[\mathcal{R}_{k'}^{\Pi,\mathcal{V}}(D)\right] = \mathcal{R}_{k}^{\Pi,\mathcal{V}}(D) + \sum_{k'=k+1}^{K} \mathbb{E}_k\left[\mathcal{R}_{k'}^{\Pi,\mathcal{V}}(D)\right].$$

Meanwhile, since $\sigma(H_k) \subseteq \sigma(H_{k'})$, the tower property of expectation yields

$$\sum_{k'=k}^{K} \mathbb{E}_k\left[\mathbb{E}_{k'}\left[\mathcal{R}_{k'+1}^{\Pi,\mathcal{V}}(D)\right]\right] = \sum_{k'=k}^{K} \mathbb{E}_k\left[\mathcal{R}_{k'+1}^{\Pi,\mathcal{V}}(D)\right] = \sum_{k'=k+1}^{K} \mathbb{E}_k\left[\mathcal{R}_{k'}^{\Pi,\mathcal{V}}(D)\right].$$

Combining the expansions results in

$$\begin{aligned}
\mathbb{E}_k\left[\sum_{k'=k}^{K} \mathbb{I}_{k'}(\widetilde{\mathcal{M}}_{k'}; \tau_{k'} \mid M_{k'})\right] &\leq \sum_{k'=k}^{K} \mathbb{E}_k\left[\mathcal{R}_{k'}^{\Pi,\mathcal{V}}(D)\right] - \sum_{k'=k}^{K} \mathbb{E}_k\left[\mathbb{E}_{k'}\left[\mathcal{R}_{k'+1}^{\Pi,\mathcal{V}}(D)\right]\right] \\
&= \mathcal{R}_{k}^{\Pi,\mathcal{V}}(D) + \sum_{k'=k+1}^{K} \mathbb{E}_k\left[\mathcal{R}_{k'}^{\Pi,\mathcal{V}}(D)\right] - \sum_{k'=k+1}^{K} \mathbb{E}_k\left[\mathcal{R}_{k'}^{\Pi,\mathcal{V}}(D)\right] \\
&= \mathcal{R}_{k}^{\Pi,\mathcal{V}}(D).
\end{aligned}$$

□

E Proof of Theorem 2

In this section, we prove a general, information-theoretic satisficing Bayesian regret bound. Central to our analysis is the information ratio in the kth episode:

$$\Gamma_k \triangleq \frac{\mathbb{E}_k\left[V^{\star}_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1}\right]^2}{\mathbb{I}_k(\widetilde{\mathcal{M}}_k; \tau_k, M_k)}, \qquad \forall k \in [K].$$

Theorem 7 (Information-Theoretic Satisficing Regret Bound)**.** *If $\Gamma_k \leq \overline{\Gamma}$, for all $k \in [K]$, then*

$$\mathbb{E}_k\left[\sum_{k=1}^{K} \mathbb{E}\left[V^{\star}_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1}\right]\right] \leq \sqrt{\overline{\Gamma} K \mathcal{R}_1^{\Pi,\mathcal{V}}(D)}.$$

Proof. The definition of the information ratio Γ_k for each term in the sum followed by the fact that $\Gamma_k \leq \overline{\Gamma}, \forall k \in [K]$ yields

$$\mathbb{E}\left[\sum_{k=1}^{K} \mathbb{E}_k\left[V^{\star}_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1}\right]\right] = \mathbb{E}\left[\sum_{k=1}^{K} \sqrt{\Gamma_k \mathbb{I}_k(\widetilde{\mathcal{M}}_k; \tau_k, M_k)}\right] \leq \sqrt{\overline{\Gamma}}\mathbb{E}\left[\sum_{k=1}^{K} \sqrt{\mathbb{I}_k(\widetilde{\mathcal{M}}_k; \tau_k, M_k)}\right].$$

Applying the tower property of expectation and Jensen's inequality in sequence yields

$$\sqrt{\overline{\Gamma}}\mathbb{E}\left[\sum_{k=1}^{K} \sqrt{\mathbb{I}_k(\widetilde{\mathcal{M}}_k; \tau_k, M_k)}\right] \leq \sqrt{\overline{\Gamma}}\mathbb{E}\left[\sum_{k=1}^{K} \sqrt{\mathbb{E}_k\left[\mathbb{I}_k(\widetilde{\mathcal{M}}_k; \tau_k, M_k)\right]}\right].$$

By the Cauchy-Schwarz inequality, we have that

$$\sqrt{\overline{\Gamma}}\mathbb{E}\left[\sum_{k=1}^{K}\sqrt{\mathbb{E}_k\left[\mathbb{I}_k(\widetilde{\mathcal{M}}_k;\tau_k,M_k)\right]}\right] \leq \sqrt{\overline{\Gamma}}\mathbb{E}\left[\sqrt{K\sum_{k=1}^{K}\mathbb{E}_k\left[\mathbb{I}_k(\widetilde{\mathcal{M}}_k;\tau_k,M_k)\right]}\right].$$

Recall that the sampled M_k by itself offers no information about $\widetilde{\mathcal{M}}_k$. Consequently, by the chain rule of mutual information, we have

$$\mathbb{I}_k(\widetilde{\mathcal{M}}_k;\tau_k,M_k) = \mathbb{I}_k(\widetilde{\mathcal{M}}_k;M_k) + \mathbb{I}_k(\widetilde{\mathcal{M}}_k;\tau_k \mid M_k) = \mathbb{I}_k(\widetilde{\mathcal{M}}_k;\tau_k \mid M_k).$$

Therefore,

$$\sqrt{\overline{\Gamma}}\mathbb{E}\left[\sqrt{K\sum_{k=1}^{K}\mathbb{E}_k\left[\mathbb{I}_k(\widetilde{\mathcal{M}}_k;\tau_k,M_k)\right]}\right] = \sqrt{\overline{\Gamma}}\mathbb{E}\left[\sqrt{K\sum_{k=1}^{K}\mathbb{E}_k\left[\mathbb{I}_k(\widetilde{\mathcal{M}}_k;\tau_k \mid M_k)\right]}\right].$$

Directly applying Lemma 6 followed by Jensen's inequality yields

$$\sqrt{\overline{\Gamma}}\mathbb{E}\left[\sqrt{K\sum_{k=1}^{K}\mathbb{E}_k\left[\mathbb{I}_k(\widetilde{\mathcal{M}}_k;\tau_k \mid M_k)\right]}\right] \leq \sqrt{\overline{\Gamma}}\mathbb{E}\left[\sqrt{K\mathcal{R}_1^{\Pi,\mathcal{V}}(D)}\right] \leq \sqrt{\overline{\Gamma}K\mathbb{E}\left[\mathcal{R}_1^{\Pi,\mathcal{V}}(D)\right]}.$$

Since the expectation is with respect to the prior $\mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_1)$ and $\mathcal{R}_1^{\Pi,\mathcal{V}}(D)$ is $\sigma(H_1)$-measurable, we have

$$\sqrt{\overline{\Gamma}K\mathbb{E}\left[\mathcal{R}_1^{\Pi,\mathcal{V}}(D)\right]} = \sqrt{\overline{\Gamma}K\mathcal{R}_1^{\Pi,\mathcal{V}}(D)},$$

as desired. □

F Proof of Lemma 1

In this section, we clarify how the shrinkage or growth of the policy class Π and value function class $\mathcal{V}$ affect the rate-distortion function at the kth episode, $\mathcal{R}_k^{\Pi,\mathcal{V}}(D)$.

Lemma 7 (Dominance with Approximate Value Equivalence)**.** *For any two Π, Π' and any $\mathcal{V}, \mathcal{V}'$ such that $\Pi' \subseteq \Pi \subseteq \{\mathcal{S} \to \Delta(\mathcal{A})\}$ and $\mathcal{V}' \subseteq \mathcal{V} \subseteq \{\mathcal{S} \to \mathbb{R}\}$, we have*

$$\mathcal{R}_k^{\Pi,\mathcal{V}}(D) \geq \mathcal{R}_k^{\Pi',\mathcal{V}'}(D), \qquad \forall k \in [K], D > 0.$$

Proof. Recall that the distortion function $d : \mathfrak{M} \times \mathfrak{M} \to \mathbb{R}_{\geq 0}$ with respect to policy class Π and value function class $\mathcal{V}$ is given by

$$d_{\Pi,\mathcal{V}}(\mathcal{M},\widehat{\mathcal{M}}) = \sup_{\substack{\pi\in\Pi\\ V\in\mathcal{V}}} ||\mathcal{B}_{\mathcal{M}}^{\pi}V - \mathcal{B}_{\widehat{\mathcal{M}}}^{\pi}V||_\infty^2 = \sup_{\substack{\pi\in\Pi\\ V\in\mathcal{V}}} \left(\max_{s\in\mathcal{S}} |\mathcal{B}_{\mathcal{M}}^{\pi}V(s) - \mathcal{B}_{\widehat{\mathcal{M}}}^{\pi}V(s)|\right)^2,$$

with an analogous definition holding for the distortion function $d_{\Pi',\mathcal{V}'}$ under Π' and $\mathcal{V}'$. In the parlance of Stjernvall [150], we have that $d_{\Pi,\mathcal{V}}$ dominates $d_{\Pi',\mathcal{V}'}$ if for all source distributions $\mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_k)$ and all distortion thresholds $D > 0$,

$$\mathcal{R}_k^{\Pi,\mathcal{V}}(D) \geq \mathcal{R}_k^{\Pi',\mathcal{V}'}(D).$$

In words, a distortion function d_1 that dominates another distortion function d_2 requires more bits of information in order to achieve the rate-distortion limit for all information sources and at all distortion thresholds. From this definition, it is clear that statement of the theorem holds if we can establish a dominance relationship between $d_{\Pi,\mathcal{V}}$ and $d_{\Pi',\mathcal{V}'}$.

Recognizing the significant amount of calculation needed to exhaustively verify a dominance relationship by hand, Stjernvall [150] prescribes six sufficient conditions for establishing dominance (with varying degrees of strength) between distortion functions; we will leverage the second of these characterizations (C2).

Fix an arbitrary source distribution $\mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_k)$ and distortion threshold $D > 0$. We denote by $\widetilde{\mathcal{M}}_k$ the MDP that achieves the rate-distortion limit $\mathcal{R}_k^{\Pi,\mathcal{V}}(D)$ under our chosen source, distortion threshold, and distortion function $d_{\Pi,\mathcal{V}}$. By definition of the supremum, we have that for any two MDPs $\mathcal{M}, \widehat{\mathcal{M}}$

$$d_{\Pi',\mathcal{V}'}(\mathcal{M}, \widehat{\mathcal{M}}) = \sup_{\substack{\pi \in \Pi' \\ V \in \mathcal{V}'}} ||\mathcal{B}_{\mathcal{M}}^\pi V - \mathcal{B}_{\widehat{\mathcal{M}}}^\pi V||_\infty^2 \leq \sup_{\substack{\pi \in \Pi \\ V \in \mathcal{V}}} ||\mathcal{B}_{\mathcal{M}}^\pi V - \mathcal{B}_{\widehat{\mathcal{M}}}^\pi V||_\infty^2 = d_{\Pi,\mathcal{V}}(\mathcal{M}, \widehat{\mathcal{M}}).$$

Consequently, since $\widetilde{\mathcal{M}}_k$ achieves the rate-distortion limit, we have

$$\mathbb{E}_k \left[d_{\Pi',\mathcal{V}'}(\mathcal{M}^\star, \widetilde{\mathcal{M}}_k) \right] \leq \mathbb{E}_k \left[d_{\Pi,\mathcal{V}}(\mathcal{M}^\star, \widetilde{\mathcal{M}}_k) \right] \leq D.$$

Observe that, since our information source and distortion threshold were arbitrary, we have that for all sources $\mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_k)$ and all thresholds $D > 0$ with $\widetilde{\mathcal{M}}_k$ achieving the rate-distortion limit under distortion $d_{\Pi,\mathcal{V}}$, there exists a Markov chain $\mathcal{M}^\star - \widetilde{\mathcal{M}}_k - \widetilde{\mathcal{M}}'_k$ such that $\widetilde{\mathcal{M}}_k = \widetilde{\mathcal{M}}'_k$ (the mapping between them is the identity function) and $\mathbb{E}\left[d_{\Pi',\mathcal{V}'}(\mathcal{M}^\star, \widetilde{\mathcal{M}}'_k) \right] \leq D$. Thus, by Theorem 2 of Stjernvall [150] (specifically, C2 $\implies$ D4), we have that $d_{\Pi,\mathcal{V}}$ dominates $d_{\Pi',\mathcal{V}'}$ for any $\Pi' \subseteq \Pi \subseteq \{\mathcal{S} \to \Delta(\mathcal{A})\}$ and $\mathcal{V}' \subseteq \mathcal{V} \subseteq \{\mathcal{S} \to \mathbb{R}\}$. As previously discussed, the claim of the theorem follows as an immediate consequence, by definition of dominance. □

G Proof of Lemma 2

Fano's inequality [64] is a key result in information theory that relates conditional entropy to the probability of error in a discrete, multi-way hypothesis testing problem. The traditional form of the result, however, determines an error as the inability to exactly recover the random variable being estimated. Naturally, given the lossy compression context of this work, a more useful analysis will use a lack of adherence to the distortion upper bound as the more appropriate notion of error. For this purpose, we require a more general result of the same flavor as those developed by Duchi and Wainwright [60]; in particular, we leverage an extension of their generalized Fano's inequality which is given as Question 7.1 in [59], whose proof we provide and adapt to our setting for completeness. We first require the following lemma:

Lemma 8. *Let P and Q be two arbitrary probability measures on the same measurable space such that $P \ll Q$. Then,*

$$D_{\mathrm{KL}}(P \,||\, Q) \geq \log\left(\frac{1}{Q(P > 0)}\right) = \log\left(\frac{1}{Q(\mathit{supp}\,(P))}\right).$$

Proof. The proof is immediate via a generalization of the traditional log-sum inequality [41]. Specifically, since $P \ll Q$, we have

$$D_{\mathrm{KL}}(P \,||\, Q) = \int \log\left(\frac{dP}{dQ}\right) dP = \int\limits_{P>0} \log\left(\frac{dP}{dQ}\right) dP \geq \left(\int dP\right) \log\left(\frac{\int dP}{\int\limits_{P>0} dQ}\right) = \log\left(\frac{1}{Q(P>0)}\right).$$

□

Theorem 8. *Take any $\Pi \subseteq \{\mathcal{S} \to \Delta(\mathcal{A})\}$ and $\mathcal{V} \subseteq \{\mathcal{S} \to \mathbb{R}\}$. For any $D \geq 0$ and any $k \in [K]$, define $\delta = \sup\limits_{\widehat{M} \in \mathfrak{M}} \mathbb{P}(d_{\Pi,\mathcal{V}}(\mathcal{M}^\star, \widehat{M}) \leq D \mid H_k)$. Then,*

$$\sup_{\widetilde{\mathcal{M}} \in \Lambda_k(D)} \mathbb{P}(d_{\Pi,\mathcal{V}}(\mathcal{M}^\star, \widetilde{\mathcal{M}}) > D \mid H_k) \geq 1 - \frac{\mathcal{R}_k^{\Pi,\mathcal{V}}(D) + \log(2)}{\log\left(\frac{1}{\delta}\right)}.$$

Proof. For any episode $k \in [K]$, recall that the agent's beliefs over the true MDP $\mathcal{M}^\star$ are distributed according to $\mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_k)$. Let $\widetilde{\mathcal{M}}$ be an arbitrary random variable denoting a compressed MDP

taking values in the set $\mathfrak{M}$ and, for a fixed distortion threshold D, we let $\mathcal{N} \subset \mathfrak{M} \times \mathfrak{M}$ denote the measurable subset of $\mathfrak{M} \times \mathfrak{M}$ that consists of all pairs of MDP which are approximately value equivalent; that is, $(M, \widehat{M}) \in \mathcal{N} \iff d_{\Pi,\mathcal{V}}(M, \widehat{M}) \leq D$. For any MDP $\widehat{M} \in \mathfrak{M}$, we define a slice

$$\mathcal{N}_{\widehat{M}} \triangleq \{M \in \mathfrak{M} \mid (M, \widehat{M}) \in \mathcal{N}\},$$

as the collection of MDPs that are approximately value equivalent to a given $\widehat{M}$. In the context of Fano's inequality and our lossy compression problem, $\mathcal{N}_{\widehat{M}}$ is the set of original or uncompressed MDPs for which a channel output of $\widehat{M}$ should not be considered an error. Furthermore, define

$$p^{\max} \triangleq \sup_{\widehat{M} \in \mathfrak{M}} \mathbb{P}(\mathcal{M}^\star \in \mathcal{N}_{\widehat{M}} \mid H_k) \qquad p^{\min} \triangleq \inf_{\widehat{M} \in \mathfrak{M}} \mathbb{P}(\mathcal{M}^\star \in \mathcal{N}_{\widehat{M}} \mid H_k).$$

Recall that for $p \in [0,1]$, we have the binary entropy function $h_2(p) = -p\log(p) - (1-p)\log(1-p)$.

Define the indicator random variable $E = \mathbb{1}((\mathcal{M}^\star, \widetilde{\mathcal{M}}) \notin \mathcal{N})$. Recalling that

$$\mathbb{I}(X;Y) = \mathbb{E}\left[D_{\mathrm{KL}}(\mathbb{P}(Y \in \cdot \mid X) \,||\, \mathbb{P}(Y \in \cdot))\right],$$

we have

$$\begin{aligned}
&\mathbb{I}_k(\mathcal{M}^\star; (\widetilde{\mathcal{M}}, E))\mathbb{E}\left[D_{\mathrm{KL}}(\mathbb{P}_k(\mathcal{M}^\star \in \cdot \mid \widetilde{\mathcal{M}}, E) \,||\, \mathbb{P}_k(\mathcal{M}^\star \in \cdot))\right] \\
&\qquad = \mathbb{P}_k(E=1) \cdot \mathbb{E}\left[D_{\mathrm{KL}}(\mathbb{P}_k(\mathcal{M}^\star \in \cdot \mid \widetilde{\mathcal{M}}, E=1) \,||\, \mathbb{P}_k(\mathcal{M}^\star \in \cdot))\right] \\
&\qquad\qquad + \mathbb{P}_k(E=0) \cdot \mathbb{E}\left[D_{\mathrm{KL}}(\mathbb{P}_k(\mathcal{M}^\star \in \cdot \mid \widetilde{\mathcal{M}}, E=0) \,||\, \mathbb{P}_k(\mathcal{M}^\star \in \cdot))\right].
\end{aligned}$$

At this point, we observe that for any $\widehat{M} \in \mathfrak{M}$,

$$\operatorname{supp}\left(\mathbb{P}_k(\mathcal{M}^\star \in \cdot \mid \widetilde{\mathcal{M}} = \widehat{M}, E=0)\right) \subset \mathcal{N}_{\widehat{M}} \qquad \operatorname{supp}\left(\mathbb{P}_k(\mathcal{M}^\star \in \cdot \mid \widetilde{\mathcal{M}} = \widehat{M}, E=1)\right) \subset \mathcal{N}^c_{\widehat{M}},$$

by definition of the slice $\mathcal{N}_{\widehat{M}}$. Thus,

$$\mathbb{P}(\mathcal{M}^\star \in \operatorname{supp}\left(\mathbb{P}_k(\mathcal{M}^\star \in \cdot \mid \widetilde{\mathcal{M}} = \widehat{M}, E=0)\right) \mid H_k) \leq \mathbb{P}(\mathcal{M}^\star \in \mathcal{N}_{\widehat{M}} \mid H_k)$$

$$\mathbb{P}(\mathcal{M}^\star \in \operatorname{supp}\left(\mathbb{P}_k(\mathcal{M}^\star \in \cdot \mid \widetilde{\mathcal{M}} = \widehat{M}, E=1)\right) \mid H_k) \leq \mathbb{P}(\mathcal{M}^\star \in \mathcal{N}^c_{\widehat{M}} \mid H_k) = 1 - \mathbb{P}(\mathcal{M}^\star \in \mathcal{N}_{\widehat{M}} \mid H_k)$$

and, consequently, we have by Lemma 8 that

$$D_{\mathrm{KL}}(\mathbb{P}_k(\mathcal{M}^\star \in \cdot \mid \widetilde{\mathcal{M}} = \widehat{M}, E=0) \,||\, \mathbb{P}_k(\mathcal{M}^\star \in \cdot)) \geq \log\left(\frac{1}{\mathbb{P}(\mathcal{M}^\star \in \mathcal{N}_{\widehat{M}} \mid H_k)}\right) \geq \log\left(\frac{1}{p^{\max}}\right),$$

$$D_{\mathrm{KL}}(\mathbb{P}_k(\mathcal{M}^\star \in \cdot \mid \widetilde{\mathcal{M}} = \widehat{M}, E=1) \,||\, \mathbb{P}_k(\mathcal{M}^\star \in \cdot)) \geq \log\left(\frac{1}{1 - \mathbb{P}(\mathcal{M}^\star \in \mathcal{N}_{\widehat{M}} \mid H_k)}\right) \geq \log\left(\frac{1}{1 - p^{\min}}\right).$$

Applying these lower bounds to our original mutual information term, we see that

$$\begin{aligned}
\mathbb{I}_k(\mathcal{M}^\star; (\widetilde{\mathcal{M}}, E)) &\geq \mathbb{P}(E=1 \mid H_k)\log\left(\frac{1}{1-p^{\min}}\right) + \mathbb{P}(E=0 \mid H_k)\log\left(\frac{1}{p^{\max}}\right) \\
&= \mathbb{P}(E=1 \mid H_k)\log\left(\frac{1}{1-p^{\min}}\right) + (1 - \mathbb{P}(E=1 \mid H_k))\log\left(\frac{1}{p^{\max}}\right) \\
&= \mathbb{P}(E=1 \mid H_k)\log\left(\frac{p^{\max}}{1-p^{\min}}\right) + \log\left(\frac{1}{p^{\max}}\right).
\end{aligned}$$

Now applying the chain rule of mutual information, the definition of mutual information, the non-negativity of entropy and the fact that conditioning reduces entropy in sequence, we obtain

$$\begin{aligned}
\mathbb{I}_k(\mathcal{M}^\star; (\widetilde{\mathcal{M}}, E)) &= \mathbb{I}_k(\mathcal{M}^\star; \widetilde{\mathcal{M}}) + \mathbb{I}_k(\mathcal{M}^\star; E \mid \widetilde{\mathcal{M}}) \\
&= \mathbb{I}_k(\mathcal{M}^\star; \widetilde{\mathcal{M}}) + \mathbb{H}_k(E \mid \widetilde{\mathcal{M}}) - \mathbb{H}_k(E \mid \widetilde{\mathcal{M}}, \mathcal{M}^\star) \\
&\leq \mathbb{I}_k(\mathcal{M}^\star; \widetilde{\mathcal{M}}) + \mathbb{H}_k(E \mid \widetilde{\mathcal{M}}) \\
&\leq \mathbb{I}_k(\mathcal{M}^\star; \widetilde{\mathcal{M}}) + \mathbb{H}_k(E) \\
&\leq \mathbb{I}_k(\mathcal{M}^\star; \widetilde{\mathcal{M}}) + \mathbb{H}(E)
\end{aligned}$$

Combining the upper and lower bounds while multiplying through by -1 yields

$$h_2(\mathbb{P}(E=1)) + \mathbb{P}(E=1 \mid H_k)\log\left(\frac{1-p^{\min}}{p^{\max}}\right) \geq \log\left(\frac{1}{p^{\max}}\right) - \mathbb{I}_k(\mathcal{M}^\star; \widetilde{\mathcal{M}}).$$

Recognizing that we have the following upper bounds

$$\begin{aligned}\log(2) + \mathbb{P}(E=1 \mid H_k)\log\left(\frac{1}{p^{\max}}\right) &\geq h_2(\mathbb{P}(E=1)) + \mathbb{P}(E=1 \mid H_k)\log\left(\frac{1}{p^{\max}}\right) \\ &\geq h_2(\mathbb{P}(E=1)) + \mathbb{P}(E=1 \mid H_k)\log\left(\frac{1-p^{\min}}{p^{\max}}\right),\end{aligned}$$

and re-arranging terms yields

$$\mathbb{P}(E=1 \mid H_k) \geq \frac{\log\left(\frac{1}{p^{\max}}\right) - \mathbb{I}_k(\mathcal{M}^\star; \widetilde{\mathcal{M}}) - \log(2)}{\log\left(\frac{1}{p^{\max}}\right)} = 1 - \frac{\mathbb{I}_k(\mathcal{M}^\star; \widetilde{\mathcal{M}}) + \log(2)}{\log\left(\frac{1}{\delta}\right)},$$

where $\delta = \sup_{\widehat{M} \in \mathfrak{M}} \mathbb{P}(d_{\Pi,\mathcal{V}}(\mathcal{M}^\star, \widehat{M}) \leq D \mid H_k)$. Noting that

$$\mathbb{P}(E=1 \mid H_k) = \mathbb{P}((\mathcal{M}^\star, \widetilde{\mathcal{M}}) \notin \mathcal{N} \mid H_k) = \mathbb{P}(d_{\Pi,\mathcal{V}}(\mathcal{M}^\star, \widetilde{\mathcal{M}}) > D \mid H_k),$$

and taking the supremum on both sides, we have

$$\begin{aligned}\sup_{\widetilde{\mathcal{M}} \in \Lambda_k(D)} \mathbb{P}(d_{\Pi,\mathcal{V}}(\mathcal{M}^\star, \widetilde{\mathcal{M}}) > D \mid H_k) &\geq \sup_{\widetilde{\mathcal{M}} \in \Lambda_k(D)} \left[1 - \frac{\mathbb{I}_k(\mathcal{M}^\star; \widetilde{\mathcal{M}}) + \log(2)}{\log\left(\frac{1}{\delta}\right)}\right] \\ &= 1 - \inf_{\widetilde{\mathcal{M}} \in \Lambda_k(D)} \frac{\mathbb{I}_k(\mathcal{M}^\star; \widetilde{\mathcal{M}}) + \log(2)}{\log\left(\frac{1}{\delta}\right)} \\ &= 1 - \frac{\mathcal{R}_k^{\Pi,\mathcal{V}}(D) + \log(2)}{\log\left(\frac{1}{\delta}\right)},\end{aligned}$$

as desired. □

H Proof of Theorem 3

In specializing to the tabular MDP setting, we wish to simplify our information-theoretic Bayesian regret bound (Corollary 1) into one that only depends on the standard problem-specific quantities ($|\mathcal{S}|$, $|\mathcal{A}|$, K, H). To do this, we will necessarily decompose mutual information into its constituent entropy terms. Inconveniently, while mutual information is well-defined for arbitrary random variables, entropy is infinite for continuous random variables (like the reward function and transition function random variables, $\mathcal{R}^\star$ and $\mathcal{T}^\star$). Rather than resorting to differential entropy, which lacks several desirable properties of Shannon entropy, we explicitly replace these random variables by their discretized analogues, obtained via a sufficiently-fine quantization of their ranges a priori such that the differential entropy of the original random variables is well-approximated by the associated metric entropy or ε-entropy [93], courtesy of Theorem 8.3.1 of [41].

Recall that, for any $\varepsilon > 0$, a ε-cover of a set Θ with respect to a (semi)-metric $\rho : \Theta \times \Theta \to \mathbb{R}_{\geq 0}$ is a set $\{\theta_1, \ldots, \theta_N\}$ with $\theta_i \in \Theta$, $\forall i \in [N]$, such that for any other point $\theta \in \Theta$, $\exists\, n \in [N]$ such that $\rho(\theta, \theta_n) \leq \varepsilon$. The ε-covering number of Θ is defined as

$$\mathcal{N}(\varepsilon, \Theta, \rho) \triangleq \inf\{N \in \mathbb{N} \mid \exists \text{ an } \varepsilon\text{-cover } \{\theta_1, \ldots, \theta_N\} \text{ of } \Theta\}.$$

Conversely, a ε-packing of a set Θ with respect to ρ is a set $\{\theta_1, \ldots, \theta_M\}$ with $\theta_i \in \Theta$, $\forall i \in [M]$, such that for any distinct $i, j \in [N]$, we have $\rho(\theta_i, \theta_j) \geq \varepsilon$. The ε-packing number of a set Θ is defined as

$$\mathcal{M}(\varepsilon, \Theta, \rho) \triangleq \sup\{M \in \mathbb{N} \mid \exists \text{ an } \varepsilon\text{-packing } \{\theta_1, \ldots, \theta_M\} \text{ of } \Theta\}.$$

With slight abuse of notation, for any norm $||\cdot||$ on a set Θ, we write $\mathcal{N}(\varepsilon, \Theta, ||\cdot||)$ to denote the ε-covering number under the metric induced by $||\cdot||$, and similarly for the ε-packing number $\mathcal{M}(\varepsilon, \Theta, ||\cdot||)$. Theorem IV of [93] establishes the following relationship between the ε-covering number and ε-packing number that we will use to upper bound metric entropy:

Fact 2. *For any metric space (Θ, ρ) and any $\varepsilon > 0$, $\mathcal{N}(\varepsilon, \Theta, \rho) \leq \mathcal{M}(\varepsilon, \Theta, \rho)$.*

This allows for a generalization of Lemma 7.6 of [59] to norm balls of arbitrary radius whose proof we include for completeness.

Lemma 9. *For any norm $||\cdot||$, let $\mathbb{B}^d = \{\theta \in \mathbb{R}^d \mid ||\theta|| \leq 1\}$ denote the unit $||\cdot||$-ball in $\mathbb{R}^d$. For any $r \in (0, \infty)$, we let $r\mathbb{B}^d = \{\theta \in \mathbb{R}^d \mid ||\theta|| \leq r\}$ denote the scaling of the unit ball by r or, equivalently, the $||\cdot||$-ball of radius r. Then, for any $\varepsilon \in (0, r]$,*

$$\log\left(\mathcal{N}(\varepsilon, r\mathbb{B}^d, ||\cdot||)\right) \leq d \log\left(1 + \frac{2r}{\varepsilon}\right).$$

Proof. Let $\text{Vol}(\cdot)$ be the function that denotes the volume of an input ball in $\mathbb{R}^d$ such that $\text{Vol}\left(r\mathbb{B}^d\right) = r^d$. Since an ε-packing requires filling $r\mathbb{B}^d$ with disjoint balls of diameter ε, we have

$$\mathcal{M}(\varepsilon, r\mathbb{B}^d, ||\cdot||)\text{Vol}\left(\frac{\varepsilon}{2}\mathbb{B}^d\right) = \sum_{i=1}^{\mathcal{M}(\varepsilon, r\mathbb{B}^d, ||\cdot||)} \text{Vol}\left(\frac{\varepsilon}{2}\mathbb{B}^d\right) \leq \text{Vol}\left(\left(r + \frac{\varepsilon}{2}\right)\mathbb{B}^d\right).$$

Dividing through by $\text{Vol}\left(\frac{\varepsilon}{2}\mathbb{B}^d\right)$ yields

$$\mathcal{M}(\varepsilon, r\mathbb{B}^d, ||\cdot||) \leq \frac{\text{Vol}\left(\left(r + \frac{\varepsilon}{2}\right)\mathbb{B}^d\right)}{\text{Vol}\left(\frac{\varepsilon}{2}\mathbb{B}^d\right)} = \left(\frac{r + \frac{\varepsilon}{2}}{\frac{\varepsilon}{2}}\right)^d = \left(1 + \frac{2r}{\varepsilon}\right)^d.$$

Applying Fact 2 gives us

$$\mathcal{N}(\varepsilon, r\mathbb{B}^d, ||\cdot||) \leq \mathcal{M}(\varepsilon, r\mathbb{B}^d, ||\cdot||) \leq \left(1 + \frac{2r}{\varepsilon}\right)^d,$$

and taking logarithms on both sides renders the desired inequality. □

Theorem 9. *Take any $\Pi \supseteq \{\mathcal{S} \to \mathcal{A}\}$, any $\mathcal{V} \supseteq \{V^\pi \mid \pi \in \Pi^H\}$, and let $D = 0$. For any prior distribution $\mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_1)$ over tabular MDPs, if $\Gamma_k \leq \overline{\Gamma}$ for all $k \in [K]$, then VSRL (Algorithm 2) has*

$$\textsc{BayesRegret}(K, \pi^{(1)}, \ldots, \pi^{(K)}) \leq \mathcal{O}\left(|\mathcal{S}|\sqrt{\overline{\Gamma}|\mathcal{A}|K}\right).$$

Proof. Using Fact 1, we have that

$$\mathcal{R}_1^{\Pi,\mathcal{V}}(D) \leq \mathbb{H}_1(\mathcal{M}^\star) = \mathbb{H}_1(\mathcal{R}^\star, \mathcal{T}^\star) = \mathbb{H}_1(\mathcal{R}^\star) + \mathbb{H}_1(\mathcal{T}^\star \mid \mathcal{R}^\star) = \mathbb{H}_1(\mathcal{R}^\star) + \mathbb{H}_1(\mathcal{T}^\star),$$

where the first equality recognizes that all randomness in the true MDP $\mathcal{M}^\star$ is driven by the model $(\mathcal{R}^\star, \mathcal{T}^\star)$, the second equality applies the chain rule of entropy, and the final equality recognizes that the reward function and transition function random variables are independent.

For some fixed $\varepsilon_\mathcal{R} > 0$, consider the $\frac{\varepsilon_\mathcal{R}}{2}$-cover of the unit interval $[0, 1]$ with respect to the L_1-norm $||\cdot||_1$ as a quantization into bins of width $\varepsilon_\mathcal{R}$. Observe that the true environment reward function $\mathcal{R}^\star : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is well-approximated by mapping state-action pairs onto this $\frac{\varepsilon_\mathcal{R}}{2}$-cover, for a sufficiently small $\varepsilon_\mathcal{R} > 0$. Consequently, we treat $\mathcal{R}^\star$ as a discrete random variable where $|\text{supp}(\mathcal{R}^\star)| = \mathcal{N}(\frac{\varepsilon_\mathcal{R}}{2}, [0, 1], ||\cdot||_1)^{|\mathcal{S}||\mathcal{A}|}$. Recall that, for a discrete random variable X with support on $\mathcal{X}$, $\mathbb{H}(X) \leq \log(|\mathcal{X}|)$. Applying this upper bound and Lemma 9 in sequence, we have that

$$\mathbb{H}_1(\mathcal{R}^\star) \leq |\mathcal{S}||\mathcal{A}| \log\left(\mathcal{N}(\frac{\varepsilon_\mathcal{R}}{2}, [0, 1], ||\cdot||_1)\right) \leq |\mathcal{S}||\mathcal{A}| \log\left(1 + \frac{4}{\varepsilon_\mathcal{R}}\right).$$

Applying the same sequence of steps *mutatis mutandis* for the transition function $\mathcal{T}^\star$ under a $\frac{\varepsilon_\mathcal{T}}{2}$-cover, for some fixed $\varepsilon_\mathcal{T} > 0$, we also have

$$\mathbb{H}_1(\mathcal{T}^\star) \leq |\mathcal{S}|^2|\mathcal{A}| \log\left(\mathcal{N}(\frac{\varepsilon_\mathcal{T}}{2}, [0, 1], ||\cdot||_1)\right) \leq |\mathcal{S}|^2|\mathcal{A}| \log\left(1 + \frac{4}{\varepsilon_\mathcal{T}}\right).$$

Applying these bounds following the earlier rate-distortion function upper bound to the result of Corollary 1 with $D = 0$, we have

$$\textsc{BayesRegret}(K, \pi^{(1)}, \ldots, \pi^{(K)}) \leq \sqrt{\overline{\Gamma}K\left(|\mathcal{S}||\mathcal{A}| \log\left(1 + \frac{4}{\varepsilon_\mathcal{R}}\right) + |\mathcal{S}|^2|\mathcal{A}| \log\left(1 + \frac{4}{\varepsilon_\mathcal{T}}\right)\right)}.$$

□

I Proof of Theorem 4

Our proof of Theorem 4 utilizes the following fact, widely known as the performance-difference lemma, adapted to the finite-horizon setting whose proof we replicate here.

Lemma 10 (Performance-Difference Lemma [89]). *For any finite-horizon MDP $\langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \beta, H \rangle$ and any two non-stationary policies $\pi_1, \pi_2 \in \Pi^H$, let $\rho^{\pi_2}(\tau)$ denote the distribution over trajectories induced by policy π_2. Then,*

$$V_1^{\pi_1} - V_1^{\pi_2} = \mathbb{E}_{\tau \sim \rho^{\pi_2}} \left[\sum_{h=1}^{H} \left(V_h^{\pi_1}(s_h) - Q_h^{\pi_1}(s_h, a_h) \right) \right].$$

Proof.

$$\begin{aligned}
V_1^{\pi_1} - V_1^{\pi_2} &= \mathbb{E}_{s_1 \sim \beta} \left[V_1^{\pi_1}(s_1) - V_1^{\pi_2}(s_1) \right] \\
&= \mathbb{E}_{s_1 \sim \beta} \left[V_1^{\pi_1}(s_1) - \mathbb{E}_{\tau \sim \rho^{\pi_2}} \left[\sum_{h=1}^{H} \mathcal{R}(s_h, a_h) \mid s_1 \right] \right] \\
&= \mathbb{E}_{\tau \sim \rho^{\pi_2}} \left[V_1^{\pi_1}(s_1) - \sum_{h=1}^{H} \mathcal{R}(s_h, a_h) \right] \\
&= \mathbb{E}_{\tau \sim \rho^{\pi_2}} \left[V_1^{\pi_1}(s_1) + \sum_{h=2}^{H} V_h^{\pi_1}(s_h) - \sum_{h=1}^{H} \left(\mathcal{R}(s_h, a_h) - V_{h+1}^{\pi_1}(s_{h+1}) \right) \right] \\
&= \mathbb{E}_{\tau \sim \rho^{\pi_2}} \left[\sum_{h=1}^{H} V_h^{\pi_1}(s_h) - \left(\mathcal{R}(s_h, a_h) + V_{h+1}^{\pi_1}(s_{h+1}) \right) \right] \\
&= \mathbb{E}_{\tau \sim \rho^{\pi_2}} \left[\sum_{h=1}^{H} \left(V_h^{\pi_1}(s_h) - \left(\mathcal{R}(s_h, a_h) + \mathbb{E} \left[V_{h+1}^{\pi_1}(s_{h+1}) \mid s_h, a_h \right] \right) \right) \right] \\
&= \mathbb{E}_{\tau \sim \rho^{\pi_2}} \left[\sum_{h=1}^{H} \left(V_h^{\pi_1}(s_h) - Q_h^{\pi_1}(s_h, a_h) \right) \right],
\end{aligned}$$

where the penultimate line invokes the tower property of expectation. □

Theorem 10. *Fix any $D \geq 0$ and, for each episode $k \in [K]$, let $\widetilde{\mathcal{M}}_k$ be any MDP that achieves the rate-distortion limit of $\mathcal{R}_k^{Q^\star}(D)$ with information source $\mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_k)$ and distortion function $d_{Q^\star}$. Then,*

$$\textsc{BayesRegret}(K, \pi^{(1)}, \ldots, \pi^{(K)}) \leq \mathbb{E} \left[\sum_{k=1}^{K} \mathbb{E}_k \left[V^\star_{\widetilde{\mathcal{M}}_k, 1} - V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k, 1} \right] \right] + (2H+2)K\sqrt{D}.$$

Proof. By applying definitions from Section 3 and applying the tower property of expectation, we have that

$$\textsc{BayesRegret}(K, \pi^{(1)}, \ldots, \pi^{(K)}) = \mathbb{E} \left[\sum_{k=1}^{K} \mathbb{E}_k \left[\Delta_k \right] \right].$$

Examining the kth episode in isolation and applying the definition of episodic regret, we have

$$\begin{aligned}
\mathbb{E}_k\left[\Delta_k\right] &= \mathbb{E}_k\left[V^{\star}_{\mathcal{M}^{\star},1} - V^{\pi^{(k)}}_{\mathcal{M}^{\star},1}\right] \\
&= \mathbb{E}_k\left[V^{\star}_{\mathcal{M}^{\star},1} - V^{\star}_{\widetilde{\mathcal{M}}_k,1} + V^{\star}_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1} + V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\mathcal{M}^{\star},1}\right] \\
&= \mathbb{E}_k\left[V^{\star}_{\mathcal{M}^{\star},1} - V^{\star}_{\widetilde{\mathcal{M}}_k,1} + V^{\star}_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1} + \underbrace{V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1} - V^{\star}_{\widetilde{\mathcal{M}}_k,1}}_{\leq 0} + V^{\star}_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\mathcal{M}^{\star},1}\right] \\
&\leq \mathbb{E}_k\left[V^{\star}_{\mathcal{M}^{\star},1} - V^{\star}_{\widetilde{\mathcal{M}}_k,1} + V^{\star}_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1} + V^{\star}_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\mathcal{M}^{\star},1}\right] \\
&= \mathbb{E}_k\left[V^{\star}_{\mathcal{M}^{\star},1} - V^{\star}_{\widetilde{\mathcal{M}}_k,1} + V^{\star}_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1} + V^{\star}_{\widetilde{\mathcal{M}}_k,1} - V^{\star}_{\mathcal{M}^{\star},1} + V^{\star}_{\mathcal{M}^{\star},1} - V^{\pi^{(k)}}_{\mathcal{M}^{\star},1}\right].
\end{aligned}$$

Observe that

$$\begin{aligned}
\mathbb{E}_k\left[V^{\star}_{\mathcal{M}^{\star},1} - V^{\star}_{\widetilde{\mathcal{M}}_k,1}\right] &\leq \mathbb{E}_k\left[||V^{\star}_{\mathcal{M}^{\star},1} - V^{\star}_{\widetilde{\mathcal{M}}_k,1}||_{\infty}\right] \\
&= \mathbb{E}_k\left[\max_{s\in\mathcal{S}} |V^{\star}_{\mathcal{M}^{\star},1}(s) - V^{\star}_{\widetilde{\mathcal{M}}_k,1}(s)|\right] \\
&= \mathbb{E}_k\left[\max_{s\in\mathcal{S}} |\max_{a\in\mathcal{A}} Q^{\star}_{\mathcal{M}^{\star},1}(s,a) - \max_{a'\in\mathcal{A}} Q^{\star}_{\widetilde{\mathcal{M}}_k,1}(s,a')|\right] \\
&\leq \mathbb{E}_k\left[\max_{s\in\mathcal{S}} \max_{a\in\mathcal{A}} |Q^{\star}_{\mathcal{M}^{\star},1}(s,a) - Q^{\star}_{\widetilde{\mathcal{M}}_k,1}(s,a)|\right] \\
&= \mathbb{E}_k\left[||Q^{\star}_{\mathcal{M}^{\star},1} - Q^{\star}_{\widetilde{\mathcal{M}}_k,1}||_{\infty}\right] \\
&= \mathbb{E}_k\left[\sqrt{||Q^{\star}_{\mathcal{M}^{\star},1} - Q^{\star}_{\widetilde{\mathcal{M}}_k,1}||^2_{\infty}}\right] \\
&\leq \sqrt{\mathbb{E}_k\left[||Q^{\star}_{\mathcal{M}^{\star},1} - Q^{\star}_{\widetilde{\mathcal{M}}_k,1}||^2_{\infty}\right]} \\
&\leq \sqrt{\mathbb{E}_k\left[\sup_{h\in H} ||Q^{\star}_{\mathcal{M}^{\star},h} - Q^{\star}_{\widetilde{\mathcal{M}}_k,h}||^2_{\infty}\right]} \\
&= \sqrt{\mathbb{E}_k\left[d_{Q^{\star}}(\mathcal{M}^{\star}, \widetilde{\mathcal{M}}_k)\right]} \\
&\leq \sqrt{D},
\end{aligned}$$

where the penultimate inequality is due to Jensen's inequality and the final inequality holds as $\widetilde{\mathcal{M}}_k$ achieves the rate-distortion limit under $d_{Q^{\star}}$, by assumption. Moreover, the exact argument can be repeated to see that

$$\begin{aligned}
\mathbb{E}_k\left[V^{\star}_{\widetilde{\mathcal{M}}_k,1} - V^{\star}_{\mathcal{M}^{\star},1}\right] &\leq \mathbb{E}_k\left[||V^{\star}_{\widetilde{\mathcal{M}}_k,1} - V^{\star}_{\mathcal{M}^{\star},1}||_{\infty}\right] \\
&= \mathbb{E}_k\left[||V^{\star}_{\mathcal{M}^{\star},1} - V^{\star}_{\widetilde{\mathcal{M}}_k,1}||_{\infty}\right] \\
&\leq \sqrt{D}.
\end{aligned}$$

Combining these two inequalities yields

$$\begin{aligned}
\mathbb{E}_k\left[\Delta_k\right] &\leq \mathbb{E}_k\left[V^{\star}_{\mathcal{M}^{\star},1} - V^{\star}_{\widetilde{\mathcal{M}}_k,1} + V^{\star}_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1} + V^{\star}_{\widetilde{\mathcal{M}}_k,1} - V^{\star}_{\mathcal{M}^{\star},1} + V^{\star}_{\mathcal{M}^{\star},1} - V^{\pi^{(k)}}_{\mathcal{M}^{\star},1}\right] \\
&\leq \mathbb{E}_k\left[V^{\star}_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1} + V^{\star}_{\mathcal{M}^{\star},1} - V^{\pi^{(k)}}_{\mathcal{M}^{\star},1}\right] + 2\sqrt{D}.
\end{aligned}$$

Observe that by virtue of posterior sampling [135, 121, 120] the compressed MDP being targeted by the agent $\widetilde{\mathcal{M}}_k$ and the sampled MDP M_k are identically distributed, conditioned upon the information available within any history H_k, and so we have

$$\mathbb{E}_k\left[V^{\star}_{\mathcal{M}^{\star},1} - V^{\pi^{(k)}}_{\mathcal{M}^{\star},1}\right] = \mathbb{E}_k\left[V^{\star}_{\mathcal{M}^{\star},1} - V^{\pi^{\star}_{M_k}}_{\mathcal{M}^{\star},1}\right] = \mathbb{E}_k\left[V^{\star}_{\mathcal{M}^{\star},1} - V^{\pi^{\star}_{\widetilde{\mathcal{M}}_k}}_{\mathcal{M}^{\star},1}\right].$$

Now applying the performance-difference lemma (Lemma 10), we see that

$$\begin{aligned}
\mathbb{E}_k\left[V^\star_{\mathcal{M}^\star,1} - V^{\pi^\star_{\widetilde{\mathcal{M}}_k}}_{\mathcal{M}^\star,1}\right] &= \mathbb{E}_k\left[\mathbb{E}_{\rho^{\pi^\star_{\widetilde{\mathcal{M}}_k}}}\left[\sum_{h=1}^H \left(V^\star_{\mathcal{M}^\star,h}(s_h) - Q^\star_{\mathcal{M}^\star,h}(s_h,a_h)\right)\right]\right] \\
&= \mathbb{E}_k\left[\mathbb{E}_{\rho^{\pi^\star_{\widetilde{\mathcal{M}}_k}}}\left[\sum_{h=1}^H \left(\max_{a\in\mathcal{A}} Q^\star_{\mathcal{M}^\star,h}(s_h,a) - Q^\star_{\mathcal{M}^\star,h}(s_h,a_h)\right)\right]\right] \\
&\leq \mathbb{E}_k\left[\mathbb{E}_{\rho^{\pi^\star_{\widetilde{\mathcal{M}}_k}}}\left[\sum_{h=1}^H \left|\max_{a\in\mathcal{A}} Q^\star_{\mathcal{M}^\star,h}(s_h,a) - Q^\star_{\mathcal{M}^\star,h}(s_h,a_h)\right|\right]\right].
\end{aligned}$$

Define $a^\star = \arg\max_{a\in\mathcal{A}} Q^\star_{\mathcal{M}^\star,h}(s_h,a)$ such that

$$\begin{aligned}
\mathbb{E}_k\left[V^\star_{\mathcal{M}^\star,1} - V^{\pi^\star_{\widetilde{\mathcal{M}}_k}}_{\mathcal{M}^\star,1}\right] &= \mathbb{E}_k\left[\mathbb{E}_{\rho^{\pi^\star_{\widetilde{\mathcal{M}}_k}}}\left[\sum_{h=1}^H \left|\max_{a\in\mathcal{A}} Q^\star_{\mathcal{M}^\star,h}(s_h,a) - Q^\star_{\mathcal{M}^\star,h}(s_h,a_h)\right|\right]\right] \\
&= \mathbb{E}_k\left[\mathbb{E}_{\rho^{\pi^\star_{\widetilde{\mathcal{M}}_k}}}\left[\sum_{h=1}^H \left|Q^\star_{\mathcal{M}^\star,h}(s_h,a^\star) - Q^\star_{\mathcal{M}^\star,h}(s_h,a_h)\right|\right]\right] \\
&= \mathbb{E}_k\left[\mathbb{E}_{\rho^{\pi^\star_{\widetilde{\mathcal{M}}_k}}}\left[\sum_{h=1}^H \left|Q^\star_{\mathcal{M}^\star,h}(s_h,a^\star) - Q^\star_{\widetilde{\mathcal{M}}_k,h}(s_h,a^\star) + Q^\star_{\widetilde{\mathcal{M}}_k,h}(s_h,a^\star) - Q^\star_{\mathcal{M}^\star,h}(s_h,a_h)\right|\right]\right].
\end{aligned}$$

Applying the triangle inequality and examining each difference in isolation, we have

$$\begin{aligned}
\mathbb{E}_k\left[\mathbb{E}_{\rho^{\pi^\star_{\widetilde{\mathcal{M}}_k}}}\left[\sum_{h=1}^H \left|Q^\star_{\mathcal{M}^\star,h}(s_h,a^\star) - Q^\star_{\widetilde{\mathcal{M}}_k,h}(s_h,a^\star)\right|\right]\right] &\leq \mathbb{E}_k\left[\mathbb{E}_{\rho^{\pi^\star_{\widetilde{\mathcal{M}}_k}}}\left[\sum_{h=1}^H ||Q^\star_{\mathcal{M}^\star,h} - Q^\star_{\widetilde{\mathcal{M}}_k,h}||_\infty\right]\right] \\
&\leq H\mathbb{E}_k\left[\sup_{h\in H} ||Q^\star_{\mathcal{M}^\star,h} - Q^\star_{\widetilde{\mathcal{M}}_k,h}||_\infty\right] \\
&= H\mathbb{E}_k\left[\sup_{h\in H} \sqrt{||Q^\star_{\mathcal{M}^\star,h} - Q^\star_{\widetilde{\mathcal{M}}_k,h}||_\infty^2}\right] \\
&\leq H\sqrt{\mathbb{E}_k\left[\sup_{h\in H} ||Q^\star_{\mathcal{M}^\star,h} - Q^\star_{\widetilde{\mathcal{M}}_k,h}||_\infty^2\right]} \\
&= H\sqrt{\mathbb{E}_k\left[d_{Q^\star}(\mathcal{M}^\star,\widetilde{\mathcal{M}}_k)\right]} \\
&\leq H\sqrt{D},
\end{aligned}$$

where the penultimate inequality follows from Jensen's inequality and the final inequality follows since $\widetilde{\mathcal{M}}_k$ achieves the rate-distortion limit.

For the remaining term, we have

$$\begin{aligned}
\mathbb{E}_k\left[V^\star_{\mathcal{M}^\star,1} - V^{\pi^\star_{\widetilde{\mathcal{M}}_k}}_{\mathcal{M}^\star,1}\right] &\leq H\sqrt{D} + \mathbb{E}_k\left[\mathbb{E}_{\rho^{\pi^\star_{\widetilde{\mathcal{M}}_k}}}\left[\sum_{h=1}^H \left|Q^\star_{\widetilde{\mathcal{M}}_k,h}(s_h,a^\star) - Q^\star_{\mathcal{M}^\star,h}(s_h,a_h)\right|\right]\right] \\
&= H\sqrt{D} + \mathbb{E}_k\left[\mathbb{E}_{\rho^{\pi^\star_{\widetilde{\mathcal{M}}_k}}}\left[\sum_{h=1}^H \left|Q^\star_{\widetilde{\mathcal{M}}_k,h}(s_h,a^\star) - Q^\star_{\widetilde{\mathcal{M}}_k,h}(s_h,a_h) + Q^\star_{\widetilde{\mathcal{M}}_k,h}(s_h,a_h) - Q^\star_{\mathcal{M}^\star,h}(s_h,a_h)\right|\right]\right] \\
&\leq H\sqrt{D} + \mathbb{E}_k\left[\mathbb{E}_{\rho^{\pi^\star_{\widetilde{\mathcal{M}}_k}}}\left[\sum_{h=1}^H \left|Q^\star_{\widetilde{\mathcal{M}}_k,h}(s_h,a_h) - Q^\star_{\mathcal{M}^\star,h}(s_h,a_h)\right|\right]\right],
\end{aligned}$$

where the inequality follows since a_h is drawn from the optimal policy of $\widetilde{\mathcal{M}}_k$, $\pi^\star_{\widetilde{\mathcal{M}}_k}$, and so $Q^\star_{\widetilde{\mathcal{M}}_k,h}(s_h,a_h) \geq Q^\star_{\widetilde{\mathcal{M}}_k,h}(s_h,a^\star)$. Repeating the identical argument from above yields

$$\mathbb{E}_k\left[\mathbb{E}_{\rho^{\pi^\star_{\widetilde{\mathcal{M}}_k}}}\left[\sum_{h=1}^H \left|Q^\star_{\widetilde{\mathcal{M}}_k,h}(s_h,a_h) - Q^\star_{\mathcal{M}^\star,h}(s_h,a_h)\right|\right]\right] \leq \mathbb{E}_k\left[\mathbb{E}_{\rho^{\pi^\star_{\widetilde{\mathcal{M}}_k}}}\left[\sum_{h=1}^H ||Q^\star_{\mathcal{M}^\star,h} - Q^\star_{\widetilde{\mathcal{M}}_k,h}||_\infty\right]\right]$$
$$\leq H\sqrt{D}.$$

Substituting back, we see that

$$\mathbb{E}_k\left[V^\star_{\mathcal{M}^\star,1} - V^{\pi^{(k)}}_{\mathcal{M}^\star,1}\right] \leq \mathbb{E}_k\left[\mathbb{E}_{\rho^{\pi^\star_{\widetilde{\mathcal{M}}_k}}}\left[\sum_{h=1}^H \left|\max_{a\in\mathcal{A}} Q^\star_{\mathcal{M}^\star,h}(s_h,a) - Q^\star_{\mathcal{M}^\star,h}(s_h,a_h)\right|\right]\right] \leq 2H\sqrt{D}.$$

Thus, we may complete our bound as

$$\begin{aligned}
\mathbb{E}_k\left[\Delta_k\right] &\leq \mathbb{E}_k\left[V^\star_{\mathcal{M}^\star,1} - V^\star_{\widetilde{\mathcal{M}}_k,1} + V^\star_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1} + V^\star_{\widetilde{\mathcal{M}}_k,1} - V^\star_{\mathcal{M}^\star,1} + V^\star_{\mathcal{M}^\star,1} - V^{\pi^{(k)}}_{\mathcal{M}^\star,1}\right] \\
&\leq \mathbb{E}_k\left[V^\star_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1} + V^\star_{\mathcal{M}^\star,1} - V^{\pi^{(k)}}_{\mathcal{M}^\star,1}\right] + 2\sqrt{D} \\
&\leq \mathbb{E}_k\left[V^\star_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1}\right] + (2H+2)\sqrt{D}.
\end{aligned}$$

Applying this upper bound on episodic regret in each episode yields

$$\begin{aligned}
\textsc{BayesRegret}(K,\pi^{(1)},\ldots,\pi^{(K)}) &= \mathbb{E}\left[\sum_{k=1}^K \mathbb{E}_k\left[\Delta_k\right]\right] \\
&\leq \mathbb{E}\left[\sum_{k=1}^K \mathbb{E}_k\left[V^\star_{\widetilde{\mathcal{M}}_k,1} - V^{\pi^{(k)}}_{\widetilde{\mathcal{M}}_k,1}\right]\right] + 2K(H+1)\sqrt{D},
\end{aligned}$$

as desired. □

J Proof of Lemma 3

To show Lemma 3, we prove the following more general result which applies whenever a distortion function adheres to a specific functional form.

Let $V, \widehat{V}$ be two arbitrary random variables defined on the same measurable space $(\mathcal{V}, \mathbb{V})$ and define the associated rate-distortion function as

$$\mathcal{R}(D) = \inf_{\widehat{V}\in\Lambda(D)} \mathbb{I}(V;\widehat{V}) = \inf_{\widehat{V}\in\Lambda(D)} D_{\mathrm{KL}}(\mathbb{P}((V,\widehat{V})\in\cdot) \,||\, \mathbb{P}(V\in\cdot)\times\mathbb{P}(\widehat{V}\in\cdot)),$$

where the distortion function $d : \mathcal{V}\times\mathcal{V} \to \mathbb{R}_{\geq 0}$ has the form $d(v,\widehat{v}) = \ell(f(v), f(\widehat{v}))$ for any two known, deterministic functions, $f : \mathcal{V} \to \mathcal{Z}$ and a semi-metric $\ell : \mathcal{Z}\times\mathcal{Z} \to \mathbb{R}_{\geq 0}$. Effectively, this structural constraint says that our distortion measure between the original V and compressed $\widehat{V}$ only depends on the statistics $f(V)$ and $f(\widehat{V})$. Under such a constraint, we may prove the following lemma

Lemma 11. *If $D = 0$ and $\widehat{V}$ achieves the rate-distortion limit, then we have the Markov chain $V \to f(V) \to \widehat{V}$*

Proof. Assume for the sake of contradiction that there exists a random variable $\widehat{V}$ that achieves the rate-distortion limit with $D = 0$ but does not induce the Markov chain $V \to f(V) \to \widehat{V}$. Since mutual information is non-negative and $\mathbb{I}(V;\widehat{V} \mid f(V)) = 0$ implies the Markov chain $V \to f(V) \to \widehat{V}$, it must be the case that $\mathbb{I}(V;\widehat{V} \mid f(V)) > 0$. Consider an independent random variable $\widehat{V}' \sim \mathbb{P}(\widehat{V} \mid f(V))$ such that

$$\mathbb{I}(V;\widehat{V}') = \mathbb{I}(V;\widehat{V}) - \underbrace{\mathbb{I}(V;\widehat{V} \mid f(V))}_{>0} < \mathbb{I}(V;\widehat{V}) = \mathcal{R}(D).$$

Clearly, we have retained all bits of information needed to preserve $f(V)$ in $\widehat{V}'$, thereby achieving the same expected distortion constraint. However, this implies that $\widehat{V}'$ achieves a strictly lower rate, contradicting our assumption that $\widehat{V}$ achieves the rate-distortion limit. Therefore, it must be the case that when $D = 0$ and $\widehat{V}$ achieves the rate-distortion limit, we have $\mathbb{I}(V; \widehat{V} \mid f(V)) = 0$ which implies the Markov chain $V \to f(V) \to \widehat{V}$.

□

Lemma 12. *For each episode $k \in [K]$ and for $D = 0$, let $\widetilde{\mathcal{M}}_k$ be a MDP that achieves the rate-distortion limit of $\mathcal{R}_k^{Q^\star}(D)$ with information source $\mathbb{P}(\mathcal{M}^\star \mid H_k)$ and distortion function $d_{Q^\star}$. Then, we have the Markov chain $\mathcal{M}^\star \to Q^\star_{\mathcal{M}^\star} \to \widetilde{\mathcal{M}}_k$, where $Q^\star_{\mathcal{M}^\star} = \{Q^\star_{\mathcal{M}^\star,h}\}_{h\in[H]}$ is the collection of random variables denoting the optimal action-value functions of $\mathcal{M}^\star$.*

Proof. Recall that our distortion function,

$$d_{Q^\star}(\mathcal{M}, \widehat{\mathcal{M}}) = \sup_{h\in[H]} ||Q^\star_{\mathcal{M},h} - Q^\star_{\widehat{\mathcal{M}},h}||_\infty^2 = \sup_{h\in[H]} \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} |Q^\star_{\mathcal{M},h}(s,a) - Q^\star_{\widehat{\mathcal{M}},h}(s,a)|^2,$$

only depends on the MDPs $\mathcal{M}$ and $\widehat{\mathcal{M}}$ through their respective optimal action-value functions, $\{Q^\star_{\mathcal{M},h}\}_{h\in[H]}$ and $\{Q^\star_{\widehat{\mathcal{M}},h}\}_{h\in[H]}$. Consequently, the claim holds immediately by applying Lemma 11 where f computes the optimal action-value functions of an input MDP for each timestep $h \in [H]$ and ℓ is the metric induced by the infinity norm on $\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$. □

K Proof of Theorem 5

Our proof of Theorem 5 proceeds by leveraging Lemma 3 (instead of Fact 1) before following the same style of argument as used in Theorem 3.

Theorem 11. *For $D = 0$ and any prior distribution $\mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_1)$ over tabular MDPs, if $\Gamma_k \le \overline{\Gamma}$ for all $k \in [K]$, then VSRL with distortion function $d_{Q^\star}$ has*

$$\textsc{BayesRegret}(K, \pi^{(1)}, \ldots, \pi^{(K)}) \le \widetilde{\mathcal{O}}\left(\sqrt{\overline{\Gamma}|\mathcal{S}||\mathcal{A}|KH}\right).$$

Proof. Starting with the information-theoretic regret bound in Corollary 2, observe that for $\mathcal{M}^\star \sim \mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_1)$, we have the Markov chain $\mathcal{M}^\star \to Q^\star_{\mathcal{M}^\star} \to \widetilde{\mathcal{M}}_1$, by virtue of Lemma 3. By the data-processing inequality, we immediately recover the following chain of inequalities:

$$\mathcal{R}_1^{Q^\star}(D) \le \mathbb{I}_1(\mathcal{M}^\star; \widetilde{\mathcal{M}}_1) \le \mathbb{I}_1(\mathcal{M}^\star; Q^\star_{\mathcal{M}^\star}).$$

Recognizing that the optimal value functions are a deterministic function of the MDP $\mathcal{M}^\star$ itself, we have

$$\mathbb{I}_1(\mathcal{M}^\star; Q^\star_{\mathcal{M}^\star}) = \mathbb{H}_k(Q^\star_{\mathcal{M}^\star}) - \mathbb{H}_k(Q^\star_{\mathcal{M}^\star} \mid \mathcal{M}^\star) = \mathbb{H}_k(Q^\star_{\mathcal{M}^\star}) = \mathbb{H}_k(Q^\star_{\mathcal{M}^\star,1}, \ldots, Q^\star_{\mathcal{M}^\star,H}) \le \sum_{h=1}^{H} \mathbb{H}_k(Q^\star_{\mathcal{M}^\star,h}),$$

where the final inequality follows by applying the chain rule of entropy and the fact that conditioning reduces entropy, in sequence.

At this point, recalling the salient exposition in the proof of Theorem 3 concerning the use of metric entropy for such function-valued random variables, we proceed to consider the $\varepsilon_{Q^\star}$-cover of the interval $[0, H]$ with respect to the L_1-norm $||\cdot||_1$, for some fixed $0 < \varepsilon_{Q^\star} < H$. Since, for a sufficiently small choice of $\varepsilon_{Q^\star}$, $Q^\star_{\mathcal{M}^\star,h}$ is well-approximated as a discrete random variable for any $h \in [H]$, we recall that the entropy of a discrete random variable X taking values on $\mathcal{X}$ is bounded as $\mathbb{H}(X) \le \log(|\mathcal{X}|)$. Applying this upper bound and Lemma 9 in sequence, we have that

$$\sum_{h=1}^{H} \mathbb{H}_k(Q^\star_{\mathcal{M}^\star,h}) \le |\mathcal{S}||\mathcal{A}|H \log\left(\mathcal{N}(\frac{\varepsilon_{Q^\star}}{2}, [0,H], ||\cdot||_1)\right) \le |\mathcal{S}||\mathcal{A}|H \log\left(1 + \frac{4H}{\varepsilon_{Q^\star}}\right).$$

Applying these upper bounds to the result of Corollary 5 and recalling that $D = 0$, we have

$$\textsc{BayesRegret}(K, \pi^{(1)}, \ldots, \pi^{(K)}) \leq \sqrt{\overline{\Gamma} K |\mathcal{S}||\mathcal{A}| H \log\left(1 + \frac{4H}{\varepsilon_{Q^\star}}\right)}.$$

□