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

Anonymous Author(s)
Affiliation
Address
email

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 [110, 111, 78, 79, 112], it has been usually treated as contrary to rational analysis [5]. In particular, modern resource-rational analyses, both utility-theoretic [50, 73] and information-theoretic [115, 113, 43], 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 [7, 10]; 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 [32] to resolve the underlying exploration-exploitation trade-off.

2 Rate-Distortion Theory

Here we offer a brief, high-level overview of rate-distortion theory [108, 18, 26]. 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 $d : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$ which quantifies the loss of fidelity by using an element $z \in \mathcal{Z}$ 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

$$\mathcal{R}(D) = \inf_{Z \in \mathcal{Z}} \mathbb{I}(X; Z) \triangleq \inf_{Z \in \mathcal{Z}} \mathbb{E}\left[D_{\mathrm{KL}}(\mathbb{P}\left(X \in \cdot \mid Z\right) \,||\, \mathbb{P}(X \in \cdot))\right] \text{ such that } \mathbb{E}\left[d(X, Z)\right] \leq D,$$

Submitted to Information-Theoretic Principles in Cognitive Systems Workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

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, $\mathbb{E}\left[d(X,Z)\right] \leq D$. 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 [27, 31], for example) or conditional distributions (such as in [20, 63]), 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 [27]. In certain contexts, it can be more suitable to employ the inverse of $\mathcal{R}(D)$, the distortion-rate function: $\mathcal{D}(R) = \inf\limits_{Z \in \mathcal{Z}} \mathbb{E}\left[d(X,Z)\right]$ 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, $\mathcal{D}(R)$ 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) [17, 96] defined by $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{U}, \mathcal{T}, \beta, H \rangle$. 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 $r_h = \mathcal{U}(s_h, a_h) \in [0,1]$, 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 $\mathcal{M}$ at timestep $h \in [H]$ when starting at state $s \in \mathcal{S}$ and taking action $a \in \mathcal{A}$ is assessed by its associated action-value function $Q^{\pi}_{\mathcal{M},h}(s,a) = \mathbb{E}\left[\sum\limits_{h'=h}^{H} \mathcal{U}(s_{h'}, a_{h'}) \mid s_h = s, a_h = a\right]$, where the expectation integrates over randomness in the action selections and transition dynamics. Taking the corresponding value function as $V^{\pi}_{\mathcal{M},h}(s) = \mathbb{E}_{a \sim \pi_h(\cdot \mid s)}\left[Q^{\pi}_{\mathcal{M},h}(s,a)\right]$, we define the optimal policy $\pi^{\star} = (\pi^{\star}_1, \pi^{\star}_2, \ldots, \pi^{\star}_H)$ as achieving supremal value $V^{\star}_{\mathcal{M},h}(s) = \sup\limits_{\pi \in \{\mathcal{S} \to \Delta(\mathcal{A})\}^H} V^{\pi}_{\mathcal{M},h}(s)$ for all $s \in \mathcal{S}$, $h \in [H]$. We let $\tau_k = (s^{(k)}_1, a^{(k)}_1, r^{(k)}_1, \ldots, s^{(k)}_H, a^{(k)}_H, r^{(k)}_H, s^{(k)}_{H+1})$ be the random variable denoting the trajectory experienced by the agent in the kth episode. Meanwhile, $H_k = \{\tau_1, \tau_2, \ldots, \tau_{k-1}\} \in \mathcal{H}_k$ is the random variable representing the entire history of the agent's interaction within the environment at the start of the kth episode.

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 [111, 78, 79, 112, 48, 129, 50, 46, 57, 73, 19, 23, 56]. Crucially, our focus is on a recurring theme throughout this literature of modeling these limitations on cognitive capabilities as being information-theoretic in nature [115, 92, 91, 21, 113, 114, 131, 55, 45, 43, 76, 66, 44, 58, 15]. Broadly speaking, these approaches all center around the perspective that a policy $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$ 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 [93, 94, 123, 97]. Paralleling the distortion-rate function $\mathcal{D}(R)$, the resulting policy-optimization objective follows as $\sup\limits_{\pi \in \{\mathcal{S} \to \Delta(\mathcal{A})\}^H} \mathbb{E}\left[Q^{\pi}(S,A)\right]$ such that $\mathbb{I}(S;A) \leq R$. Depending on the precise work, subtle variations on this optimization problem exist from choosing a fixed state distribution for the random variable S [93, 94], incorporating the state visitation distribution of the

policy being optimized [116, 43, 66], or assuming access to the generative model of the MDP and decomposing the objective across a finite state space [123, 97]. In all of these cases, the end empirical result tends to converge by using variants of the classic Blahut-Arimoto algorithm [20, 6] to solve the Lagrangian associated with the constrained optimization 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^π 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 [119, 68] has led to various reinforcement-learning approaches that draw direct inspiration from this line of thinking [65, 84, 116, 83, 107, 121, 70, 71, 3], most notably including parallel connections to work on "control as inference" or KL-regularized reinforcement learning [124, 127, 62, 72, 132, 39, 53, 54, 40, 122]. Nevertheless, despite their empirical successes, such approaches lack principled mechanisms for addressing the exploration challenge [81]. While exploration is quintessentially studied in the multi-armed bandit setting [67, 69], we focus our main discussion on reinforcement learning and defer consideration of bandit learning to Appendix B.

Similar to human decision making [41, 106, 42], provably-efficient reinforcement-learning algorithms have historically relied upon one of two possible exploration strategies: optimism in the face of uncertainty [64, 22, 61, 12, 16, 117, 59, 29, 14, 30, 60, 130, 33] or posterior sampling [87, 86, 4, 74, 75]. 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 [87–89, 86, 90, 36]. 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 [47], 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 $\mathcal{M}^\star = (\mathcal{U}^\star, \mathcal{T}^\star)$ is reflected by a prior distribution $\mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_1)$. Since the regret is a random variable due to our uncertainty in $\mathcal{M}^\star$, we integrate over this randomness to arrive at the Bayesian regret over K episodes: $\textsc{BayesRegret}(K, \pi^{(1:K)}) = \mathbb{E}\left[\textsc{Regret}(K, \pi^{(1)}, \ldots, \pi^{(K)}, \mathcal{M}^\star)\right] = \mathbb{E}\left[\sum_{k=1}^{K}\left(V^\star_{\mathcal{M}^\star,1}(s_1) - V^{\pi^{(k)}}_{\mathcal{M}^\star,1}(s_1)\right)\right]$.

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 $\mathbb{H}_k(X) \triangleq \mathbb{H}(X \mid H_k = H_k)$ 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 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: $\mathbb{E}_k\left[X\right] \triangleq \mathbb{E}\left[X|H_k\right]$.

A natural starting point for addressing the exploration challenge in a principled manner is via Thompson sampling [120, 104]. The Posterior Sampling for Reinforcement Learning (PSRL) [118, 87, 85, 2, 4, 86, 74] algorithm does this by, in each episode $k \in [K]$, sampling a candidate MDP $\mathcal{M}_k \sim \mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_k)$ and executing its optimal policy in the environment $\pi^{(k)} = \pi^\star_{\mathcal{M}_k}$; notably, such posterior sampling guarantees the hallmark probability-matching principle of Thompson sampling: $\mathbb{P}(\mathcal{M}_k = M \mid H_k) = \mathbb{P}(\mathcal{M}^\star = M \mid H_k)$, $\forall M \in \mathfrak{M}, k \in [K]$. The resulting trajectory τ_k leads to a new history $H_{k+1} = H_k \cup \tau_k$ and an updated posterior over the true MDP $\mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_{k+1})$.

We recognize that, for complex environments, pursuit of the exact MDP $\mathcal{M}^\star$ may be an entirely infeasible goal. A MDP representing control of a real-world, physical system, for example, suggests that learning the associated transition function requires the agent to internalize laws of physics and

motion with near-perfect accuracy. More formally, identifying $\mathcal{M}^\star$ demands the agent obtain exactly $\mathbb{H}_1(\mathcal{M}^\star)$ bits of information from the environment which, under an uninformative prior, may either be prohibitively large by far exceeding the agent's capacity constraints or be simply impractical under time and resource constraints [75]. Consequently, an agent must embrace a satisficing solution and Arumugam and Van Roy [9, 10] employ the following rate-distortion function to identify a suitable lossy compression of the underlying MDP $\widetilde{\mathcal{M}} \in \mathfrak{M}$ whose information an agent may prioritize as an alternative learning target to $\mathcal{M}^\star$: $\mathcal{R}_k(D) = \inf_{\widetilde{\mathcal{M}} \in \mathfrak{M}} \mathbb{I}_k(\mathcal{M}^\star; \widetilde{\mathcal{M}})$ such that $\mathbb{E}_k[d(\mathcal{M}^\star, \widetilde{\mathcal{M}})] \leq D$.

Here, the rate-distortion function is indexed by each episode $k \in [K]$ as the agent takes $\mathbb{P}(\mathcal{M}^\star \in \cdot \mid H_k)$ 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 [126, 125]. By definition, the $\widetilde{\mathcal{M}}$ that achieves this rate-distortion limit will demand that the agent acquire fewer bits of information than what is needed to identify $\mathcal{M}^\star$. Since, the rate-distortion function is a non-negative; convex; and non-increasing function in its argument [26], the preceding claim is guaranteed for all $k \in [K]$ and any $D > 0$: $\mathcal{R}_k(D) \leq \mathcal{R}_k(0) \leq \mathbb{I}_k(\mathcal{M}^\star; \mathcal{M}^\star) = \mathbb{H}_k(\mathcal{M}^\star)$.

Arumugam and Van Roy [10] 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 [109, 38, 82, 11, 37, 34, 1, 28, 13, 105, 77, 80, 128] through the value-equivalence principle [51, 52], 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:

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

Crucially, Arumugam and Van Roy [10] establish an information-theoretic Bayesian regret bound for a posterior-sampling algorithm that performs probability matching with respect to $\widetilde{\mathcal{M}}$ instead of $\mathcal{M}^\star$: $\textsc{BayesRegret}(K, \pi^{(1:K)}) \leq \sqrt{\overline{\Gamma} K \mathcal{R}_1^{Q^\star}(D)} + 2K(H+1)\sqrt{D}$. 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 $\widetilde{\mathcal{M}}$ which it aspires to learn [25]. 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 $\mathcal{R}_1^{Q^\star}(D)$ and guaranteed to be less than the entropy of $\mathcal{M}^\star$[1]. 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: $\textsc{BayesRegret}(K, \pi^{(1:K)}) \leq \sqrt{\overline{\Gamma} K R} + 2K(H+1)\sqrt{\mathcal{D}_1^{Q^\star}(R)}$.

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 explicitly having a satisficing agent employ a fundamentally different exploration strategy 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 the process of learning itself.

[1] Here $\overline{\Gamma} < \infty$ is an (assumed) uniform upper bound to the information ratio [99, 98, 100] that emerges as an artifact of the analysis.

References

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

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

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

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

[5] John R Anderson. The adaptive character of thought. 1990.

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

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

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

[9] Dilip Arumugam and Benjamin Van Roy. Between rate-distortion theory & value equivalence in model-based reinforcement learning. *Proceedings of Reinforcement Learning and Decision Making (RLDM)*, 2022.

[10] Dilip Arumugam and Benjamin Van Roy. Deciding what to model: Value-equivalent sampling for reinforcement learning. *arXiv preprint arXiv:2206.02072*, 2022.

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

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

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

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

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

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

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

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

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

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] Claire Cook, Noah D Goodman, and Laura E Schulz. Where science starts: Spontaneous experiments in preschoolers' exploratory play. *Cognition*, 120(3):341–349, 2011.

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

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

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

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

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

[31] Amir Dembo and L Kontoyiannis. Source coding, large deviations, and approximate pattern matching. *IEEE Transactions on Information Theory*, 48(6):1590–1615, 2002.

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

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

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

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

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

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

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

[39] 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*, pages 202–211, 2016.

[40] 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.

[41] Samuel J Gershman. Deconstructing the human algorithms for exploration. *Cognition*, 173: 34–42, 2018.

[42] Samuel J Gershman. Uncertainty and exploration. *Decision*, 6(3):277, 2019.

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

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

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

[46] 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.

[47] 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.

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

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

[50] 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.

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

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

[53] 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*, pages 1352–1361. JMLR. org, 2017.

[54] 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*, pages 1861–1870, 2018.

[55] 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*, pages 1300–1307. AAAI press, 2020.

[56] 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.

[57] Thomas Icard and Noah D Goodman. A resource-rational approach to the causal frame problem. In *CogSci*, 2015.

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

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

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

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

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

[63] Tsutomu Kawabata and Amir Dembo. The rate-distortion dimension of sets and measures. *IEEE Transactions on Information Theory*, 40(5):1564–1572, 1994.

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

[65] 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, pages 128–135. IEEE, 2005.

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

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

[68] 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.

[69] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

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

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

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

[73] 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.

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

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

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

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

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

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

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

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

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

[83] 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.

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

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

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

[87] 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.

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

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

[90] 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.

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

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

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

[94] Daniel Polani. An informational perspective on how the embodiment can relieve cognitive burden. In *2011 IEEE Symposium on Artificial Life (ALIFE)*, pages 78–85. IEEE, 2011.

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

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

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

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

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

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

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

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

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

[104] 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.

[105] 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.

[106] Eric Schulz and Samuel J Gershman. The algorithmic architecture of exploration in the human brain. *Current opinion in neurobiology*, 55:7–14, 2019.

[107] 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.

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

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

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

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

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

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

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

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

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

[117] 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.

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

[119] 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.

[120] 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.

[121] Stas Tiomkin and Naftali Tishby. A unified Bellman equation for causal information and value in Markov Decision Processes. *arXiv preprint arXiv:1703.01585*, 2017.

[122] 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.

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

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

[125] Michael Tomasello. *The cultural origins of human cognition*. Harvard University Press, 1999.

[126] Michael Tomasello, Ann Cale Kruger, and Hilary Horn Ratner. Cultural learning. *Behavioral and Brain Sciences*, 16(3):495–511, 1993.

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

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

[129] 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.

[130] Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312. PMLR, 2019.

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

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

A Information Theory

Here we introduce various concepts in probability theory and information theory used throughout this paper. We encourage readers to consult [26, 49, 95, 35] 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 [26], 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 $\{\mathcal{X} \to \mathcal{Y}\} \triangleq \{f \mid f : \mathcal{X} \to \mathcal{Y}\}$.

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

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

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

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

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

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

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

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

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

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

B Multi-Armed Bandits

A special case of the finite-horizon episodic MDP formulated earlier is the multi-armed bandit (MAB) problem [67, 24, 69] which consists of exactly one state $|\mathcal{S}| = 1$ 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 $\mathcal{E} = \langle \mathcal{A}, \mathcal{U} \rangle$ 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 $A^\star = \max_{a \in \mathcal{A}} \mathcal{U}(a)$. 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:

$$\textsc{BayesRegret}(T, \pi^{(1:T)}) = \mathbb{E}\left[\sum_{t=1}^{T} \left(\mathcal{U}(A^\star) - \mathcal{U}(A_t)\right)\right].$$

A standard algorithm for solving MABs is Thompson sampling [120, 104] 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 $\min_{\pi \in \Delta(\mathcal{A})} \mathbb{E}\left[\mathcal{U}(A^\star) - \mathcal{U}(A_t) \mid \mathcal{E} = \theta\right]$. 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: $\mathbb{P}(A_t = a \mid H_t) = \mathbb{P}(A^\star = a \mid H_t)$. Moreover, this algorithm yields provably-efficient exploration via an information-theoretic Bayesian regret bound [99] in terms of the agent's prior entropy over $A^\star$: $\textsc{BayesRegret}(T, \pi^{(1:T)}) \leq \sqrt{\frac{1}{2}|\mathcal{A}|\mathbb{H}_1(A^\star)T}$. Since the optimal action $A^\star$ is a deterministic function of the environment $\mathcal{E}$, $\mathbb{H}_1(A^\star \mid \mathcal{E}) = 0$ and so the total amount of information an agent must acquire from the environment $\mathcal{E}$ over the course of learning to identify $A^\star$ is given by $\mathbb{I}_1(\mathcal{E}; A^\star) = \mathbb{H}_1(A^\star) - \mathbb{H}_1(A^\star \mid \mathcal{E}) = \mathbb{H}_1(A^\star)$. Consequently, the regret bound affirms that an agent who begins with a strong, well-specified prior over $A^\star$ 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 $\mathbb{H}_1(A^\star)$ bits of information needed for optimal decisions. This reality coupled with the observation that $A^\star = f(\mathcal{E})$ is merely a (deterministic) statistic $f(\cdot)$ of the unknown environment $\mathcal{E}$ motivates the consideration of a learning target $\chi \in \mathcal{A}$ [75] that, at the most abstract level, is merely some other statistic of the environment. Of course, $A^\star$ is a statistic of $\mathcal{E}$ that has a certain desirable property: no other action can achieve mean reward higher than $A^\star$. What characteristic(s) does one want encapsulated in a learning target χ? Intuitively, a good learning target should likely strike some kind of balance between two desiderata: **(1)** be easier to learn than $A_\star$ by having $\mathbb{I}_1(\mathcal{E}; \chi) \leq \mathbb{I}_1(\mathcal{E}; A^\star) = \mathbb{H}_1(A^\star)$ and **(2)** have bounded performance shortfall or regret relative to $A^\star$, $\mathbb{E}\left[\mathcal{U}(A_\star) - \mathcal{U}(\chi)\right]$.

To this end, Russo et al. [103], Russo and Van Roy [101, 102] consider a learning target, parameterized by $\varepsilon > 0$, as $A_\varepsilon = \text{Uniform}(\{a \in \mathcal{A} \mid \mathcal{U}(A^\star) - \mathcal{U}(a) \leq \varepsilon\})$ and introduce satisficing Thompson sampling (STS) where probability matching occurs with respect to A_ε: $\mathbb{P}(A_t = a \mid H_t) = \mathbb{P}(A_\varepsilon = a \mid H_t)$, $\forall t \in [T], a \in \mathcal{A}$. While it is intuitive that using an ε-optimal action A_ε 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 $A^\star$ and A_ε are chosen and held fixed for the duration of learning, one could imagine adapting a learning target χ_t incrementally as knowledge of the underlying environment accumulates through $\mathbb{P}(\mathcal{E} \in \cdot \mid H_t)$.

Arumugam and Van Roy [7] 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:

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

Assuming that $\mathcal{E}$ takes values in a set Θ, 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:

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

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 [26], it follows that we also ensure the first criterion of a learning target for any $D > 0$: $\mathcal{R}_t(D) \leq \mathcal{R}_t(0) \leq \mathbb{I}_t(\mathcal{E}; A^\star) = \mathbb{H}_t(A^\star)$. Arumugam and Van Roy [7, 8] employ the classic Blahut-Arimoto algorithm [20, 6] 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 $A^\star$ or A_ε, 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 [7] prove the following information-theoretic Bayesian regret bound for their Blahut-Arimoto Satisficing Thompson Sampling (BLASTS) algorithm:

$$\textsc{BayesRegret}(T, \pi^{(1:T)}) \leq \sqrt{2|\mathcal{A}|\mathcal{R}_1(D)T} + 2T\sqrt{D},$$

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 [99] 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.