On Rate-Distortion Theory in Capacity-Limited Cognition & Reinforcement Learning

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

Mark K. Ho  
Department of Computer Science  
Princeton University  
mho@princeton.edu

Noah D. Goodman  
Department of Psychology  
Department of Computer Science  
Stanford University  
ngoodman@stanford.edu

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

Abstract

Throughout the cognitive-science literature, there is widespread agreement that decision-making agents operating in the real world do so under limited information-processing capabilities and without access to unbounded cognitive or computational resources. Prior work has drawn inspiration from this fact and leveraged an information-theoretic model of such behaviors or policies as communication channels operating under a bounded rate constraint. Meanwhile, a parallel line of work also capitalizes on the same principles from rate-distortion theory to formalize capacity-limited decision making through the notion of a learning target, which facilitates Bayesian regret bounds for provably-efficient learning algorithms. In this paper, we aim to elucidate this latter perspective by presenting a brief survey of these information-theoretic models of capacity-limited decision making in biological and artificial agents.

1 Motivation

In a perfect world, decision makers of any shape or form would always be capable of producing optimal behaviors. In reality, however, the inescapable constraints of an overwhelmingly-complex environment force agents to seek out alternative behaviors that are sufficiently satisfying or, more succinctly, satisficing. While satisficing is a longstanding, well-studied idea about how to understand resource-limited cognition [Simon, 1955, 1956, Newell et al., 1958, Newell and Simon, 1972, Simon, 1982], it has been usually treated as contrary to rational analysis [Anderson, 1990]. In particular, modern resource-rational analyses, both utility-theoretic [Griffiths et al., 2015, Lieder and Griffiths, 2020] and information-theoretic [Sims, 2003, 2016, Gershman, 2020], still aim to find an optimal policy but to do so within resource constraints. An alternative view for the resource-rational learning of satisficing agents comes from recent theoretical work in bandit learning and reinforcement learning [Arumugam and Van Roy, 2021a, 2022b]; this perspective focuses on learning to achieve a deliberately-sub-optimal, satisficing policy that requires obtaining fewer bits of information from the environment. Unlike other approaches to resource-rational analysis, this methodology is accompanied by guarantees about provably-efficient learning through its use of epistemic uncertainty [Der Kiureghian and Ditlevsen, 2009] to resolve the underlying exploration-exploitation trade-off.

2 Rate-Distortion Theory

Here we offer a brief, high-level overview of rate-distortion theory [Shannon, 1959, Berger, 1971, Cover and Thomas, 2012]. Due to space constraints, precise mathematical details are relegated to Appendix A. A lossy compression problem consumes as input a fixed information source \( \mathbb{P}(X \in \cdot) \) and a distortion function \( d : \mathcal{X} \times Z \to \mathbb{R}_{\geq 0} \) which quantifies the loss of fidelity by using an element \( z \in Z \) in lieu of \( x \in \mathcal{X} \). Then, for any \( D \in \mathbb{R}_{\geq 0} \), the rate-distortion function quantifies the fundamental limit of lossy compression as

\[
\mathcal{R}(D) = \inf_{Z} \mathbb{E} \left[ D_{KL}(\mathbb{P}(X \in \cdot) \parallel \mathbb{P}(X \in \cdot)) \right] \text{ such that } \mathbb{E}[d(X,Z)] \leq D,
\]

where the infimum is taken over all random variables \( Z \) (representing the output of a channel given \( X \) as input) that incur bounded expected distortion, \( \mathbb{E}[d(X,Z)] \leq D \). As an aside, we note that one may equivalently define \( \mathcal{R}(D) \) as an infimum over a constrained collection of either joint distributions (as done in [Csiszár, 1974, Dembo and Kontoyiannis, 2002], for example) or conditional distributions (such as in [Blahut, 1972, Kawabata and Dembo, 1994]), representing the channel or lossy compression itself. Naturally, \( \mathcal{R}(D) \) represents the minimum number of bits of information that must be retained on average from \( X \) in order to achieve this bound on the expected loss of fidelity. Moreover, \( \mathcal{R}(D) \) is well-defined for arbitrary information source and channel output random variables taking values on abstract spaces [Csiszá, 1974]. In certain contexts, it can be more suitable to employ the inverse of \( \mathcal{R}(D) \), the distortion-rate function: \( D(R) = \inf_{Z} \mathbb{E}[d(X,Z)] \) such that \( \mathbb{I}(X;Z) \leq R \).

For a given upper limit \( R \in \mathbb{R}_{\geq 0} \) on the bits of information that can be transmitted, \( D(R) \) quantifies the minimum achievable distortion of the resulting compression.

3 Problem Formulation

We formulate a sequential decision-making problem as an episodic, finite-horizon Markov Decision Process (MDP) [Bellman, 1957, Puterman, 1994] defined by \( M = (\mathcal{S}, \mathcal{A}, \mathcal{U}, \mathcal{T}, \beta, H) \). Here \( \mathcal{S} \) denotes a set of states, \( \mathcal{A} \) is a set of actions, \( \mathcal{U} : \mathcal{S} \times \mathcal{A} \to [0,1] \) is a deterministic reward or utility function providing evaluative feedback signals, \( \mathcal{T} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S}) \) 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 length or horizon. Within each one of \( K \in \mathbb{N} \) episodes, the agent acts for exactly \( H \) steps beginning with an initial state \( s_1 \sim \beta \). For each timestep \( h \in [H] \), 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 \( r_h = U(s_h, a_h) \in [0,1] \), 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 behavior as a mapping from states to distributions over 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 \( M \) 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 \( Q^\pi_{M,h}(s,a) = \mathbb{E} \left[ \sum_{h'=h}^H U(s_{h'}, a_{h'}) \mid s_h = s, a_h = a \right] \), where the expectation integrates over randomness in the action selections and transition dynamics. Taking the corresponding value function as \( V^\pi_{M,h}(s) = \mathbb{E}_{a \sim \pi_h(\cdot \mid s)} \left[ Q^\pi_{M,h}(s,a) \right] \), we define the optimal policy \( \pi^* = (\pi_1^*, \pi_2^*, \ldots, \pi_H^*) \) as achieving supremal value \( V^*_{M,h}(s) = \sup_{\pi \in \{\mathcal{S} \to \Delta(\mathcal{A})\}^H} V^\pi_{M,h}(s) \) for all \( s \in \mathcal{S}, h \in [H] \). We let \( \tau_k = (s_1^{(k)}, a_1^{(k)}, r_1^{(k)}, \ldots, s_H^{(k)}, a_H^{(k)}, r_H^{(k)}, s_{H+1}^{(k)}) \) be the random variable denoting the trajectory experienced by the agent in the \( k \)th 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 \( k \)th episode.

4 Capacity Limitation as a Policy Information Bottleneck

There is a long, rich literature exploring the natural limitations on time, knowledge, and cognitive capacity faced by human (and animal) decision makers [Simon, 1956, Newell et al., 1958, Newell and Simon, 1972, Simon, 1982, Gigerenzer and Goldstein, 1996, Vul et al., 2014, Griffiths et al.,]...
2015, Gershman et al., 2015, Icard and Goodman, 2015, Lieder and Griffiths, 2020, Bhui et al., 2021, Brown et al., 2022, Ho et al., 2022]. Crucially, our focus is on a recurring theme throughout this literature of modeling these limitations on cognitive capabilities as being information-theoretic in nature [Sims, 2003, Peng, 2005, Parush et al., 2011, Botvinick et al., 2015, Sims, 2016, 2018, Zenon et al., 2019, Ho et al., 2020, Gershman and Lai, 2020, Gershman, 2020, Mikhael et al., 2021, Lai and Gershman, 2021, Gershman, 2021, Jakob and Gershman, 2022, Bari and Gershman, 2022]. Broadly speaking, these approaches all center around the perspective that a policy $\pi_h: S \rightarrow \Delta(A)$ should be modeled as a communication channel that, like a human decision-maker with limited information processing capability, is subject to a constraint on the maximal number of bits that may be sent across it. Consequently, an agent aspiring to maximize returns must do so subject to this constraint on policy complexity; conversely, an agent ought to transmit the minimum amount of information possible while it endeavors to reach a desired level of performance [Polani, 2009, 2011, Tishby and Polani, 2011, Rubin et al., 2012]. Parallelizing the distortion-rate function $D(R)$, the resulting policy-optimization objective follows as \[
\sup_{\pi \in (S \rightarrow \Delta(A))^H} \mathbb{E}[Q^\pi(S, A)] \quad \text{such that} \quad \mathbb{H}(S; A) \leq R.
\]
Depending on the precise work, subtle variations on this optimization problem exist from choosing a fixed state distribution for the random variable $S$ [Polani, 2009, 2011], incorporating the state visitation distribution of the policy being optimized [Still and Precup, 2012, Gershman, 2020, Lai and Gershman, 2021], or assuming access to the generative model of the MDP and decomposing the objective across a finite state space [Tishby and Polani, 2011, Rubin et al., 2012]. In all of these cases, the end empirical result tends to converge by using variants of the classic Blahut-Arimoto algorithm [Blahut, 1972, Arimoto, 2004] and produce policies that exhibit higher entropy across states under an excessively limited rate $R$, with a gradual convergence towards the greedy optimal policy as $R$ increases.

The alignment between this optimization problem and that of the distortion-rate function is slightly wrinkled by the non-stationarity of the distortion function (here, $Q^\pi$ is used as an analogue to distortion which changes as the policy or channel does) and, when using the policy visitation distribution for $S$, the non-stationarity of the information source. Despite these slight, subtle mismatches with the core rate-distortion problem, the natural synergy between cognitive and computational decision making [Tenenbaum et al., 2011, Lake et al., 2017] has led to various reinforcement-learning approaches that draw direct inspiration from this line of thinking [Klyubin et al., 2005, Ortega and Braun, 2011, Still and Precup, 2012, Ortega and Braun, 2013, Shafieepoorfard et al., 2016, Tionkink and Tishby, 2017, Lerch and Sims, 2018, 2019, Abel et al., 2019], most notably including parallel connections to work on “control as inference” or KL-regularized reinforcement learning [Todorov, 2007, Toussaint, 2009, Kappen et al., 2012, Levine, 2018, Ziebart, 2010, Fox et al., 2016, Haarnoja et al., 2017, Galashov et al., 2019, Tirumala et al., 2019]. Nevertheless, despite their empirical successes, such approaches lack principled mechanisms for addressing the exploration challenge [O’Donoghue et al., 2020]. While exploration is quintessentially studied in the multi-armed bandit setting [Lai and Robbins, 1985, Lattimore and Szepesvári, 2020], we focus our main discussion on reinforcement learning and defer consideration of bandit learning to Appendix B.

Similar to human decision making [Gershman, 2018, Schulz and Gershman, 2019, Gershman, 2019], provably-efficient reinforcement-learning algorithms have historically relied upon one of two possible exploration strategies: optimism in the face of uncertainty [Kearns and Singh, 2002, Brafman and Tennenholtz, 2002, Kakade, 2003, Auer et al., 2009, Bartlett and T ewari, 2009, Strehl et al., 2009, Jaksch et al., 2010, Dann and Brunskill, 2015, Azar et al., 2017, Dann et al., 2017, Jin et al., 2018, Zanette and Brunskill, 2019, Dong et al., 2022] or posterior sampling [Osband et al., 2013, Osband and Van Roy, 2017, Agrawal and Jia, 2017, Lu and Van Roy, 2019, Lu et al., 2021]. While both paradigms have laid down solid theoretical foundations, a line of work has demonstrated how posterior-sampling methods can be more favorable both in theory and in practice [Osband et al., 2013, 2016a,b, Osband and Van Roy, 2017, Osband et al., 2019, Dwaracherla et al., 2020]. In the next section, we outline how these latter posterior-sampling algorithms still lack a consideration for agents acting under limited capacity constraints and demonstrate an alternative utilization of rate-distortion theory to help account for such limitations.

5 Learning Targets for Capacity-Limited Decision Making

As is standard in Bayesian reinforcement learning [Bellman and Kalaba, 1959, Duff, 2002, Ghavamzadeh et al., 2015], neither the transition function nor the reward function are known to
the agent and, consequently, both are treated as random variables. An agent’s initial uncertainty in
the (unknown) true MDP $\mathcal{M}^* = (\mathcal{U}^*, T^*)$ is reflected by a prior distribution $\mathcal{P}(\mathcal{M}^* \in \cdot \mid H_1)$. Since
the regret is a random variable due to our uncertainty in $\mathcal{M}^*$, we integrate over this
randomness to arrive at the Bayesian regret over $K$ episodes: $\text{BAYESREGRET}(K, \pi^{(1:K)}) =
\mathbb{E}
\left[\text{REGRET}(K, \pi^{(1)}, \ldots, \pi^{(K)}, \mathcal{M}^*)\right] = \mathbb{E} \left[ \sum_{k=1}^{K} \left( V_{\mathcal{M}^*, 1}^\ast(s_1) - V_{\mathcal{M}^*, 1}^{(k)}(s_1) \right) \right].$

In the following, 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 $\mathbb{H}_k(X) \triangleq \mathbb{H}(X \mid H_k)$ and
$\mathbb{H}_k(X \mid Y) \triangleq \mathbb{H}(X \mid Y, H_k = H_k)$, for two arbitrary random variables $X$ and $Y$. This notation will
also apply analogously to mutual information: $\mathbb{I}_k(X; Y) \triangleq \mathbb{I}(X; Y \mid H_k = H_k) = \mathbb{H}_k(X) - \mathbb{H}_k(X \mid Y) = \mathbb{H}_k(Y) - \mathbb{H}_k(Y \mid X)$.

The dependence on the realization of a random history $H_k$ makes $\mathbb{I}_k(X; Y)$ a random variable and the usual conditional mutual information arises by integrating over this randomness: $\mathbb{E}\left[\mathbb{I}_k(X; Y)\right] = \mathbb{I}(X; Y \mid H_k)$. Additionally, we will also adopt a similar notation to express a conditional expectation given the random history $H_k$: $\mathbb{E}_k[X] \triangleq \mathbb{E}[X | H_k]$.

A natural starting point for addressing the exploration challenge in a principled manner is via
Thompson sampling [Thompson, 1933, Russo et al., 2018]. The Posterior Sampling for Reinforcement
Learning (PSRL) [Strens, 2000, Osband et al., 2013, Osband and Van Roy, 2014, Abbasi-Yadkori
and 2021, 2020, Abachi et al., 2020, Cui et al., 2020, Ayoub et al., 2020, Schrittwieser et al.,
2020, Nair et al., 2017, Farahmand et al., 2017, Oh et al., 2017, Asadi et al., 2018, Farahmand,
2018, D’Oro et al., 2019, Tomasello, 1999]. By definition, the
rate-distortion function is a non-negative; convex; and non-increasing function in its argument [Cover
et al., 1993] which discards irrelevant environment features and only
exchange for some approximate surrogate, which

Here, the rate-distortion function is indexed by each episode
$k \in [K]$ as the agent takes $\mathbb{P}(\mathcal{M}^* \in \cdot \mid H_k)$ for the information source to be compressed, allowing for incremental refinement of the learning
target as knowledge of the environment accumulates and is transmitted to the “next” agent [Tomasetti
et al., 1993, Tomasetti, 1999]. By definition, the $\hat{\mathcal{M}}$ that achieves this rate-distortion limit will demand
that the agent acquire fewer bits of information than what is needed to identify $\mathcal{M}^*$. Since, the
rate-distortion function is a non-negative; convex; and non-increasing function in its argument [Cover
and Thomas, 2012], the preceding claim is guaranteed for all $k \in [K]$ and any $D > 0$: $\mathcal{R}_k(D) \leq
\mathbb{I}_k(\mathcal{M}^*; \hat{\mathcal{M}}) \leq \mathbb{I}_k(\mathcal{M}^*; \hat{\mathcal{M}})$.

Arumugam and Van Roy [2022a,b] study two distinct choices of distortion function to assess the loss of
fidelity incurred by learning a compressed MDP over the original; the first of these provides an
information-theoretic account of recent successes in deep model-based reinforcement learning [Silver
et al., 2017, Farahmand et al., 2017, Oh et al., 2017, Asadi et al., 2018, Farahmand, 2018, D’Oro et al.,
2020, Abachi et al., 2020, Cui et al., 2020, Ayoub et al., 2020, Schrittwieser et al., 2020, Nair et al.,
2020, Nikishin et al., 2022, Voelcker et al., 2022] through the value-equivalence principle [Grimm
et al., 2020, 2021], where agents deliberately forego learning the true model of the environment in
exchange for some approximate surrogate, which discards irrelevant environment features and only
models dynamics of the world critical to agent performance. In the interest of space, we focus on the
second distortion function which has a simpler form:

$$
\mathcal{d}_Q(\mathcal{M}, \hat{\mathcal{M}}) = \sup_{h \in [H]} ||Q_{\mathcal{M}, h} - Q_{\hat{\mathcal{M}}, h}^{\ast}||_\infty^2 = \sup_{h \in [H]} \sup_{(s, a) \in S \times A} |Q_{\mathcal{M}, h}(s, a) - Q_{\hat{\mathcal{M}}, h}^{\ast}(s, a)|^2.
$$
Crucially, Arumugam and Van Roy [2022b] establish an information-theoretic Bayesian regret bound for a posterior-sampling algorithm that performs probability matching with respect to $\mathcal{M}$ instead of $\mathcal{M}^*$: $\text{BAYESREGRET}(K, \pi^{(1:K)}) \leq \sqrt{K R^Q_1(D)} + 2K(H + 1)\sqrt{D}$. Such an algorithm, by virtue of probability matching, explicitly links an agent’s exploration strategy not only to its epistemic uncertainty but also to that $\mathcal{M}$ which it aspires to learn [Cook et al., 2011]. The bound communicates that an agent with limited capacity must tolerate a higher distortion threshold $D$ and pursue the resulting compressed MDP that bears less fidelity to the original MDP; in exchange, the resulting number of bits needed from the environment to identify such a simplified model of the world is given as $R^Q_1(D)$ and guaranteed to be less than the entropy of $\mathcal{M}^*$. Additionally, one can express a near-identical result through the associated distortion-rate function that explicitly takes an agent’s capacity of only being able to acquire $R \in \mathbb{R}_{\geq 0}$ bits into account: $\text{BAYESREGRET}(K, \pi^{(1:K)}) \leq \sqrt{K R + 2K(H + 1)\sqrt{D^Q_1(R)}}$.

6 Conclusion

A number of recent proposals have approached resource-rationality by combining tools from rate-distortion theory with those of sequential decision-making. Here, we have reviewed a parallel line of work that uses related ideas but within the framework of satisficing and Bayesian reinforcement learning. The distinctive feature of this approach is that it precisely characterizes how capacity-limited learners optimally balance model complexity and value distortion in a manner that gives rise to learning targets which reflect rational satisficing—that is, intelligently choosing how sub-optimal to be based on information acquired during learning. A key challenge for future work will be to translate insights about provably-efficient learning algorithms from this literature into plausible models of human decision-making. More broadly, by modulating the exploration of a satisficing agent to identify a fundamentally different learning target than that of an unconstrained agent, these analyses can provide a framework for understanding how information-theoretic considerations not only shape the products of learning (such as perception, memory, and decisions), but also the process of learning itself.

References


\footnote{Here $\overline{\Gamma} < \infty$ is an (assumed) uniform upper bound to the information ratio [Russo and Van Roy, 2016, 2014, 2018a] that emerges as an artifact of the analysis.}


Daniel Polyanskiy, Yihong Wu. Lecture notes on information theory. 2019.


A Information Theory

Here we introduce various concepts in probability theory and information theory used throughout this paper. We encourage readers to consult [Cover and Thomas, 2012, Gray, 2011, Polyanisky and Wu, 2019, Duchi, 2021] for more background. For readers unfamiliar with measure-theoretic probability, we simply note that while one should feel free to use more familiar definitions of the subsequent information-theoretic quantities [Cover and Thomas, 2012], the more general formulation of this section allows us to be agnostic as to whether the associated random variables are discrete, continuous, or even some mixture thereof. Consequently, the results described in this work hold regardless of whether the state-action space is discrete or continuous; naturally, however, such details do have consequences for practical instantiations of these ideas.

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 \(\{\mathcal{X} \to \mathcal{Y}\} \triangleq \{f : f : \mathcal{X} \to \mathcal{Y}\}\).

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

\[
I(X; Y) = D_{KL}(\mathbb{P}((X, Y) \in \cdot) \| \mathbb{P}(X \in \cdot) \times \mathbb{P}(Y \in \cdot)) \quad D_{KL}(P \| Q) = \left\{ \begin{array}{ll} 
\int dP \frac{dP}{dQ} & P \ll Q \\
+\infty & P \not\ll Q
\end{array} \right.
\]

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 \(X, Y, Z\):

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

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

\[
\mathbb{H}(X) = I(X; X) \quad \mathbb{H}(Y \mid X) = \mathbb{H}(Y) - I(X; Y).
\]

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

\[
I(X; Y) = \mathbb{H}(X) - \mathbb{H}(X \mid Y) = \mathbb{H}(Y) - \mathbb{H}(Y \mid X), \\
I(X; Y \mid Z) = \mathbb{H}(X \mid Z) - \mathbb{H}(X \mid Y, Z) = \mathbb{H}(Y \mid Z) - \mathbb{H}(Y \mid X, Z).
\]

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

\[
I(X; Y) = \mathbb{E} [D_{KL}(\mathbb{P}(Y \in \cdot \mid X) \| \mathbb{P}(Y \in \cdot))]).
\]

B Multi-Armed Bandits

A special case of the finite-horizon episodic MDP formulated earlier is the multi-armed bandit (MAB) problem [Lai and Robbins, 1985, Bubeck et al., 2012, Lattimore and Szepesvári, 2020] which consists of exactly one state \(|S| = 1\) and a horizon \(H = 1\); for simplicity, we will further assume a finite number of arms \(|\mathcal{A}| < \infty\) and follow the standard convention of denoting the total number of time periods as \(T = K \in \mathbb{N}\). Compactly, the MAB is now characterized by an environment \(\mathcal{E} = (\mathcal{A}, \mathcal{U})\) where \(\mathcal{U} : \mathcal{A} \to [0, 1]\) returns the mean reward associated with the input action. While \(\mathcal{A}\) is considered a known quantity, it is the agent’s uncertainty in the underlying rewards of the environment \(\mathcal{E}\) that drive its uncertainty in the optimal action \(A^* = \max_{a \in \mathcal{A}} \mathcal{U}(a)\). At each time period \(t \in [T]\), the agent’s initial/refined uncertainties in \(\mathcal{E}\) are given by the prior/posterior distribution \(\mathbb{P}(\mathcal{E} \in \cdot \mid H_t)\) and it aims to minimize the Bayesian regret: \(\text{BAYESREGRET}(T, \pi^{(1:T)}) = \mathbb{E} \left[ \sum_{t=1}^{T} (\mathcal{U}(A^*) - \mathcal{U}(A_t)) \right]\).

A standard algorithm for solving MABs is Thompson sampling [Thompson, 1933, Russo et al., 2018] where, at each time period, the agent draws a single posterior sample from its current beliefs over the environment \(\theta \sim \mathbb{P}(\mathcal{E} \in \cdot \mid H_t)\) and then acts optimally with respect to this sample \(\pi \in \Delta(\mathcal{A})\). Such an action selection strategy embodies the hallmark
probability-matching principle of Thompson sampling where, for all \( t \in [T] \) and \( a \in \mathcal{A} \), actions are chosen according to their probability of being optimal: \( \mathbb{P}(A_t = a \mid H_t) = \mathbb{P}(A^* = a \mid H_t) \). Moreover, this algorithm yields provably-efficient exploration via an information-theoretic Bayesian regret bound [Russo and Van Roy, 2016] in terms of the agent’s prior entropy over \( A^* \): BAYESREGRET\((T, \pi^{(1:T)}) \leq \sqrt{\frac{2}{\epsilon^2} A \mathbb{H}_1(A^*) T} \). Since the optimal action \( A^* \) is a deterministic function of the environment \( \mathcal{E} \), \( \mathbb{H}_1(A^* \mid \mathcal{E}) = 0 \) and so the total amount of information an agent must acquire from the environment \( \mathcal{E} \) over the course of learning to identify \( A^* \) is given by \( \mathbb{I}_1(\mathcal{E}; A^*) = \mathbb{H}_1(A^*) - \mathbb{H}_1(A^* \mid \mathcal{E}) = \mathbb{H}_1(A^*) \). Consequently, the regret bound affirms that an agent who begins with a strong, well-specified prior over \( A^* \) at the start of learning incurs far less regret than an agent operating under a less informative prior.

Critically, however, decision-making agents operating under constraints on time and resources may not be capable of acquiring all \( \mathbb{H}_1(\tilde{A}^*) \) bits of information needed for optimal decisions. This reality coupled with the observation that \( A^* = f(\mathcal{E}) \) is merely a (deterministic) statistic \( f(\cdot) \) of the unknown environment \( \mathcal{E} \) motivates the consideration of a learning target \( \chi \in \mathcal{A} \) [Lu et al., 2021] that, at the most abstract level, is merely some other statistic of the environment. Of course, \( A^* \) is a statistic of \( \mathcal{E} \) that has a certain desirable property: no other action can achieve mean reward higher than \( A^* \). What characteristic(s) does one want encapsulated in a learning target \( \chi \)? Intuitively, a good learning target should likely strike some kind of balance between two desiderata: (1) be easier to learn than \( A_t \), by having \( \mathbb{I}_1(\mathcal{E}; \chi) \leq \mathbb{I}_1(\mathcal{E}; A^*) = \mathbb{H}_1(A^*) \) and (2) have bounded performance shortfall or regret relative to \( A^* \), \( \mathbb{E}[\mathcal{U}(A_t) - \mathcal{U}(\chi)] \).

To this end, Russo et al. [2017], Russo and Van Roy [2018b, 2022] consider a learning target, parameterized by \( \epsilon > 0 \), as \( A_\epsilon = \text{Uniform}\{\{a \in \mathcal{A} \mid \mathcal{U}(A^*) - \mathcal{U}(a) \leq \epsilon\}\} \) and introduce satisficing Thompson sampling (STS) where probability matching occurs with respect to \( A_\epsilon \): \( \mathbb{P}(A_t = a \mid H_t) = \mathbb{P}(A_t = a \mid H_t), \forall t \in [T], a \in \mathcal{A} \). While it is intuitive that using an \( \epsilon \)-optimal action \( A_\epsilon \) in this manner strikes some kind of trade-off between the two desiderata above, it’s not clear that this is by any means the best trade-off. Furthermore, while \( A^* \) and \( A_\epsilon \) are chosen and held fixed for the duration of learning, one could imagine adapting a learning target \( \chi_t \) incrementally as knowledge of the underlying environment accumulates through \( \mathbb{P}(\mathcal{E} \in \cdot \mid H_t) \).

Arumugam and Van Roy [2021a] leverage the tools of rate-distortion theory to define such a learning target \( \tilde{A}_t \) that an agent should strive for in each time period \( t \in [T] \) according to the rate-distortion function:

\[
\mathcal{R}_t(D) = \inf_{\tilde{A} \in \mathcal{A}} \mathbb{I}_1(\mathcal{E}; \tilde{A}) \text{ such that } \mathbb{E}_t\left[d(\tilde{A}, \mathcal{E})\right] \leq D.
\]

Assuming that \( \mathcal{E} \) takes values in a set \( \Theta \), the distortion function \( d_t : \mathcal{A} \times \Theta \to \mathbb{R}_{\geq 0} \) is defined as the expected squared regret of the candidate action:

\[
d_t(a, \theta) = \mathbb{E}_t\left[\left(\mathcal{U}(A^*) - \mathcal{U}(a)\right)^2 \mid \mathcal{E} = \theta\right], \quad \forall a \in \mathcal{A}, \theta \in \Theta.
\]

Clearly, achieving bounded distortion for any \( D \in \mathbb{R}_{\geq 0} \) addresses the second criterion for learning targets. Since, for all \( t \in [T] \) and any \( D > 0 \), the rate-distortion function \( \mathcal{R}_t(D) \) is a non-negative; convex; and non-increasing function in its argument [Cover and Thomas, 2012], it follows that we also ensure the first criterion of a learning target for any \( D > 0 \): \( \mathcal{R}_t(D) \leq \mathcal{R}_t(0) \leq \mathbb{I}_t(\mathcal{E}; A^*) = \mathbb{H}_t(A^*) \).

Arumugam and Van Roy [2021a,b] employ the classic Blahut-Arimoto algorithm [Blahut, 1972, Arimoto, 1972] for computing these target actions \( \tilde{A}_t \) in each time period and demonstrate a spectrum of regret curves corresponding to various satisficing and optimal policies. Conveniently and unlike with \( A^* \) or \( A_\epsilon \), the onus does not fall upon the agent designer to specify the exact form of the learning target in each time period and, instead, they need only select an appropriate distortion function and threshold for the Blahut-Arimoto algorithm to produce a \( \tilde{A}_t \) that strikes the optimal trade-off between ease of learnability and near optimality. Arumugam and Van Roy [2021a] prove the following information-theoretic Bayesian regret bound for their Blahut-Arimoto Satisficing Thompson Sampling (BLASTS) algorithm:

\[
\text{BAYESREGRET}(T, \pi^{(1:T)}) \leq \sqrt{2A|\mathcal{R}_1(D)|T + 2T \sqrt{D}},
\]

alongside computational experiments demonstrating a wide spectrum of possible target actions that can be realized. Naturally, a setting of \( D = 0 \) recovers the information-theoretic regret bound of Russo and Van Roy [2016] up to a constant factor. With rewards assumed to be in the unit interval
[0, 1], a setting of $D = 1$ permits any action to achieve the rate-distortion limit, requiring 0 bits of information overall and rendering the regret bound vacuous. For intermediate settings of $D$, however, the bound critically provides the information-theoretically optimal balance in selecting a learning target that requires the minimum amount of information to deliver a minimum tolerable amount of sub-optimality.