arXiv:2407.12185v1 [cs.LG] 16 Jul 2024

Satisficing Exploration for Deep Reinforcement Learning

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

Saurabh Kumar
Department of Computer Science
Stanford University
szk@stanford.edu

Ramki Gummadi
Google Research, Brain Team
gsrk@google.com

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

Abstract

A default assumption in the design of reinforcement-learning algorithms is that a decision-making agent always explores to learn optimal behavior. In sufficiently complex environments that approach the vastness and scale of the real world, however, attaining optimal performance may in fact be an entirely intractable endeavor and an agent may seldom find itself in a position to complete the requisite exploration for identifying an optimal policy. Recent work has leveraged tools from information theory to design agents that deliberately forgo optimal solutions in favor of sufficiently-satisfying or *satisficing* solutions, obtained through lossy compression. Notably, such agents may employ fundamentally different exploratory decisions to learn satisficing behaviors more efficiently than optimal ones that are more data intensive. While supported by a rigorous corroborating theory, the underlying algorithm relies on model-based planning, drastically limiting the compatibility of these ideas with function approximation and high-dimensional observations. In this work, we remedy this issue by extending an agent that directly represents uncertainty over the optimal value function allowing it to both bypass the need for model-based planning and to learn satisficing policies. We provide simple yet illustrative experiments that demonstrate how our algorithm enables deep reinforcement-learning agents to achieve satisficing behaviors. In keeping with previous work on this setting for multi-armed bandits, we additionally find that our algorithm is capable of synthesizing optimal behaviors, when feasible, more efficiently than its non-information-theoretic counterpart.

1 Introduction

In recent years, there has been a tectonic shift in the most celebrated successes of deep reinforcement learning that transcends the initial arcade games (Tesauro, 1995; Bellemare et al., 2013; Mnih et al., 2015; Silver et al., 2016; Schrittwieser et al., 2020) where the sub-field began in favor of real-world applications (Dulac-Arnold et al., 2021); an incomplete list of some notable examples includes classic robotic control problems (Lillicrap et al., 2016; Schulman et al., 2016; Akkaya et al., 2019), theorem provers (Kaliszyk et al., 2018), molecular design (Popova et al., 2018; Zhou et al., 2019), superpressure balloon flight controllers (Bellemare et al., 2020), computer chip layout design (Mirhoseini et al., 2021), matrix multiplication algorithms (Fawzi et al., 2022), and sophisticated dialogue agents (Stiennon et al., 2020; Ouyang et al., 2022).

While these advances are exciting, they also bring awareness to some of the harsh realities that real-world decision-making agents must inevitably face as they continue to proliferate and push the frontier into novel application areas. It is important to recognize that these agents are computationally bounded and, in many problems of interest, must contend with additional real-world constraints on time and other resources. As a concrete example of a scenario that behooves satisficing solutions over optimal ones, consider an environment designed for a sequential decision-making agent to autonomously complete tasks on the Internet (Shi et al., 2017; Yao et al., 2022), such as making e-commerce purchases or querying for pieces of information. Given the wealth of knowledge sources and vendors that can be accessed through the Internet, any one query or desired item for purchase will likely yield a minimum of hundreds, if not thousands, of potential results. As has been well known from years of research on provably-efficient exploration in reinforcement learning (Kearns & Singh, 2002; Kakade, 2003; Strehl et al., 2009; Jin et al., 2018), identifying the globally optimal solution technically requires sifting through this vast number of candidates from start to finish, lest the agent settle prematurely and miss out on the best purchase deal or most accurate answer for an input query. However, these environments (Shi et al., 2017; Yao et al., 2022) are incredibly rich, complex, and contain a wealth of information that likely exceeds the capacity of any one agent; pursuing optimal behaviors in such environments may no longer be a tractable endeavor as agents are forced to continually explore in order to obtain the potentially unbounded amount of information needed for synthesizing an optimal policy.

Recent work by Arumugam & Van Roy (2022) examines this capacity-limited setting and introduces a Bayesian reinforcement-learning algorithm for prioritizing exploration around an alternative, surrogate learning target (Lu et al., 2023), which strikes a balance between being sufficiently simple to learn while also being suitably performant for the task at hand; crucially, they ground this procedure for learning *satisficing* behaviors (Simon, 1955; 1956; Newell et al., 1958; 1972; Simon, 1982) formally through lossy compression and rate-distortion theory (Shannon, 1959), resulting in a posterior-sampling algorithm that is capable of making fundamentally different exploratory choices than those of an agent purely seeking optimal behavior without regard for its own capacity constraints. One limitation of their proposed algorithm is that it is model-based, hindering its application to large-scale problems by the prerequisite of approximate, near-optimal planning.

While progress on the open problem of high-dimensional model-based planning continues (Wang et al., 2019), this paper develops a model-free approach inspired by a long line of work that performs Thompson sampling (Thompson, 1933; Russo et al., 2018) over the optimal action-value function (Osband et al., 2016a; Osband, 2016; Osband et al., 2016b; 2019; 2023), bypassing the need for planning. We introduce an algorithm that amortizes the lossy compression of Arumugam & Van Roy (2021a) across the state space, allowing classic algorithms from the information theory community for computing the rate-distortion function to become viable (Blahut, 1972; Arimoto, 1972). We demonstrate this precisely through computational experiments which showcase a single deep reinforcement-learning algorithm that can be parameterized to achieve a broad spectrum of satisficing solutions, which attain varying degrees of performance.

The paper is organized as follows: we outline the problem formulation in Section 2 before presenting our deep reinforcement-learning algorithm in Section 3. In Section 4, we outline the core hypothesis that underlies our empirical investigation before presenting the complementary experimental results. In the interest of space, all algorithms, an overview of related work, and additional details around empirical results are relegated to the appendix.

2 Problem Formulation

We formulate a sequential decision-making problem as an infinite-horizon, discounted Markov Decision Process (MDP) (Bellman, 1957; Puterman, 1994) defined by $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \rho, \mathcal{T}, \mu, \gamma \rangle$. Here $\mathcal{S}$ denotes a set of states, $\mathcal{A}$ is a set of actions, $\rho : \mathcal{S} \times \mathcal{A} \to [0, 1]$ 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, $\mu \in \Delta(\mathcal{S})$ is an initial state distribution, and $\gamma \in [0,1)$ is the discount factor communicating a preference for near-term versus long-term rewards. Beginning with an initial state $s_0 \sim \mu$, for each timestep $t \in \mathbb{N}$, the agent observes the current state $s_t \in \mathcal{S}$, selects action $a_t \sim \pi(\cdot \mid s_t) \in \mathcal{A}$, enjoys a reward $r_t = \rho(s_t, a_t) \in [0,1]$, and transitions to the next state $s_{t+1} \sim \mathcal{T}(\cdot \mid s_t, a_t) \in \mathcal{S}$.

A stationary, stochastic policy $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$, encodes a pattern of behavior mapping individual states to distributions over possible actions whose overall performance in any MDP $\mathcal{M}$ when starting at state $s \in \mathcal{S}$ and taking action $a \in \mathcal{A}$ is assessed by its associated action-value function $Q^\pi(s,a) = \mathbb{E}\left[\sum\limits_{t=0}^{\infty} \gamma^t \rho(s_t, a_t) \mid s_0 = s, a_0 = a\right]$, where the expectation integrates over randomness in the action selections and transition dynamics. Taking the corresponding value function as $V^\pi(s) = \mathbb{E}_{a \sim \pi(\cdot|s)}\left[Q^\pi(s,a)\right]$ and letting $\Pi \triangleq \{\mathcal{S} \rightarrow \Delta(\mathcal{A})\}$ denote the set of all stochastic policies, we define the optimal policy $\pi^\star$ as achieving supremal value with $V^\star(s) = \sup\limits_{\pi \in \Pi} V^\pi(s) = \max\limits_{a^\star \in \mathcal{A}} Q^\star(s, a^\star)$ and $Q^\star(s,a) = \sup\limits_{\pi \in \Pi} Q^\pi(s,a)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. As the agent interacts with the environment over the course of $K \in \mathbb{N}$ episodes, we let $\tau_k = (s_1^{(k)}, a_1^{(k)}, r_1^{(k)}, \ldots)$ be the random variable denoting the trajectory experienced by the agent in the kth episode, for any $k \in [K]$. 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. Abstractly, a reinforcement-learning algorithm is a sequence of policies $\{\pi^{(k)}\}_{k \in [K]}$ where, for each episode $k \in [K]$, $\pi^{(k)} : \mathcal{H}_k \rightarrow \Pi$ is a function of the current history H_k.

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: $\mathbb{E}_k[X] \triangleq \mathbb{E}[X|H_k]$.

3 Satisficing with Randomized Value Functions

In this section, we extend the preceding problem formulation to the Bayesian reinforcement learning setting used throughout this work. We then introduce a deep reinforcement-learning algorithm that yields satisficing solutions via lossy compression of the optimal action-value function, $Q^\star$.

3.1 Randomized Value Functions

Our work operates in the Bayesian reinforcement learning (Bellman & Kalaba, 1959; Duff, 2002; Ghavamzadeh et al., 2015) setting, wherein the underlying MDP the agent interacts with is unknown and, therefore, a random variable. An agent's initial uncertainty in this unknown, true MDP $\mathcal{M}$ is reflected by a prior distribution $\mathbb{P}(\mathcal{M} \in \cdot \mid H_1)$. A typical objective for this setting is to design a provably-efficient reinforcement-learning algorithm that incurs bounded Bayesian regret, which simply takes the traditional notion of regret and applies an outer expectation under the agent's prior to account for the unknown environment. Unlike prior work (Russo & Van Roy, 2022; Arumugam & Van Roy, 2021a; 2022), which successfully leverages rate-distortion theory in this regard, the focus of the present work is to give rise to a practical agent design that is compatible with deep reinforcement learning; up to now, this has only been realized for multi-armed bandit problems (Arumugam & Van Roy, 2021a;b).

In the absence of (or without regard for) any limitations on agent capacity, one fruitful strategy for addressing this setting involves reducing the agent's epistemic uncertainty (Der Kiureghian & Ditlevsen, 2009) through Thompson sampling (Thompson, 1933; Russo et al., 2018), resulting in the well-studied Posterior Sampling for Reinforcement Learning (PSRL) algorithm (Strens, 2000; Osband et al., 2013; Osband & Van Roy, 2014; Abbasi-Yadkori & Szepesvari, 2014; Agrawal & Jia, 2017; Osband & Van Roy, 2017; Lu & Van Roy, 2019) that enjoys rigorous theoretical guarantees for provably-efficient exploration. Despite this fact, PSRL poses a significant computational hurdle for deployment in large-scale environments as acting optimally with respect to a single posterior sample in each episode (per Thompson sampling) requires MDP planning for the optimal policy. Even while PSRL can still support efficient learning when the policy used in each episode is only near-optimal (see Algorithm 5.1 of (Osband, 2016)), obtaining such a policy is itself a computationally-intensive procedure, especially in conjunction with function approximation (Wang et al., 2019).

Due to these challenges, recent work that combines principled, uncertainty-based exploration with deep reinforcement learning has been almost exclusively driven by the Randomized Value Functions (RVF) algorithm (Osband et al., 2016b; Osband, 2016; O'Donoghue et al., 2018; Osband et al., 2019; 2023) which begins with a prior distribution over the optimal action-value function (rather than the MDP model $\langle \rho, \mathcal{T} \rangle$ as in PSRL) and, for each episode $k \in [K]$, performs Thompson sampling with respect to the current posterior distribution $\widehat{Q} \sim \mathbb{P}(Q^\star \in \cdot \mid H_k)$. While maintaining an equivalence to PSRL in the tabular MDP setting (see Theorem 7.1 of (Osband, 2016)) alongside a complementary regret bound, RVF avoids the computational inefficiencies of PSRL by reasoning over statistically-plausible value functions and, at each episode, behaving greedily with respect to the sampled action-value function: $\pi^{(k)}(s) \in \arg\max\limits_{a \in \mathcal{A}} \widehat{Q}(s, a)$.

While preliminary instantiations of RVF with deep neural networks relied on maintaining an approximate posterior over $Q^\star$ via computationally-inefficient ensembles (Osband et al., 2016a; 2018; Dwaracherla et al., 2022) or hypernetworks (Dwaracherla et al., 2020), recent follow-up work has gone on to address these limitations and retain the benefits of RVF with only a modest increase in computational effort (Osband et al., 2021; 2023).

3.2 Blahut-Arimoto Randomized Value Functions

Just as an agent employing PSRL will relentlessly explore to identify the underlying MDP $\mathcal{M}$ (Arumugam & Van Roy, 2022), a RVF agent will engage in a similar pursuit of the optimal action-value function $Q^\star$. While this reality reflects a desire to always act in search of optimal behavior, real-world reinforcement learning must instead contend with a simple, computationally-bounded agent interacting within an overwhelmingly-complex environment where limitations on time and resources may cause optimal behavior to no longer reside within the agent's means (Arumugam & Van Roy, 2021a; Russo & Van Roy, 2022; Lu et al., 2023; Arumugam & Van Roy, 2022). A line of prior work (Russo et al., 2017; Russo & Van Roy, 2018b; Arumugam & Van Roy, 2021a;b) has studied and addressed this issue for the multi-armed bandit setting (Lai & Robbins, 1985; Bubeck et al., 2012; Lattimore & Szepesvári, 2020) while progress for reinforcement learning has been exclusively theoretical (Arumugam & Van Roy, 2022) in nature.

Rather than unrealistically presuming an agent has the (potentially unlimited) capacity needed to negotiate amongst these numerous choices and acquire the requisite bits of information for identifying the best option, one might instead take inspiration from human decision makers (Tenenbaum et al., 2011; Lake et al., 2017), whose cognition is known to be resource-limited (Simon, 1956; Newell et al., 1958; 1972; Simon, 1982; Gigerenzer & Goldstein, 1996; Vul et al., 2014; Griffiths et al., 2015; Gershman et al., 2015; Lieder & Griffiths, 2020; Bhui et al., 2021; Brown et al., 2022; Ho et al., 2022), and settle for a near-optimal or satisficing solution. In this paper, we build upon a long-line of work in the cognitive-science literature that formalizes the limits of bounded decision-makers using the tools of information theory and doing so in tandem with reinforcement learning (Sims, 2003; Peng, 2005; Parush et al., 2011; Botvinick et al., 2015; Sims, 2016; 2018; Zenon et al., 2019; Ho et al., 2020; Gershman & Lai, 2020; Gershman, 2020; Mikhael et al., 2021; Lai & Gershman, 2021; Gershman, 2021; Jakob & Gershman, 2022; Bari & Gershman, 2022; Arumugam et al., 2024). Crucially, however, we do so in a manner meant to retain and generalize the elegant theoretical properties of RVF while also maintaining its compatibility with deep reinforcement learning.

We design an agent that solves a rate-distortion optimization on a per-timestep basis once the current state has already been observed. Consequently, the learning target computed by each lossy compression problem is a target action $\widetilde{A}_t$ (Arumugam & Van Roy, 2021a;b; Arumugam et al., 2024) that leverages current knowledge of $Q^\star$ to achieve satisficing performance when executed from each state. Thus, for any state $s \in \mathcal{S}$ and distortion threshold $D \in \mathbb{R}_{\geq 0}$, the resulting rate-distortion function is given by $\mathcal{R}_k(s, D) = \inf\limits_{\widetilde{A} \in \mathcal{A}} \mathbb{I}_k(Q^\star; \widetilde{A})$ s.t. $\mathbb{E}_k\left[d_s(Q^\star, \widetilde{A})\right] \leq D$, where the distortion function $d_s : \mathcal{Q} \times \mathcal{A} \rightarrow \mathbb{R}_{\geq 0}$ induced by any state $s \in \mathcal{S}$ is defined as $d_s(Q^\star, \widetilde{a}) = \left(\max\limits_{a \in \mathcal{A}} Q^\star(s,a) - Q^\star(s, \widetilde{a})\right)^2$.

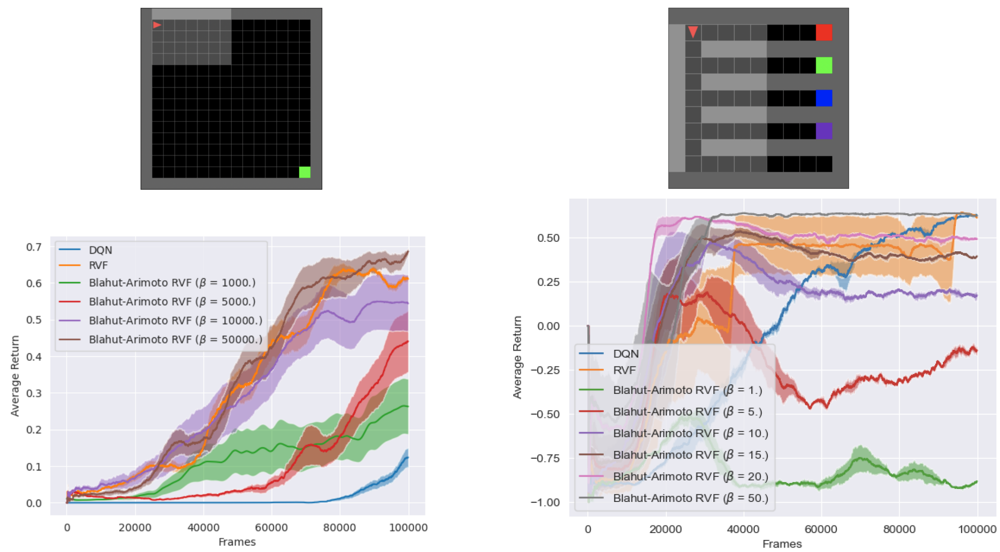
From an information-theoretic perspective, this formulation is akin to lossy source coding with side information available at the decoder (Wyner & Ziv, 1976; Berger & Gibson, 1998). Thinking about the extremes of this rate-distortion trade-off, notice that exclusive concern with rate minimization ($D \uparrow \infty$) yields a uniform distribution over all actions at each state; meanwhile, exclusive concern with distortion minimization ($D = 0$) recovers the greedy action for the particular realization of $Q^\star$, as RVF would. An agent only concerned with optimal behavior must obtain all $\mathbb{H}_1(Q^\star)$ bits of information in order for each of the rate-distortion optimization problems to recover the optimal action at each state. In contrast, an agent that is only interested in target actions that are easy to learn obtains a uniform distribution over actions in every state. Naturally, the intermediate region between these extremes reflects a spectrum of satisficing policies that, within each state, focuses on learning a more tractable, near-optimal action in a manner analogous to satisficing algorithms in the multi-armed bandit setting (Arumugam & Van Roy, 2021a;b).

We may employ the classic Blahut-Arimoto algorithm (Blahut, 1972; Arimoto, 1972) to compute the channel achieving the rate-distortion limit at each timestep; rather than having an explicit distortion threshold D, this algorithm consumes as input a Lagrange multiplier $\beta \in \mathbb{R}_{\geq 0}$ that communicates an implicit preference for the desired trade-off between rate and distortion. While only computationally feasible for a discrete information source and a discrete channel output, the latter requirement is immediately satisfied for MDPs with a discrete action space ($|\mathcal{A}| < \infty$) or a suitably-fine quantization of a continuous action space. To address the former constraint around the information source, we may employ the so-called "plug-in estimator" of the rate-distortion function (Harrison & Kontoyiannis, 2008), which replaces a continuous information source with the discrete empirical distribution obtained via Monte-Carlo sampling and is not only asymptotically consistent (Harrison & Kontoyiannis, 2008) but also admits a finite-sample approximation guarantee (Palaiyanur & Sahai, 2008).

The resulting Blahut-Arimoto Randomized Value Functions (BA-RVF) algorithm is given as Algorithm 1 in Appendix B. It is worth mentioning that BA-RVF does bear increased computational cost per-timestep in comparison to traditional RVF and future work may benefit from using ideas like distillation (Rusu et al., 2015) to reduce these costs during rollouts in exchange for increased computational overhead between episodes. An additional drawback of BA-RVF is the dependence of agent performance on the hyperparameter, β; unfortunately, as β is a Lagrange multiplier (Boyd & Vandenberghe, 2004), it must be tuned on a per-problem basis although future work may benefit from finding heuristic schemes for tuning or adapting β over time that work well across a broad range of problems.

4 Experiments

While the typical empirical evaluation for deep reinforcement-learning agents centers around demonstrating efficient acquisition of optimal behaviors, our experiments instead aim to elucidate how our proposed Blahut-Arimoto RVF algorithm (Algorithm 1) yields a successful generalization of the



(a) MiniGrid-Empty (b) MiniGrid-CorridorEnv

Figure 1: (Top) MiniGrid environments used in our empirical evaluation of Blahut-Arimoto RVF. An observation is a partial image of the whole grid indicated by the shaded region. Black tiles represent empty squares, gray tiles represent walls, and colored tiles represent goal states. The agent begins in the upper left corner and an episode terminates when the agent either reaches a goal state or takes 100 steps. (Bottom) Learning curves of DQN, RVF, and Blahut-Arimoto RVF.

standard RVF algorithm capable of recovering a broad spectrum of satisficing solutions while still retaining the ability to gracefully address the challenge of exploration. To this end, we begin with results on two simple yet illustrative tasks that serve as unit tests of our core empirical hypothesis and leave the task of orchestrating a large-scale empirical demonstration of our algorithm to future work. In particular, we build our evaluation around two MiniGrid environments Chevalier-Boisvert et al. (2018): (1) MiniGrid-Empty-16x16-v0, a standard MiniGrid domain with a single terminal goal state providing sparse positive reward, and (2) MiniGrid-Corridor-v0, a custom designed environment containing multiple goal states which provide rewards proportional to their distance from the agent's initial position. As noted in Figure 1, both environments are partially observable where the agent is given a limited egocentric view of the world and has three movement actions available for execution: RotateLeft, RotateRight, and Forward.

In both domains, we train Blahut-Arimoto RVF agents with different values of the Lagrange multiplier $\beta \in \mathbb{R}_{\geq 0}$ to verify that this parameter successfully controls the trade-off between rate and distortion; concretely, larger values of β should yield agents more concerned with learning near-optimal policies. To contextualize the results achieved by each of these Blahut-Arimoto RVF agents, we also include baseline results attained by standard DQN (Mnih et al., 2015) and RVF agents, where the latter uses an epistemic neural network (Osband et al., 2021) for representing uncertainty over $Q^\star$ in a computationally-efficient manner. We train all agents for $100,000$ frames and report the average un-discounted episodic return achieved throughout training over multiple independent trials (8 seeds on MiniGrid-Empty-16x16-v0 and 3 seeds on MiniGrid-CorridorEnv-v0).

In order to further underscore how our Blahut-Arimoto RVF agent achieves satisficing behaviors while retaining the strategic effectiveness of deep exploration, we offer an additional experiment using a variant of the classic RiverSwim environment (Strehl & Littman, 2008) show in Figure 2 from Osband et al. (2013).

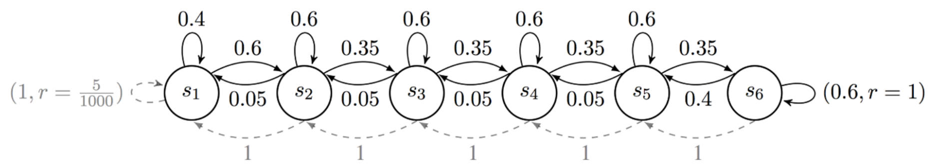


Figure 2: The RiverSwim MDP of Strehl & Littman (2008) as studied by Osband et al. (2013).

Specifically, we provide results for an environment called ConfluenceSwim[1] which consists of three instances of the RiverSwim environment joined by a common initial state. Importantly, the "current" of each river that governs the success of the agent's movements towards the rewarding state at the opposite end varies and leads to three different levels of difficulty. The hardest level of difficulty adopts the transition structure shown in Figure 2 while the other rivers allow the agent to more easily swim upstream. Naturally, swimming up the rivers with weaker currents yields a smaller reward than engaging with the hardest river. Strategic deep exploration, however, is still needed to successfully traverse any one of the rivers.

Figure 3 shows how varying the Lagrange multiplier hyperparameter in Blahut-Arimoto RVF still yields a spectrum of satisficing behaviors where the agent may opt to settle for the smaller reward at one of the easier rivers instead of traversing the most challenging river to obtain an optimal policy. Moreover, just as in prior work on rate-distortion-theoretic exploration for multi-armed bandit problems (Arumugam & Van Roy, 2021a;b), we observe the existence of a β value for BA-RVF that synthesizes optimal behavior more efficiently than classic RVF, which can be interpreted as running BA-RVF with an incredibly large $\beta = 10^6$ such that the resulting target action at each state is the greedy action for a single, fixed posterior sample.

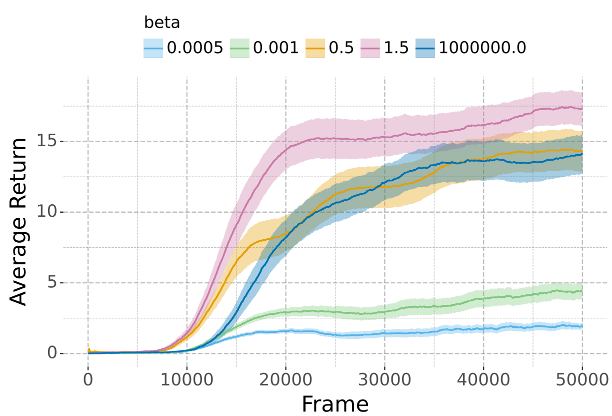


Figure 3: Learning curves for BA-RVF varying β valuse in the ConfluenceSwim environment.

5 Conclusion

In this work, we challenge a core premise of agent design in deep reinforcement learning: that an agent should orient its exploration in pursuit of optimal behavior without regard for the complexity of the underlying environment. Using rate-distortion theory, we offer an agent designed to prioritize exploration towards satisficing behaviors and successfully dovetails with deep reinforcement learning. Our computational results demonstrate the efficacy of this agent in not only generalizing to accommodate satisficing solutions while retaining a graceful handling of the exploration challenge but also in synthesizing optimal solutions more efficiently than its non-satisficing counterpart. Future work still remains to precisely clarify how data efficiency factors into learning these satisficing behaviors.

[1] A confluence is the point where multiple rivers meet.

References

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

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

David Abel, David Hershkowitz, and Michael Littman. Near optimal behavior via approximate state abstraction. In *International Conference on Machine Learning*, pp. 2915–2923. PMLR, 2016.

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, pp. 3134–3142, 2019.

Alekh Agarwal and Tong Zhang. Model-based RL with Optimistic Posterior Sampling: Structural Conditions and Sample Complexity. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Alekh Agarwal, Mikael Henaff, Sham Kakade, and Wen Sun. PC-PG: Policy Cover Directed Exploration for Provable Policy Gradient Learning. *Advances in Neural Information Processing Systems*, 33:13399–13412, 2020a.

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. FLAMBE: Structural Complexity and Representation Learning of Low Rank MDPs. *Advances in Neural Information Processing Systems*, 33:20095–20107, 2020b.

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *The Journal of Machine Learning Research*, 22(1):4431–4506, 2021.

Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: Worst-case regret bounds. In *Advances in Neural Information Processing Systems*, pp. 1184–1194, 2017.

Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

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

Dilip Arumugam and Satinder Singh. Planning to the Information Horizon of BAMDPs via Epistemic State Abstraction. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Dilip Arumugam and Benjamin Van Roy. Deciding What to Learn: A Rate-Distortion Approach. In *International Conference on Machine Learning*, pp. 373–382. PMLR, 2021a.

Dilip Arumugam and Benjamin Van Roy. The Value of Information When Deciding What to Learn. *Advances in Neural Information Processing Systems*, 34:9816–9827, 2021b.

Dilip Arumugam and Benjamin Van Roy. Deciding What to Model: Value-Equivalent Sampling for Reinforcement Learning. *Advances in Neural Information Processing Systems*, 35, 2022.

Dilip Arumugam, Peter Henderson, and Pierre-Luc Bacon. An Information-Theoretic Perspective on Credit Assignment in Reinforcement Learning. *arXiv preprint arXiv:2103.06224*, 2021.

Dilip Arumugam, Mark K. Ho, Noah D. Goodman, and Benjamin Van Roy. On Rate-Distortion Theory in Capacity-Limited Cognition & Reinforcement Learning. In *NeurIPS 2022 Workshop on Information-Theoretic Principles in Cognitive Systems*, 2022.

Dilip Arumugam, Mark K Ho, Noah D Goodman, and Benjamin Van Roy. Bayesian Reinforcement Learning with Limited Cognitive Load. *Open Mind*, 8:395–438, 2024.

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*, pp. 19–26, 2009.

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

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

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

Bilal A Bari and Samuel J Gershman. Undermatching is a consequence of policy compression. *bioRxiv*, 2022.

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*, pp. 35–42, 2009.

Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The Arcade Learning Environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

Marc G Bellemare, Salvatore Candido, Pablo Samuel Castro, Jun Gong, Marlos C Machado, Subhodeep Moitra, Sameera S Ponda, and Ziyu Wang. Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature*, 588(7836):77–82, 2020.

Richard Bellman. A Markovian Decision Process. *Journal of Mathematics and Mechanics*, pp. 679–684, 1957.

Richard Bellman and Robert Kalaba. On Adaptive Control Processes. *IRE Transactions on Automatic Control*, 4(2):1–9, 1959.

Toby Berger. *Rate Distortion Theory: A Mathematical Basis for Data Compression.* Prentice-Hall, 1971.

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

Rahul Bhui, Lucy Lai, and Samuel J Gershman. Resource-rational decision making. *Current Opinion in Behavioral Sciences*, 41:15–21, 2021.

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

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

Matthew Botvinick, Ari Weinstein, Alec Solway, and Andrew Barto. Reinforcement learning, efficient coding, and the statistics of natural tasks. *Current opinion in behavioral sciences*, 5:71–77, 2015.

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.

Stephen P. Boyd and Lieven Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

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.

Vanessa M Brown, Michael N Hallquist, Michael J Frank, and Alexandre Y Dombrovski. Humans adaptively resolve the explore-exploit dilemma under cognitive constraints: Evidence from a multi-armed bandit task. *Cognition*, 229:105233, 2022.

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.

Nuttapong Chentanez, Andrew Barto, and Satinder Singh. Intrinsically motivated reinforcement learning. *Advances in neural information processing systems*, 17, 2004.

Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. Minimalistic gridworld environment for gymnasium, 2018. URL https://github.com/Farama-Foundation/Minigrid.

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

Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

James P Crutchfield and David P Feldman. Synchronizing to the environment: Information-theoretic constraints on agent learning. *Advances in Complex Systems*, 4(02n03):251–264, 2001.

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

Imre Csiszár. On an extremum problem of information theory. *Studia Scientiarum Mathematicarum Hungarica*, 9, 1974b.

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*, pp. 2818–2826, 2015.

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*, pp. 5717–5727, 2017.

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*, pp. 1429–1439, 2018.

Christoph Dann, Mehryar Mohri, Tong Zhang, and Julian Zimmert. A Provably Efficient Model-Free Posterior Sampling Method for Episodic Reinforcement Learning. *Advances in Neural Information Processing Systems*, 34:12040–12051, 2021.

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

Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

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

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*, pp. 1665–1674. PMLR, 2019.

John C. Duchi. *Lecture Notes for Statistics 311/Electrical Engineering 377.* Stanford University, 2021.

Michael O'Gordon Duff. *Optimal Learning: Computational Procedures for Bayes-adaptive Markov Decision Processes.* University of Massachusetts Amherst, 2002.

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*, pp. 1–50, 2021.

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

Vikranth Dwaracherla, Zheng Wen, Ian Osband, Xiuyuan Lu, Seyed Mohammad Asghari, and Benjamin Van Roy. Ensembles for Uncertainty Estimation: Benefits of Prior Functions and Bootstrapping. *arXiv preprint arXiv:2206.03633*, 2022.

Alhussein Fawzi, Matej Balog, Aja Huang, Thomas Hubert, Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Francisco J R Ruiz, Julian Schrittwieser, Grzegorz Swirszcz, et al. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610(7930):47–53, 2022.

Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The Statistical Complexity of Interactive Decision Making. *arXiv preprint arXiv:2112.13487*, 2021.

Roy Fox, Ari Pakman, and Naftali Tishby. Taming the noise in reinforcement learning via soft updates. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, pp. 202–211, 2016.

Alexandre Galashov, Siddhant M Jayakumar, Leonard Hasenclever, Dhruva Tirumala, Jonathan Schwarz, Guillaume Desjardins, Wojciech M Czarnecki, Yee Whye Teh, Razvan Pascanu, and Nicolas Heess. Information asymmetry in KL-regularized RL. In *International Conference on Learning Representations*, 2019.

Samuel J Gershman. Origin of perseveration in the trade-off between reward and complexity. *Cognition*, 204:104394, 2020.

Samuel J Gershman. The rational analysis of memory. *Oxford Handbook of Human Memory.*, 2021.

Samuel J Gershman and Lucy Lai. The reward-complexity trade-off in schizophrenia. *bioRxiv*, 2020.

Samuel J Gershman, Eric J Horvitz, and Joshua B Tenenbaum. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245):273–278, 2015.

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.

Gerd Gigerenzer and Daniel G Goldstein. Reasoning the fast and frugal way: models of bounded rationality. *Psychological review*, 103(4):650, 1996.

Anirudh Goyal, Riashat Islam, DJ Strouse, Zafarali Ahmed, Hugo Larochelle, Matthew Botvinick, Yoshua Bengio, and Sergey Levine. InfoBot: Transfer and Exploration via the Information Bottleneck. In *International Conference on Learning Representations*, 2018.

Anirudh Goyal, Yoshua Bengio, Matthew Botvinick, and Sergey Levine. The Variational Bandwidth Bottleneck: Stochastic Evaluation on an Information Budget. In *International Conference on Learning Representations*, 2020.

Robert M. Gray. *Entropy and Information Theory*. Springer Science & Business Media, 2011.

Thomas L Griffiths, Falk Lieder, and Noah D Goodman. Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in cognitive science*, 7(2): 217–229, 2015.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1352–1361. JMLR. org, 2017.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pp. 1861–1870, 2018.

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

Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, pp. 2681–2691. PMLR, 2019.

Mark K Ho, David Abel, Jonathan D Cohen, Michael L Littman, and Thomas L Griffiths. The efficiency of human cognition reflects planned information processing. In *34th AAAI Conference on Artificial Intelligence, AAAI 2020*, pp. 1300–1307. AAAI press, 2020.

Mark K Ho, David Abel, Carlos G Correa, Michael L Littman, Jonathan D Cohen, and Thomas L Griffiths. People construct simplified mental representations to plan. *Nature*, 606(7912):129–136, 2022.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. VIME: Variational information maximizing exploration. *Advances in neural information processing systems*, 29, 2016.

Kazunori Iwata, Kazushi Ikeda, and Hideaki Sakai. A new criterion using information gain for action selection strategy in reinforcement learning. *IEEE Transactions on Neural Networks*, 15 (4):792–799, 2004.

Anthony MV Jakob and Samuel J Gershman. Rate-distortion theory of neural coding and its implications for working memory. *bioRxiv*, 2022.

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

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*, pp. 1704–1713. PMLR, 2017.

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*, pp. 4868–4878, 2018.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020.

Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman Eluder Dimension: New rich classes of RL problems, and sample-efficient algorithms. *Advances in Neural Information Processing Systems*, 34:13406–13418, 2021.

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.

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

Cezary Kaliszyk, Josef Urban, Henryk Michalewski, and Miroslav Olšák. Reinforcement learning of theorem proving. *Advances in Neural Information Processing Systems*, 31, 2018.

Hilbert J Kappen, Vicenç Gómez, and Manfred Opper. Optimal control as a graphical model inference problem. *Machine learning*, 87(2):159–182, 2012.

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

Alexander S Klyubin, Daniel Polani, and Chrystopher L Nehaniv. Empowerment: A universal agent-centric measure of control. In *2005 ieee congress on evolutionary computation*, volume 1, pp. 128–135. IEEE, 2005.

Alexander S Klyubin, Daniel Polani, and Chrystopher L Nehaniv. Keep your options open: An information-based driving principle for sensorimotor systems. *PloS one*, 3(12):e4018, 2008.

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

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

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*, pp. 1848–1856, 2016.

Lucy Lai and Samuel J Gershman. Policy compression: An information bottleneck in action selection. In *Psychology of Learning and Motivation*, volume 74, pp. 195–232. Elsevier, 2021.

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

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2017.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Rachel A. Lerch and Chris R. Sims. Policy generalization in capacity-limited reinforcement learning, 2018.

Rachel A. Lerch and Chris R. Sims. Rate-distortion theory and computationally rational reinforcement learning. *Proceedings of Reinforcement Learning and Decision Making (RLDM)*, pp. 7–10, 2019.

Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.

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

Falk Lieder and Thomas L Griffiths. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, 43, 2020.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations*, 2016.

Bo Liu and Sridhar Mahadevan. Compressive reinforcement learning with oblique random projections. Technical report, Citeseer, 2011.

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

Xiuyuan Lu, Benjamin Van Roy, Vikranth Dwaracherla, Morteza Ibrahimi, Ian Osband, Zheng Wen, et al. Reinforcement Learning, Bit by Bit. *Foundations and Trends® in Machine Learning*, 16(6): 733–865, 2023.

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

John G Mikhael, Lucy Lai, and Samuel J Gershman. Rational inattention and tonic dopamine. *PLoS computational biology*, 17(3):e1008659, 2021.

Azalia Mirhoseini, Anna Goldie, Mustafa Yazgan, Joe Wenjie Jiang, Ebrahim Songhori, Shen Wang, Young-Joon Lee, Eric Johnson, Omkar Pathak, Azade Nazi, et al. A graph placement methodology for fast chip design. *Nature*, 594(7862):207–212, 2021.

Sanjoy Mitter and Anant Sahai. Information and control: Witsenhausen revisited. In *Learning, Control and Hybrid Systems*, pp. 281–293. Springer, 1999.

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

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Shakir Mohamed and Danilo Jimenez Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. *Advances in neural information processing systems*, 28, 2015.

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*, pp. 2505–2508. IEEE, 2009.

Allen Newell, John Calman Shaw, and Herbert A Simon. Elements of a theory of human problem solving. *Psychological review*, 65(3):151, 1958.

Allen Newell, Herbert Alexander Simon, et al. *Human problem solving*, volume 104. Prentice-hall Englewood Cliffs, NJ, 1972.

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

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

Brendan O'Donoghue, Ian Osband, and Catalin Ionescu. Making Sense of Reinforcement Learning and Probabilistic Inference. In *International Conference on Learning Representations*, 2020.

Pedro A Ortega and Daniel A Braun. Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 469(2153):20120683, 2013.

Pedro Alejandro Ortega and Daniel Alexander Braun. Information, utility and bounded rationality. In *International Conference on Artificial General Intelligence*, pp. 269–274. Springer, 2011.

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

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

Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *International Conference on Machine Learning*, pp. 2701–2710. PMLR, 2017.

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.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via Bootstrapped DQN. In *Advances in Neural Information Processing Systems*, pp. 4026–4034, 2016a.

Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pp. 2377–2386, 2016b.

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

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.

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

Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Morteza Ibrahimi, Xiuyuan Lu, and Benjamin Van Roy. Approximate Thompson Sampling via Epistemic Neural Networks. In *Uncertainty in Artificial Intelligence*, pp. 1586–1595. PMLR, 2023.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

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

Naama Parush, Naftali Tishby, and Hagai Bergman. Dopaminergic balance between reward maximization and policy complexity. *Frontiers in Systems Neuroscience*, 5:22, 2011.

Lin Peng. Learning with information capacity constraints. *Journal of Financial and Quantitative Analysis*, 40(2):307–329, 2005.

Daniel Polani. Information: currency of life? *HFSP journal*, 3(5):307–316, 2009.

Daniel Polani, Thomas Martinetz, and Jan Kim. An information-theoretic approach for the quantification of relevance. In *Advances in Artificial Life: 6th European Conference, ECAL 2001 Prague, Czech Republic, September 10–14, 2001 Proceedings 6*, pp. 704–713. Springer, 2001.

Yury Polyanskiy and Yihong Wu. *Information Theory: From Learning to Coding.* Cambridge University Press, 2022.

Mariya Popova, Olexandr Isayev, and Alexander Tropsha. Deep reinforcement learning for de novo drug design. *Science advances*, 4(7):eaap7885, 2018.

Pascal Poupart and Craig Boutilier. Value-directed compression of POMDPs. *Advances in neural information processing systems*, 15, 2002.

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

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

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

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

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

Daniel Russo and Benjamin Van Roy. Satisficing in time-sensitive bandit learning. *arXiv preprint arXiv:1803.02855*, 2018b.

Daniel Russo and Benjamin Van Roy. Satisficing in time-sensitive bandit learning. *Mathematics of Operations Research*, 2022.

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

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.

Andrei A Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy Distillation. *arXiv preprint arXiv:1511.06295*, 2015.

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

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.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*, 2016.

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.

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

Tianlin Shi, Andrej Karpathy, Linxi Fan, Jonathan Hernandez, and Percy Liang. World of bits: An open-domain platform for web-based agents. In *International Conference on Machine Learning*, pp. 3135–3144. PMLR, 2017.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Herbert A Simon. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69 (1):99–118, 1955.

Herbert A Simon. Rational choice and the structure of the environment. *Psychological review*, 63 (2):129, 1956.

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

Chris R Sims. Rate–distortion theory and human perception. *Cognition*, 152:181–198, 2016.

Chris R Sims. Efficient coding explains the universal law of generalization in human perception. *Science*, 360(6389):652–656, 2018.

Christopher A Sims. Implications of rational inattention. *Journal of Monetary Economics*, 50(3): 665–690, 2003.

Satinder Singh, Richard L Lewis, and Andrew G Barto. Where do rewards come from? In *Proceedings of the annual conference of the cognitive science society*, pp. 2601–2606. Cognitive Science Society, 2009.

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*, pp. 564–571, 2010.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

Susanne Still. Information-theoretic approach to interactive learning. *Europhysics Letters*, 85(2): 28005, 2009.

Susanne Still and Doina Precup. An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131(3):139–148, 2012.

Jan Storck, Sepp Hochreiter, Jürgen Schmidhuber, et al. Reinforcement driven information acquisition in non-deterministic environments. In *Proceedings of the International Conference on Artificial Neural Networks, Paris*, volume 2, pp. 159–164, 1995.

Alexander L Strehl and Michael L Littman. An Analysis of Model-Based Interval Estimation for Markov Decision Processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

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.

Malcolm JA Strens. A Bayesian framework for reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 943–950, 2000.

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*, pp. 2898–2933. PMLR, 2019.

Yi Sun, Faustino J Gomez, and Jürgen Schmidhuber. Planning to be surprised: Optimal Bayesian exploration in dynamic environments. In *AGI*, pp. 41–51. Springer, 2011.

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

Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285, 2011.

Gerald Tesauro. Temporal difference learning and TD-Gammon. *Communications of the ACM*, 38 (3):58–68, 1995.

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.

Dhruva Tirumala, Hyeonwoo Noh, Alexandre Galashov, Leonard Hasenclever, Arun Ahuja, Greg Wayne, Razvan Pascanu, Yee Whye Teh, and Nicolas Heess. Exploiting hierarchy for learning and transfer in KL-regularized RL. *arXiv preprint arXiv:1903.07438*, 2019.

Naftali Tishby and Daniel Polani. Information theory of decisions and actions. In *Perception-action cycle*, pp. 601–636. Springer, 2011.

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

Emanuel Todorov. Linearly-solvable Markov decision problems. In *Advances in neural information processing systems*, pp. 1369–1376, 2007.

Marc Toussaint. Robot trajectory optimization using approximate inference. In *Proceedings of the 26th annual international conference on machine learning*, pp. 1049–1056, 2009.

Sander van Dijk, Daniel Polani, and Chrystopher L Nehaniv. Hierarchical behaviours: Getting the most bang for your bit. *Lecture Notes in Computer Science*, 2011.

Sander G van Dijk and Daniel Polani. Look-ahead relevant information: Reducing cognitive burden over prolonged tasks. In *2011 IEEE Symposium on Artificial Life (ALIFE)*, pp. 46–53. IEEE, 2011.

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

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.

Edward Vul, Noah Goodman, Thomas L Griffiths, and Joshua B Tenenbaum. One and done? Optimal decisions from very few samples. *Cognitive Science*, 38(4):599–637, 2014.

Tingwu Wang, Xuchan Bao, Ignasi Clavera, Jerrick Hoang, Yeming Wen, Eric Langlois, Shunshi Zhang, Guodong Zhang, Pieter Abbeel, and Jimmy Ba. Benchmarking model-based reinforcement learning. *arXiv preprint arXiv:1907.02057*, 2019.

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

Aaron Wyner and Jacob Ziv. The rate-distortion function for source coding with side information at the decoder. *IEEE Transactions on Information Theory*, 22(1):1–10, 1976.

Shunyu Yao, Howard Chen, John Yang, and Karthik R Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. In *Advances in Neural Information Processing Systems*, 2022.

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

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*, pp. 7304–7312. PMLR, 2019.

Alexandre Zenon, Oleg Solopchuk, and Giovanni Pezzulo. An information-theoretic perspective on the costs of cognition. *Neuropsychologia*, 123:5–18, 2019.

Zhenpeng Zhou, Steven Kearnes, Li Li, Richard N Zare, and Patrick Riley. Optimization of molecules via deep reinforcement learning. *Scientific reports*, 9(1):1–10, 2019.

Brian D Ziebart. *Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy*. PhD thesis, Carnegie Mellon University, 2010.

A Preliminaries

In this section, we provide brief background on information theory, rate-distortion theory, and details on our notation. All random variables are defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. 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 functions mapping from $\mathcal{X}$ to $\mathcal{Y}$ as $\{\mathcal{X} \to \mathcal{Y}\} \triangleq \{f \mid f : \mathcal{X} \to \mathcal{Y}\}$.

Here we introduce various concepts in probability theory and information theory used throughout this paper. We encourage readers to consult (Cover & Thomas, 2012; Gray, 2011; Duchi, 2021; Polyanskiy & Wu, 2022) for more background. We define the mutual information between any two random variables X, Y through the Kullback-Leibler (KL) divergence:

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

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. We define the entropy and conditional entropy for any two random variables X, Y as $\mathbb{H}(X) = \mathbb{I}(X;X)$ and $\mathbb{H}(Y \mid X) = \mathbb{H}(Y) - \mathbb{I}(X;Y)$, respectively. This yields the following identity for the conditional mutual information of any three arbitrary random variables X, Y, and Z: $\mathbb{I}(X;Y|Z) = \mathbb{H}(X|Z) - \mathbb{H}(X \mid Y, Z) = \mathbb{H}(Y|Z) - \mathbb{H}(Y|X,Z)$. Through the chain rule of the KL-divergence, we obtain the chain rule of mutual information: $\mathbb{I}(X;Y_1, \ldots, Y_n) = \sum\limits_{i=1}^{n} \mathbb{I}(X;Y_i \mid Y_1, \ldots, Y_{i-1})$.

Here we offer a high-level overview of rate-distortion theory (Shannon, 1959; Berger, 1971) and encourage readers to consult (Cover & Thomas, 2012) for more details. 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

$$\mathcal{R}(D) = \inf_{Z \in \mathcal{Z}} \mathbb{I}(X;Z) \text{ such that } \mathbb{E}\left[d(X,Z)\right] \leq D,$$

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 and, conveniently, is well-defined for abstract information source and channel output random variables (Csiszár, 1974b). Moreover, the rate-distortion function has useful structural properties:

Fact 1. *$\mathcal{R}(D)$ is a non-negative, convex, and non-increasing function of $D \in \mathbb{R}_{\geq 0}$.*

B Algorithms

Here we present the algorithm introduced in the main body of the paper.

Algorithm 1 Blahut-Arimoto Randomized Value Functions

Input: Prior $\mathbb{P}(Q^\star \in \cdot \mid H_1)$, Lagrange multiplier $\beta \in \mathbb{R}_{\geq 0}$, Posterior samples $Z \in \mathbb{N}$
for $k \in [K]$ **do**
  Draw posterior sample $Q_1 \sim \mathbb{P}(Q^\star \in \cdot \mid H_k)$
  for $t = 1, 2, \ldots,$ **do**
    Observe current state s_t and draw samples $Q_2, \ldots, Q_Z \overset{i.i.d.}{\sim} \mathbb{P}(Q^\star \in \cdot \mid H_k)$
    Compute distortions $d_{s_t}(Q_z, \tilde{a})$, $\forall z \in [Z], \tilde{a} \in \mathcal{A}$
    Run Blahut-Arimoto algorithm with β to compute $\mathbb{P}(\tilde{A} \in \cdot \mid Q^\star)$ achieving $\mathcal{R}_k(s_t, D)$ limit
    Sample action a_t from policy $\pi^{(k)}(a_t \mid s_t) = \mathbb{P}(\tilde{A} = a_t \mid Q^\star = Q_1)$
  end for
  $H_{k+1} = H_k \cup \{\tau_k\}$ and update posterior $\mathbb{P}(Q^\star \in \cdot \mid H_{k+1})$
end for

C Related Work

Overall, this paper touches upon two rich veins of prior work in the reinforcement-learning literature: (1) provably-efficient reinforcement learning and (2) information-theoretic reinforcement learning. While there are numerous works that fall under each of these areas, we isolate a relevant, comprehensive subset below to allow for a suitably clear juxtaposition with our approach.

On the side of provably-efficient reinforcement learning, there are various approaches that have been developed over the course of the last two decades (Kearns & Singh, 2002; Brafman & Tennenholtz, 2002; Kakade, 2003; Auer et al., 2009; Bartlett & Tewari, 2009; Strehl et al., 2009; Jaksch et al., 2010; Osband et al., 2013; Dann & Brunskill, 2015; Krishnamurthy et al., 2016; Osband & Van Roy, 2017; Azar et al., 2017; Dann et al., 2017; Agrawal & Jia, 2017; Jiang et al., 2017; Jin et al., 2018; Dann et al., 2018; Zanette & Brunskill, 2019; Du et al., 2019; Sun et al., 2019; Agarwal et al., 2020b;a; Jin et al., 2020; Dong et al., 2021; Agarwal et al., 2021; Jin et al., 2021; Foster et al., 2021; Lu et al., 2023) which vary both along the type of analyses and guarantees (bounds on PAC-MDP sample complexity, regret, or iteration complexity) as well as the underlying structural assumptions leveraged to obtain those guarantees (tabular MDPs, linear function approximation, low-rank transition structure, bounded Bellman rank, *etc.*). Due to our information-theoretic analysis, our results are quite general and can be applied to any MDP without structural assumptions such as finiteness of the state-action space. Additionally, our guarantees are obtained through a regret-analysis, although translations of these bounds to corresponding PAC-MDP bounds (Kakade, 2003; Strehl et al., 2009) may be feasible (Dann et al., 2017; Jin et al., 2018).

Within this narrowed field of view, methods in this space can largely be segregated according to their use of optimism in the face of uncertainty or posterior sampling as the tool of choice for handling the exploration challenge, though recent work has considered blending ideas from both regimes (Dann et al., 2021; Agarwal & Zhang, 2022); while both algorithm classes are theoretically sound and empirically effective, there are cases where the latter Bayesian reinforcement learning methods can be more favorable both in theory (Osband & Van Roy, 2017) as well as in practice (Osband et al., 2016a; 2018). While some of the former optimism-based methods admit high-probability sample-complexity guarantees that depend on a sub-optimality parameter, the corresponding agents are designed to pursue optimal policies; in contrast, a core premise of this work is that an agent designer aware of a preference for satisficing behaviors over optimal ones can embed such considerations explicitly into the design of the agent which can, in turn, make fundamentally different exploratory choices during learning based on the objective of satisficing, rather than purely optimizing.

Traditionally, the Bayesian reinforcement learning setting (Bellman & Kalaba, 1959; Duff, 2002; Ghavamzadeh et al., 2015) employed by posterior-sampling methods relies on the Bayes-Adaptive MDP (BAMDP) formulation (Duff, 2002), which is notoriously intractable even in the tabular setting (see the discussion of Arumugam & Singh (2022) for some conditions under which BAMDP planning may be resolved efficiently). The main innovation behind posterior-sampling methods (Strens, 2000) is a lazy updating of the agent's epistemic uncertainty, removing the aforementioned intractability. Alternatively, one can consult Ghavamzadeh et al. (2015) for details on methods that employ other approximation techniques or adhere to the PAC-BAMDP notion of efficient Bayesian reinforcement learning (Kolter & Ng, 2009; Asmuth et al., 2009; Sorg et al., 2010).

The core mechanism that underlies many posterior-sampling algorithms is Thompson sampling (Thompson, 1933; Russo et al., 2018) which, by design, is exclusively focused on obtaining optimal solutions. Even subsequent improvements to Thompson sampling that account for information gain (Russo & Van Roy, 2014; 2018a) also retain this property. Lu et al. (2023) introduce the idea of a *learning target* as a mechanism through which an agent may prioritize its exploration when the underlying environment is too immense and complex for the agent to be endlessly curious and pursue all available bits of information; crucially, however, their regret bounds presume that such a learning target has been computed *a priori*. In contrast, the agents discussed in this work adaptively compute and refine the learning target as the agent's knowledge of the underlying environment accumulates. Most related to the present work are the algorithms of Arumugam & Van Roy (2021a;b) and Arumugam & Van Roy (2022), where the former algorithms employ learning targets adaptively computed via rate-distortion theory exclusively to multi-armed bandit problems while the latter translates these ideas over to PSRL, but without a concrete roadmap to a computationally-feasible instantiation. Our work remedies this final issue by maintaining and resolving epistemic uncertainty over the optimal action-value function, rather than the underlying model of the unknown MDP.

Setting aside work on provably-efficient reinforcement learning with guarantees obtained via information-theoretic analyses, the topic of information-theoretic reinforcement learning is largely an empirical body of work, focusing on how information-theoretic quantities may be practically applied by decision-making agents to address the fundamental challenges of generalization, exploration, and credit assignment. As the algorithms of this paper are primarily designed to address the challenge of exploration, we only mention in passing that work coupling information theory to address the challenges of temporal credit assignment is nascent (van Dijk et al., 2011; Arumugam et al., 2021). Meanwhile, there is a substantial literature on information-theoretic methods to aid generalization, largely situated around the information bottleneck principle (Tishby et al., 2000), which instantiates a particular rate-distortion optimization to formalize the notion of a learned data representation that is both maximally compressive while also retaining the requisite information needed for task performance. This perspective leads to a information-theoretic formulation of classic state abstraction in reinforcement learning (Li et al., 2006; Van Roy, 2006; Abel et al., 2016; Abel, 2020) for lossy compression of the original MDP state space (Liu & Mahadevan, 2011; Shafieepoorfard et al., 2016; Lerch & Sims, 2018; 2019; Abel et al., 2019). A similar but distinct problem also manifests in the partially-observable MDP (POMDP) (Åström, 1965; Kaelbling et al., 1998) setting where classic work in control theory models observational capacity limitations as an information-theoretic rate constraint (Witsenhausen, 1971; Mitter & Sahai, 1999; Mitter, 2001; Borkar et al., 2001; Crutchfield & Feldman, 2001; Poupart & Boutilier, 2002; Tatikonda & Mitter, 2004; Kostina & Hassibi, 2019) and asks how well one can control a system subject to such a rate limit.

Aligned with the perspective of this work, the bulk of the information-theoretic reinforcement-learning literature is aimed at addressing the challenge of exploration, logically expecting that novel information can serve as a useful form of intrinsic motivation (Chentanez et al., 2004; Singh et al., 2009) to guide the agent (Storck et al., 1995; Polani et al., 2001; Iwata et al., 2004; Klyubin et al., 2005; Todorov, 2007; Klyubin et al., 2008; Still, 2009; Polani, 2009; Ortega & Braun, 2011; van Dijk & Polani, 2011; Still & Precup, 2012; Tishby & Polani, 2011; Sun et al., 2011; Rubin et al., 2012; Ortega & Braun, 2013; Mohamed & Jimenez Rezende, 2015; Houthooft et al., 2016; Goyal et al., 2018; 2020). Most notable among these methods are those that have inspired popular modern deep reinforcement-learning algorithms through the "control as inference" or KL-regularized reinforcement learning perspective (Toussaint, 2009; Kappen et al., 2012; Levine, 2018; Ziebart, 2010; Fox et al., 2016; Haarnoja et al., 2017; 2018; Galashov et al., 2019; Tirumala et al., 2019). We refer readers to the short survey of Arumugam et al. (2022) for an overview of how the principled Bayesian methods used in this work compare relative to these latter methods inspired by the information bottleneck principle though, in short, the fundamental issue boils down to a lack of properly alleviating the burdens of exploration (O'Donoghue et al., 2020); that said, guarantees do exist for such maximum entropy exploration schemes in the absence of reward (Hazan et al., 2019) although, translated into our real-world reinforcement learning setting where agents are fundamentally bounded, naively attempting to maximize entropy would be an ill-defined objective as all bits of information that could be acquired from the environment cannot be retained within the agent's capacity limitations, by assumption.

Finally, we conclude by noting that the practical deep reinforcement-learning algorithm developed in this work relies on the classic Blahut-Arimoto algorithm (Blahut, 1972; Arimoto, 1972) for computing the rate-distortion function when both the information source and channel output random variables are discrete. While the algorithm is known to be theoretically-sound in general (Csiszár, 1974b) and globally-convergent (Csiszár, 1974a) under these conditions, various techniques have been developed in order to accelerate the Blahut-Arimoto algorithm and make it applicable to continuous information sources (Boukris, 1973; Rose, 1994; Sayir, 2000; Matz & Duhamel, 2004; Chiang & Boyd, 2004; Dauwels, 2005; Niesen et al., 2007; Vontobel et al., 2008; Naja et al., 2009; Yu, 2010). While we do not explore any of these extensions here, the reality that our proposed deep reinforcement-learning agent runs the Blahut-Arimoto algorithm on a per-timestep basis suggests that these works could serve as a useful basis for future work which more carefully studies large-scale deployment of our agent.

D Additional Minigrid Experiment Details

Algorithm Implementations

Hyperparameters for DQN	
Parameter	Value
discount (γ)	0.99
number of frames	$100,000$
optimizer	Adam
learning rate	$5 \cdot 10^{-4}$
replay buffer capacity	$25,000$
τ (used for soft target net updates)	$1 \cdot 10^{-3}$
batch size	128
max ϵ	1.0
min ϵ	0.0
warmup	100 frames
ϵ decay rate	$\frac{1}{0.95 \cdot (\text{number of frames} - \text{warmup})}$

Table 1: DQN Hyperparameters.

Hyperparameters for RVF	
Parameter	Value
discount (γ)	0.99
number of frames	$100,000$
optimizer	Adam
learning rate	$5 \cdot 10^{-4}$
replay buffer capacity	$25,000$
τ (used for soft target net updates)	$1 \cdot 10^{-3}$
batch size	128
index dimension	30
noise scale	0.1
prior scale	0.1

Table 2: RVF Hyperparameters.

DQN

We implement the DQN agent as a convolutional neural network with the following architecture:

1. Conv Layer: Kernel Size = 2×2, output channels = 16, RELU activation
2. Max Pool 2×2
3. Conv Layer: Kernel Size = 2×2, output channels = 16, RELU activation
4. Conv Layer: Kernel Size = 2×2, output channels = 64, RELU activation
5. Fully-connected layer, output dimension = 128, RELU activation
6. Fully-connected layer, output dimension = 3 (number of actions)

RVF & Blahut-Arimoto RVF

We implement RVF, and consequently Blahut-Arimoto RVF, as an epistemic neural network comprised of a base network and an epinet Osband et al. (2021). For the base network, we use the same architecture as the DQN network. The epinet consists of a learnable network and a prior network and we guide our design choices around both components based on details reported by Osband et al. (2021).

In order to reliably induce a distribution over functions, an epinet operates with an epistemic index $z \sim \mathcal{N}(0, I_d) \in \mathbb{R}^d$, where the index dimension $d \in \mathbb{N}$ is a hyperparameter. Intuitively, it is the stochasticity in the underlying reference distribution of z that produces the stochasticity needed to represent a full distribution over functions. Both the learnable network and the prior network take as input a single sampled epistemic index concatenated with the embedding obtained by the convolutional layers of the base network as well as the penultimate layer activations of the base network. These last two components are first passed through a stop-gradient layer to avoid propagation of epinet gradients into the base network.

For a finite action space $|\mathcal{A}| < \infty$, both the base network and the learnable network produce outputs in $\mathbb{R}^{|\mathcal{A}|}$. Meanwhile, the prior network is itself a small ensemble of MLPs where the size of the ensemble is directly determined by the index dimension d. Consequently, the output of the entire prior network is determined by first feeding the inputs through all elements of the ensemble, stacking them together into a $|\mathcal{A}| \times d$ matrix, and then finally taking the matrix-vector product with the epistemic index $z \in \mathbb{R}^d$ to obtain the action-values in $\mathbb{R}^{|\mathcal{A}|}$. The base network and learnable network output vectors are immediately added together along with the prior network output after it has been multiplied by a prior scale parameter.

As noted in Osband et al. (2018); Dwaracherla et al. (2022), the prior scale hyperparameter effectively controls the effective width or support of the initial distribution over value functions; large settings of this parameter will yield more variance between functions sampled through different epistemic indices but, simultaneously, will require prolonged interaction and stochastic gradient updates for the learnable network to compensate for the scaled prior effect. Our loss function for optimizing the epinet parameters consists of the standard TD(0)-error accompanied by Gaussian noise (Osband et al., 2016a; 2019) where the standard deviation of the zero-mean noise is also a hyperparameter, denoted as the noise scale, which aligns with Bayesian linear regression. The resulting additive noise bonus is then computed during loss computations by taking the inner product between sampled noise variables and the epistemic indices in each minibatch. We encourage readers to consult (Osband et al., 2019; 2021) for more details on epinets and inducing a distribution over optimal action-value functions in this manner.

We use the following prior network architecture:

1. Fully-connected layer: output dimension = 5, RELU activation
2. Fully-connected layer: output dimension = 5, RELU activation
3. Fully-connected layer: output dimension = 3 (number of actions)

The learnable network architecture is as follows:

1. Fully-connected layer: output dimension = 256, RELU activation
2. Fully-connected layer: output dimension = 256, RELU activation
3. Fully-connected layer: output dimension = $3 \cdot 5$ (number of actions $*$ index dimension)

The output from the epistemic network is computed as

$$\text{base network output} + \text{learnable epinet output} + \text{prior scale} \cdot \text{prior network output}$$

Hyper-parameter Selection

On each environment, we performed a grid search over the prior scale and noise scale hyper-parameters of the RVF agent using a single seed. Intuitively, tuning these parameters is commensurate with ensuring that the true optimal action-value function of each domain lies in the support of the prior distribution represented by the corresponding epinet and that there is enough signal to facilitate deep exploration throughout the domain without compromising convergence. Specifically we performed a grid search over all combinations of noise scale $\in \{0.1, 0.15, 0.2\}$ and prior scale $\in \{0.025, 0.05, 0.1, 0.25\}$. On both domains, we found noise scale = 0.1 and prior scale = 0.1 to yield the highest average return after 100,000 frames. As these hyperparameters govern the quality of how epistemic uncertainty is represented, but not how each algorithm utilizes that uncertainty to resolve exploration, we keep these tuned values fixed for all RVF and Blahut-Arimoto RVF experiments.

Blahut-Arimoto RVF consumes as input the Lagrange multiplier $\beta \in \mathbb{R}_{\geq 0}$ which controls the trade-off between rate and distortion. To determine which range of β values covers the spectrum of minimizing rate to minimizing distortion, which can differ between environments, we first trained Blahut-Arimoto RVF with $\beta \in \{1, 10, 100, 1000, 10000, 100000, 1000000\}$. On MiniGrid-Empty-16x16-v0, the performance of Blahut-Arimoto RVF with $\beta = 1000000$ surpassed that of RVF, whereas on MiniGrid-CorridorEnv-v0 this occured with a much smaller value of $\beta = 100$. Consequently, we used 1000000 and 100 as the upper bound on β values to select for training Blahut-Arimoto RVF on MiniGrid-Empty-16x16-v0 and MiniGrid-CorridorEnv-v0, respectively.

Further hyper-parameter details for all agents are in Tables 1 and 2.

E Compute Details

For our experiments, we used an n1-standard-8 Google Cloud Virtual Machine with 1 NVIDIA Tesla P100 GPU.