Deciding What to Learn: A Rate-Distortion Approach

Dilip Arumugam [1] **Benjamin Van Roy** [1]

Abstract

Agents that learn to select optimal actions represent a prominent focus of the sequential decision-making literature. In the face of a complex environment or constraints on time and resources, however, aiming to synthesize such an optimal policy can become infeasible. These scenarios give rise to an important trade-off between the information an agent must acquire to learn and the sub-optimality of the resulting policy. While an agent designer has a preference for how this trade-off is resolved, existing approaches further require that the designer translate these preferences into a fixed learning target for the agent. In this work, leveraging rate-distortion theory, we automate this process such that the designer need only express their preferences via a single hyperparameter and the agent is endowed with the ability to compute its own learning targets that best achieve the desired trade-off. We establish a general bound on expected discounted regret for an agent that decides what to learn in this manner along with computational experiments that illustrate the expressiveness of designer preferences and even show improvements over Thompson sampling in identifying an optimal policy.

1. Introduction

Learning is a process of acquiring information that reduces an agent's uncertainty about its environment. Anything that an agent may endeavor to learn requires obtaining a precise amount of information about the environment; naturally, as measured by this requisite information, some things are easier to learn than others. When interacting with a complex environment, however, the agent is spoiled for choice as there is too much to learn within any reasonable time frame, and the agent must prioritize. A simple approach is to designate a *learning target*, which can be thought of as a corpus of information that, while insufficient to fully identify the environment, suffices to guide effective decisions. Then, the agent can prioritize gathering of information about this learning target.

One possible learning target, which has dominated the bandit-learning literature (Bubeck et al., 2012; Lattimore & Szepesvári, 2020), is an action-selection policy that would be optimal given full information about the environment. While suitable for simple environments, like multi-armed bandits with few arms, this concept does not scale well with the size of the action space. Moreover, in complex environments, there is typically too much to learn about the optimal policy within any reasonable time frame.

Recent work has highlighted conditions under which it is helpful to target a near-optimal or *satisficing* policy (Russo & Van Roy, 2018b). Such a learning target is not without precedent and has been studied implicitly in a variety of contexts (Bubeck et al., 2011; Kleinberg et al., 2008; Rusmevichientong & Tsitsiklis, 2010; Ryzhov et al., 2012; Deshpande & Montanari, 2012; Berry et al., 1997; Wang et al., 2008; Bonald & Proutiere, 2013). There is an important tension between information requirements for policy learning and policy performance; as one is more permissive of increasingly sub-optimal policies, the requisite amount of information for learning such policies decreases. Crucially, a satisficing policy can be manually specified by an agent designer in order to strike the desired balance. To do so, however, it is incumbent upon the designer to have sufficient knowledge of the problem structure in order to negotiate the information-performance trade-off.

We consider the design of an agent that selects its own learning target. This shifts the agent designer's role from specifying one to endowing the agent with the ability to designate and to suitably adapt the target as learning progresses. The designer can specify the general form of this learning target as part of the scaffold for a learning algorithm. More traditional, fixed-target learning algorithms can then be repurposed as subroutines an agent may use to achieve its own goals. We introduce in this paper what is possibly the first principled approach to address a fundamental question: *how should an agent decide what to learn?*

As a first step, this work offers one concrete answer to this question by introducing an agent that adaptively learns

[1]Stanford University, California, USA. Correspondence to: Dilip Arumugam <dilip@cs.stanford.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

target actions. To endow this agent with the ability to reason about the information-performance trade-off autonomously, we employ rate-distortion theory (Shannon, 1959; Berger, 1971), building on connections to sequential decision-making made by Russo & Van Roy (2018b). With an appropriately chosen distortion measure, the canonical rate-distortion function precisely characterizes the trade-off between the information required for policy learning and policy performance. Rather than placing the burden on the agent designer to procure the solution to a single rate-distortion function on behalf the agent, we instead place the onus upon the agent to solve a rate-distortion function in each time period and gradually adapt its self-designated target action. We recognize that computation of rate-distortion functions is a well-studied problem of the information theory community for which an elegant solution already exists as the classic Blahut-Arimoto algorithm (Blahut, 1972; Arimoto, 1972). Accordingly, we begin by introducing a variant of Thompson sampling which uses the Blahut-Arimoto algorithm as a subroutine for computing a target action in each time period that achieves the rate-distortion limit. We then prove a bound on the expected discounted regret for this algorithm, differing from previous information-theoretic analyses in its treatment of a learning target that changes in each time period. Finally, we conclude with a series of computational experiments that highlight the efficacy of our procedure in enabling an agent to target desired points along the information-performance trade-off curve.

The paper proceeds as follows: in Section 2 we briefly discuss background material before clarifying the connections between our approach and rate-distortion theory in Section 3. Due to space constraints, we relegate an overview of prior work to the appendix. We introduce our main algorithm in Section 4 before finally presenting a corresponding regret analysis and supporting computational experiments in Sections 5 and 6, respectively.

2. Background

In this section, we begin with an overview of several standard quantities in information theory. For more background on information theory, see Cover & Thomas (2012). We conclude the section with a brief outline of rate-distortion theory.

2.1. Information Theory

Consider three random variables X, Y, Z defined on a probability space $(\Omega, \mathbb{F}, \mathbb{P})$. We define entropy, conditional entropy, mutual information, and conditional mutual information as follows:

$$\begin{aligned}
\mathbb{H}(X) &= -\mathbb{E}[\log(\mathbb{P}(X \in \cdot))] \\
\mathbb{H}(Y|X) &= -\mathbb{E}[\log(\mathbb{P}(Y \in \cdot|X))] \\
\mathbb{I}(X;Y) &= \mathbb{H}(X) - \mathbb{H}(X|Y) = \mathbb{H}(Y) - \mathbb{H}(Y|X) \\
\mathbb{I}(X;Y|Z) &= \mathbb{H}(X|Z) - \mathbb{H}(X|Y,Z) = \mathbb{H}(Y|Z) - \mathbb{H}(Y|X,Z)
\end{aligned}$$

Importantly, the multivariate mutual information between a single random variable X and another sequence of random variables $Z_1, \dots Z_n$ decomposes via the chain rule of mutual information:

$$\mathbb{I}(X; Z_1, \dots, Z_n) = \sum_{i=1}^{n} \mathbb{I}(X; Z_i | Z_1, \dots, Z_{i-1})$$

2.2. Rate-Distortion Theory

Rate-distortion theory is a sub-area of information theory concerned with lossy compression and the achievability of coding schemes that maximally compress while adhering to a desired upper bound on error or loss of fidelity (Shannon, 1959; Berger, 1971; Cover & Thomas, 2012). More formally, consider a random variable X with fixed distribution $p(x) = \mathbb{P}(X = x)$ that represents an information source along with a random variable $\hat{X}$ that corresponds to a channel output. Given a distortion measure $d : \mathcal{X} \times \hat{\mathcal{X}} \mapsto \mathbb{R}_{\geq 0}$ and a desired upper bound on distortion D, the rate-distortion function is defined as:

$$\mathcal{R}(D) = \inf_{\hat{X} \in \Lambda} \mathcal{I}(X; \hat{X}) \tag{1}$$

quantifying the minimum number of bits (on average) that must be communicated from X across a channel in order to adhere to the specified expected distortion threshold D. Here, the infimum is taken over $\Lambda = \{\hat{X} : \mathbb{E}\left[d(X, \hat{X})\right] \leq D\}$ representing the set of all random variables $\hat{X} : \Omega \mapsto \hat{\mathcal{X}}$ which which satisfy the constraint on expected distortion. Intuitively, a higher rate corresponds to requiring more bits of information and smaller information loss between X and $\hat{X}$, enabling higher-fidelity reconstruction (lower distortion); conversely, lower rates reflect more substantial information loss, potentially exceeding the tolerance on distortion D.

Fact 1. *$\mathcal{R}(D)$ is a non-negative, convex, and monotonically-decreasing function in D (Cover & Thomas, 2012).*

Some readers may be more familiar with the related problem of computing channel capacity; while the rate-distortion function considers a fixed information source $p(x)$ and optimizes for a channel $p(\hat{x}|x) = \mathbb{P}(\hat{X} = \hat{x}|X = x)$ that minimizes distortion, the channel-capacity function considers a fixed channel and optimizes for the information source that maximizes throughput.

3. Sequential Decision-Making & Rate-Distortion Theory

3.1. Problem Formulation

We define all random variables with respect to a common probability space $(\Omega, \mathbb{F}, \mathbb{P})$; all events are determined by a random outcome $\omega \in \Omega$. An agent interacts with an unknown environment $\mathcal{E}$, which is itself a random variable. The interaction generates a history $H_t = (A_0, O_1, A_1, O_2, \ldots, O_t)$ of actions and observations that take values in finite sets $\mathcal{A}$ and $\mathcal{O}$. Initial uncertainty about the environment is reflected by probabilities $\mathbb{P}(\mathcal{E} \in \cdot)$ where $\mathcal{E}$ has support on Θ and, as the history unfolds, what can be learned is represented by conditional probabilities $\mathbb{P}(\mathcal{E} \in \cdot | H_t)$.

Actions are independent of the environment conditioned on history, $A_{t+1} \perp \mathcal{E} | H_t$. This reflects the fact that the agent selects actions based only on history and, possibly, algorithmic randomness. It may be helpful to think of the actions as being selected by an *admissible policy* $\pi(a|H_t) = \mathbb{P}(A_t = a | H_t)$, which assigns a probability to each action $a \in \mathcal{A}$ given the history. By *admissible*, we mean that action probabilities are determined by history and do not depend on further information about the environment.

We assume that observations are independent of history conditioned on the environment and most recent action, $O_{t+1} \perp H_t | (\mathcal{E}, A_t)$. Note that this precludes delayed consequences, and we will restrict attention in this paper to such environments. Further, we assume a stationary environment such that conditional observation probabilities $\mathbb{P}(O_{t+1} | \mathcal{E}, A_t)$ do not depend on t.

Upon each observation, the agent enjoys a reward $R_{t+1} = r(A_t, O_{t+1})$ where $r : \mathcal{A} \times \mathcal{O} \mapsto \mathbb{R}$ is a deterministic function. Let $\overline{r}(a) = \mathbb{E}[R_{t+1} | A_t = a, \mathcal{E}]$ denote mean reward and note that $\overline{r}$ is itself a random variable since it depends on $\mathcal{E}$. Let $A_\star$ be an action that maximizes the expected mean reward $\mathbb{E}[\overline{r}(A_\star)]$ and let $R_\star = \overline{r}(A_\star)$. Note that $A_\star$ and $R_\star$ are random variables, as they depend on $\mathcal{E}$. It may be helpful to think of $A_\star$ as generated by an optimal policy $\pi_\star(a) = \mathbb{P}(A_t = a | \mathcal{E})$, which is *inadmissible*, in the sense that it depends on the environment, not just the history. Traditionally, the performance of an admissible policy π at any time period $\tau = 0, 1, 2, \ldots$ is quantified by its regret:

$$\mathbb{E}\left[\sum_{t=\tau}^{\infty} R_\star - R_{t+1} \middle| H_\tau\right].$$

While this is a suitable measure of asymptotic performance, we follow suit with Russo & Van Roy (2018b) and examine expected discounted regret

$$\mathbb{E}\left[\sum_{t=\tau}^{\infty} \gamma^{t-\tau}(R_\star - R_{t+1}) \middle| H_\tau\right],$$

where the discount factor $\gamma \in [0, 1)$ helps regulate the agent's preference for minimizing near-term versus long-term performance shortfall.

3.2. Target Actions

In the course of identifying an optimal policy, we take $\mathbb{H}(A_\star)$ to denote the bits of information an agent must acquire in order to identify $A_\star$. Russo & Van Roy (2016) offer a novel information-theoretic analysis of Thompson sampling (Thompson, 1933) whose corresponding regret bound depends on $\mathbb{H}(A_\star)$. Due to the non-negativity of conditional entropy, $\mathbb{H}(A_\star | \mathcal{E}) \geq 0$, it follows that the entropy of $A_\star$ upper bounds the mutual information between $A_\star$ and $\mathcal{E}$, $\mathbb{H}(A_\star) \geq \mathbb{H}(A_\star) - \mathbb{H}(A_\star | \mathcal{E}) = \mathbb{I}(A_\star; \mathcal{E})$, which is tight when the optimal action $A_\star$ is a deterministic function of $\mathcal{E}$.

When faced with a complex environment $\mathcal{E}$, acquiring these $\mathbb{H}(A_\star)$ bits of information for optimal behavior may be exceptionally difficult. While Thompson sampling is a simple yet effective algorithm with widespread empirical success in synthesizing optimal policies (Chapelle & Li, 2011; Russo et al., 2018), it can fall short in these more challenging learning settings. Russo & Van Roy (2018b) first drew awareness to this issue, highlighting several examples where Thompson sampling struggles in the face of a large, possibly infinite, action set or a time-sensitivity constraint on learning. In short, the problem stems from the fact that Thompson sampling will select new, untested actions in each time period, rapidly becoming inefficient as the number of actions grows.

Russo & Van Roy (2018b) introduce the notion of satisficing actions $\tilde{A}$, in lieu of optimal actions, as a remedy to the aforementioned issues. The core premise of this alternative learning target is that a deliberately sub-optimal action should require the agent to learn fewer bits of information about the environment in order to identify a corresponding satisficing policy. Their proposed satisficing Thompson sampling algorithm makes the natural modification of probability matching with respect to the agent's posterior beliefs over $\tilde{A}$, given the current history, such that $A_t \sim \mathbb{P}(\tilde{A} = \cdot | H_t)$. Crucially, Russo & Van Roy (2018b) draw an interesting connection between the specification of satisficing actions and rate-distortion theory. Taking the distortion function to be the instantaneous expected regret conditioned on a realization of the environment, $d(\tilde{a}, e) = \mathbb{E}[\overline{r}(A_\star) - \overline{r}(a) | \mathcal{E} = e]$, they study the corresponding rate-distortion function

$$\mathcal{R}(D) = \inf_{\tilde{A} \in \tilde{\mathcal{A}}} \mathbb{I}(\tilde{A}; \mathcal{E}) \tag{2}$$

where $\tilde{\mathcal{A}} = \{\tilde{A} : \mathbb{E}\left[d(\tilde{A}, \mathcal{E})\right] \leq D, \tilde{A} \perp H_t | \mathcal{E}, \forall t\}$ denotes the set of all random variables $\tilde{A} : \Omega \mapsto \mathcal{A}$ that are conditionally-independent from all histories given the environment $\mathcal{E}$ and adhere to the distortion constraint. Applying

Fact 1, we immediately recover the following:

Fact 2. *For any* $D > 0$, $\mathbb{H}(A_\star) \geq \mathbb{H}(A_\star) - \mathbb{H}(A_\star|\mathcal{E}) = \mathbb{I}(A_\star; \mathcal{E}) = \mathcal{R}(0) \geq \mathcal{R}(D) = \mathbb{I}(\tilde{A}; \mathcal{E})$

which confirms a crucial desideratum for satisficing actions; namely, that an agent must acquire fewer bits of information about $\mathcal{E}$ in order to learn a satisficing action, relative to learning an optimal action. Moreover, following an analogue of the information-theoretic analysis of Russo & Van Roy (2016), Russo & Van Roy (2018b) prove an information-theoretic regret bound that depends on the value of the rate-distortion function, rather than the entropy. While this performance guarantee highlights an interesting and useful link between sequential decision-making and rate-distortion theory, there is no guarantee that a manually-specified satisficing action $\tilde{A}$ will achieve the rate-distortion limit as desired. Thus, an agent that can manufacture its own satisficing actions which achieve the rate-distortion limit stands to dramatically outperform any hand-crafted $\tilde{A}$. To make the distinction between the manually-specified satisficing actions of prior work, we use the term *target actions* to refer to the agent's self-designated learning targets which explicitly differ from satisficing actions in that the are (1) computed by the agent, (2) adapted over time according the agent's current knowledge of the environment $\mathcal{E}$, and (3) achieve the rate-distortion limit in each time period.

Agents we consider can forgo the aim of learning an optimal action and instead try to learn a *target action*. Formally, a *target action* $\tilde{A}$ is a random variable that be thought of as generated by an inadmissible policy $\tilde{\pi}(a) = \mathbb{P}(\tilde{A} = a|\mathcal{E})$. Similarly with $A_\star$, a target action may depend on the environment, not just the history. Moreover, a target action is a random variable $\tilde{A}$ that satisfies $H_t \perp \tilde{A}|\mathcal{E}$ for all t. In other words, observations do not provide information about $\tilde{A}$ beyond what the environment would. As it based upon an inadmissible policy, a target action can change along with the agent's beliefs over the environment $\mathbb{P}(\mathcal{E} \in \cdot|H_t)$. This represents another key distinction between target actions that an agent can modify to reflect its updated knowledge about the environment and manually-specified satisficing actions that act as a fixed learning objective (much like optimal actions $A_\star$). We use $\tilde{A}_t$ to denote the target action computed in time period t according to the distortion function $d(a, e|H_t) = \mathbb{E}[(\overline{r}(A_\star) - \overline{r}(a))^2|\mathcal{E} = e, H_t])$. Consequently, this induces a sequence of rate-distortion functions, one for each time period, each of which is conditioned on the agent's history H_t. In the next section, we discuss a classic approach for computing a single, arbitrary rate-distortion function before introducing a variant of Thompson sampling that applies this method to compute target actions in each time period.

4. Approach

4.1. Notation

At various points going forward, it will be necessary to refer to the mutual information between two random variables conditioned upon a specific realization of an agent's history at some time period t. For convenience, we will denote this as

$$\mathbb{I}_t(X; Y) = \mathbb{I}(X; Y|H_t = H_t).$$

This notation will also apply analogously to the conditional mutual information

$$\mathbb{I}_t(X; Y|Z) = \mathbb{I}(X; Y|H_t = H_t, Z).$$

Note that their dependence on the realization of random history H_t makes both $\mathbb{I}_t(X; Y)$ and $\mathbb{I}_t(X; Y|Z)$ random variables themselves. The traditional notion of conditional mutual information which uses the random variable H_t arises by integrating over this randomness:

$$\mathbb{E}\left[\mathbb{I}_t(X; Y)\right] = \mathbb{I}(X; Y|H_t)$$
$$\mathbb{E}\left[\mathbb{I}_t(X; Y|Z)\right] = \mathbb{I}(X; Y|H_t, Z)$$

Additionally, we will also adopt a similar notation to express a conditional expectation given the random history H_t:

$$\mathbb{E}_t\left[X\right] = \mathbb{E}\left[X|H_t\right].$$

4.2. Blahut-Arimoto Satisficing Thompson Sampling

A classic algorithm for carrying out the constrained optimization problem captured in the rate-distortion function is the Blahut-Arimoto algorithm (Blahut, 1972; Arimoto, 1972). While the first step in the derivation of the algorithm is to start with the Lagrangian of the constrained objective, we will adopt a different notation to recognize the sequence of rate-distortion functions an agent must solve as its history expands. Namely, consider a loss function that, given history H_t, assesses a target action:

$$\mathcal{L}_\beta(\tilde{A}|H_t) = \mathbb{I}_t(\mathcal{E}; \tilde{A}) + \beta\mathbb{E}_t\left[(\overline{r}(A_\star) - \overline{r}(\tilde{A}))^2\right].$$

The first term can be interpreted as the number of bits of information from the environment required to identify target action, which we refer to as the *information rate* of $\tilde{A}$. The second term is a measure of distortion – the expected squared error between mean rewards generated by the target action versus an optimal action – scaled by a constant $\beta \in \mathbb{R}_{\geq 0}$ representing a Lagrange multiplier. Hence, this loss-function captures a rate-distortion trade-off. An optimal action minimizes distortion but, via Fact 2, may require a high rate. An uninformed action has a rate of zero but results in high distortion. Our goal is for an agent designer to use the β hyperparameter to express a preference for the ease of

learning versus the tolerable level of sub-optimality whereas it is the agent's responsibility to identify the appropriate target action $\tilde{A}_t$ that best reflects these preferences (Singh et al., 2010).

The Blahut-Arimoto Algorithm can be applied to identify a target action $\tilde{A}$ that minimizes this loss function. The algorithm is initialized with environment-dependent target action probabilities $\tilde{p}_0$, and generates a sequence of iterates $\tilde{p}_1, \tilde{p}_2, \ldots$, converging on probabilities $\tilde{p}_\star$ such that $\tilde{p}_\star(a|e) = \mathbb{P}(\tilde{A} = a|\mathcal{E} = e)$ for all $a \in \mathcal{A}$ and $e \in \Theta$. Each iteration carries out two steps. The first computes marginal probabilities of the target action

$$\tilde{q}_k(a) = \mathbb{E}_t[\tilde{p}_k(a|\mathcal{E})] \qquad \forall a \in \mathcal{A},$$

while the second updates environment-dependent target action probabilities, $\forall a \in \mathcal{A}, e \in \Theta$,

$$\tilde{p}_{k+1}(a|e) = \frac{\tilde{q}_k(a)\exp(-\beta\mathbb{E}_t[(\overline{r}(A_\star)-\overline{r}(a))^2|\mathcal{E}=e])}{\sum_{a'\in\mathcal{A}}\tilde{q}_k(a')\exp(-\beta\mathbb{E}_t[(\overline{r}(A_\star)-\overline{r}(a'))^2|\mathcal{E}=e])}.$$

A standard choice for the initial channel parameters $\tilde{p}_0(a|e)$ is the uniform distribution. Again, β now subsumes the role of D in Equation 2 for expressing the desired prioritization of minimizing rate (lower $\mathbb{I}(\tilde{A}_t; \mathcal{E})$) versus minimizing distortion (lower $d(a, e|H_t) = \mathbb{E}_t[(\overline{r}(A_\star) - \overline{r}(a))^2|\mathcal{E} = e]$)). Notice that as $\beta \to \infty$, $\tilde{p}_{k+1}(a|e)$ sharpens to a max, placing all probability mass on the realization of $\tilde{A}$ that minimizes distortion; consequently, $\tilde{p}_\star(a|e) = \mathbb{P}(\tilde{A} = a|\mathcal{E} = e) = \mathbb{P}(A_\star = a|\mathcal{E} = e)$ and we recover the standard learning target of Thompson sampling.

Just as Thompson sampling selects actions according to the probability of being optimal $\mathbb{P}(A_t = a|H_{t-1}) = \mathbb{P}(A^\star = a|H_{t-1})$, our BLahut-Arimoto Satisficing Thompson Sampling (BLASTS) algorithm selects actions according to their probability of being the target action $\tilde{A}_t$ that achieves the rate-distortion limit. We present the BLASTS algorithm as Algorithm 2.

5. Regret Analysis

Abstracting away the precise details of BLASTS, we can consider a coarsely-defined algorithm that selects each action A_t as follows: **(1)** identify a target action $\tilde{A}_t$ that minimizes a loss function $\mathcal{L}_\beta(\cdot|H_t)$ and **(2)** sample $A_t \sim \mathbb{P}(\tilde{A}_t = \cdot|H_t)$. Recall that the loss function is defined, for any target action $\tilde{A}$, by

$$\mathcal{L}_\beta(\tilde{A}|H_t) = \mathbb{I}_t(\mathcal{E}; \tilde{A}) + \beta\mathbb{E}_t\left[(\overline{r}(A_\star) - \overline{r}(\tilde{A}))^2\right].$$

Due to space constraints, the proofs associated with all of the following results can be found in the appendix. The following result helps establish that the expected loss of any target action decreases as observations accumulate.

Algorithm 1 Blahut-Arimoto Satisficing Thompson Sampling (BLASTS)

Input: Lagrange multiplier $\beta \in \mathbb{R}_{\geq 0}$, Blahut-Arimoto iterations $K \in \mathbb{N}$, Posterior samples $Z \in \mathbb{N}$
$H_0 = \{\}$
for $t = 0$ **to** $T - 1$ **do**
  $e_1, \ldots, e_Z \sim \mathbb{P}(\mathcal{E} \in \cdot|H_t)$
  $d(a, e|H_t) = \mathbb{E}[(\overline{r}(A_\star) - \overline{r}(a))^2|\mathcal{E} = e, H_t])$
  $\tilde{p}_0(a|e_z) = \frac{1}{|\mathcal{A}|}, \forall a \in \mathcal{A}, z \in [Z]$
  for $k = 0$ **to** $K - 1$ **do**
    $\tilde{q}_k(a) = \mathbb{E}_t[\tilde{p}_k(a|\mathcal{E})], \forall a \in \mathcal{A}$
    $\tilde{p}_{k+1}(a|e_z) \propto \tilde{q}_k(a)\exp\left(-\beta d(a, e_z \mid H_t)\right), \forall a \in \mathcal{A}, \forall z \in Z$
  end for
  $\hat{z} \sim \text{Uniform}(Z)$
  $A_t \sim \tilde{p}_K(a|e_{\hat{z}})$
  $H_{t+1} = H_t \cup \{(A_t, O_{t+1})\}$
  $R_{t+1} = r(A_t, O_{t+1})$
end for

Lemma 1. *For all $\beta > 0$, target actions $\tilde{A}$, and $t = 0, 1, 2, \ldots$,*

$$\mathbb{E}_t[\mathcal{L}_\beta(\tilde{A}|H_{t+1})] = \mathcal{L}_\beta(\tilde{A}|H_t) - \mathbb{I}_t(\tilde{A}; (A_t, O_{t+1})).$$

As a consequence of the above, the following lemma assures that expected loss decreases as target actions are adapted. It also suggests that there are two sources of decrease in loss: (1) a possible decrease in shifting from target $\tilde{A}_t$ to $\tilde{A}_{t+1}$ and (2) a decrease of $\mathbb{I}_t(\tilde{A}_t; (A_t, O_{t+1}))$ from observing the interaction (A_t, O_{t+1}). The former reflects the agent's improved ability to select a suitable target, and the latter captures information gained about the previous target. The proof of the lemma follows immediately from Lemma 1 and the fact that $\tilde{A}_{t+1}$ minimizes $\mathcal{L}_\beta(\tilde{A}_{t+1}|H_{t+1})$, by definition.

Lemma 2. *For all $\beta > 0$, target actions $\tilde{A}$, and $t = 0, 1, 2, \ldots$,*

$$\mathbb{E}[\mathcal{L}_\beta(\tilde{A}_{t+1}|H_{t+1})|H_t] \leq \mathcal{L}_\beta(\tilde{A}_t|H_t) - \mathbb{I}_t(\tilde{A}_t; (A_t, O_{t+1})).$$

Note that, for all t, loss is non-negative and bounded by mutual information between the optimal action and the environment (since optimal actions incur a distortion of 0):

$$\mathcal{L}_\beta(\tilde{A}_t|H_t) \leq \mathcal{L}_\beta(A_\star|H_t) = \mathbb{I}_t(\mathcal{E}; A_\star).$$

We therefore have the following corollary.

Corollary 1. *For all $\beta > 0$ and $\tau = 0, 1, 2, \ldots$,*

$$\mathbb{E}\left[\sum_{t=\tau}^{\infty}\mathbb{I}_t(\tilde{A}_t; (A_t, O_{t+1}))\Big|H_\tau\right] \leq \mathbb{I}_\tau(\mathcal{E}; A_\star).$$

The proof of Corollary 1 follows directly by applying the preceding inequality to the following generalization that applies to any target action.

Corollary 2. *For all $\beta > 0$, target actions $\tilde{A}$, and $\tau = 0, 1, 2, \ldots$,*

$$\mathbb{E}_\tau \left[\sum_{t=\tau}^{\infty} \mathbb{I}_t(\tilde{A}_t; (A_t, O_{t+1})) \right] \leq \mathcal{L}_\beta(\tilde{A} | H_\tau).$$

Let Γ be a constant such that

$$\Gamma \geq \frac{\mathbb{E}_t[\overline{r}(\tilde{A}) - \overline{r}(A)]^2}{\mathbb{I}_t(\tilde{A}; A, O)},$$

for all histories H_t, target actions $\tilde{A}$, if the executed action A is an independent sample drawn from the marginal distribution of $\tilde{A}$, and O is the resulting observation. Thus, Γ is an upper bound on the information ratio (Russo & Van Roy, 2014; 2016; 2018a) for which existing information-theoretic analyses of worst-case finite-arm bandits and linear bandits provide explicit values of Γ that satisfy this condition.

We can now establish our main results. We omit the proof of Theorem 1 as it is a special case of our subsequent result.

Theorem 1. *If $\beta = \frac{1-\gamma^2}{(1-\gamma)^2\Gamma}$ then, for all $\tau = 0, 1, 2, \ldots$,*

$$\mathbb{E}_\tau \left[\sum_{t=\tau}^{\infty} \gamma^{t-\tau} (\overline{r}(A_\star) - \overline{r}(A_t)) \right] \leq 2\sqrt{\frac{\Gamma \mathbb{I}_\tau(\mathcal{E}; A_\star)}{1-\gamma^2}}.$$

In a complex environment with many actions, $\mathbb{I}(\mathcal{E}; A_\star)$ can be extremely large, rendering the above result somewhat vacuous under such circumstances. The next result offers a generalization, establishing a regret bound that can depend on the information content of any target action, including of course those that are much simpler than $A_\star$.

Theorem 2. *If $\beta = \frac{1-\gamma^2}{(1-\gamma)^2\Gamma}$ then, for all target actions $\tilde{A}$ and $\tau = 0, 1, 2, \ldots$,*

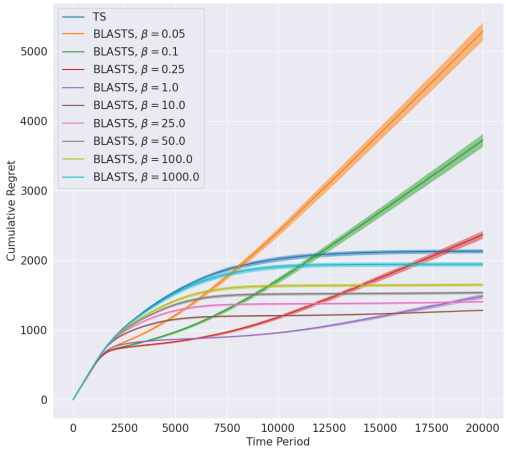
$$\mathbb{E}_\tau \left[\sum_{t=\tau}^{\infty} \gamma^{t-\tau} (\overline{r}(A_\star) - \overline{r}(A_t)) \right] \leq 2\sqrt{\frac{\Gamma \mathbb{I}_\tau(\mathcal{E}; \tilde{A})}{1-\gamma^2}} + \frac{2\epsilon}{1-\gamma},$$

where $\epsilon = \sqrt{\mathbb{E}_\tau[(\overline{r}(A_\star) - \overline{r}(\tilde{A})^2]}$.

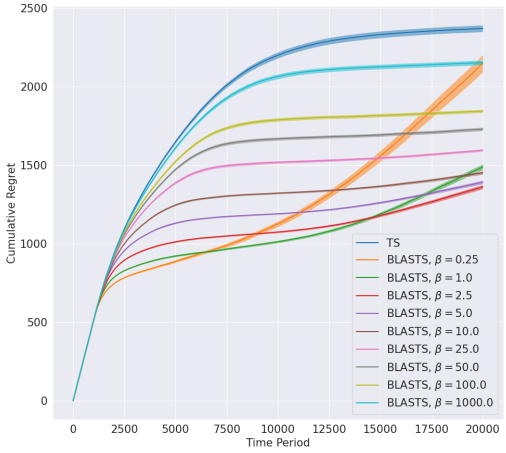
For the sake of completeness, we may derive the analogues of Corollary 2 and Theorem 2 for the more traditional finite-horizon, undiscounted regret setting.

Corollary 3. *For all $\beta > 0$, target actions $\tilde{A}$, and $\tau = 0, 1, 2, \ldots$,*

$$\mathbb{E}_\tau \left[\sum_{t=\tau}^{T+\tau} \mathbb{I}_t(\tilde{A}_t; (A_t, O_{t+1})) \right] \leq \mathcal{L}_\beta(\tilde{A} | H_\tau).$$



(a) 50 arms



(b) 250 arms

Figure 1. Bernoulli bandit with independent arms

Theorem 3. *If $\beta = \frac{T}{\Gamma}$ then, for all target actions $\tilde{A}$ and $\tau = 0, 1, 2, \ldots$,*

$$\mathbb{E}_\tau \left[\sum_{t=\tau}^{T+\tau} \overline{r}(A_\star) - \overline{r}(A_t) \right] \leq 2\sqrt{\Gamma T \mathbb{I}_\tau(\mathcal{E}; \tilde{A})} + 2T\epsilon,$$

where $\epsilon = \sqrt{\mathbb{E}[(\overline{r}(A_\star) - \overline{r}(\tilde{A})^2 | H_\tau]}$.

Notably, the information-theoretic regret bounds of Theorems 2 and 3 align with that of (Russo & Van Roy, 2018b) as a sum of the difficulty associated with learning $\tilde{A}$ and the associated performance shortfall between $\tilde{A}$ and $A_\star$.

6. Experiments

In this section, we outline two sets of computational experiments that evaluate BLASTS against traditional Thompson sampling (TS). The primary goal of our experiments is to illustrate how BLASTS enables an agent to navigate the

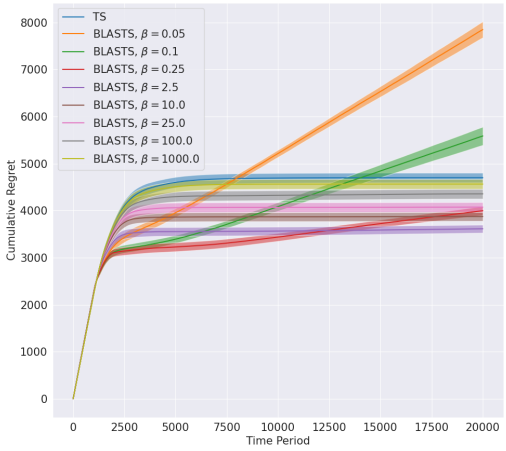
information-performance trade-off through the specification of β. To this end, we examine two commonly-studied multi-armed bandit problems and sweep across several values of β, benchmarking performance relative to Thompson sampling. In the course of doing so, we find that both settings offer a range of β values which allow the agent to converge on the optimal policy with greater efficiency than Thompson sampling.

In all of our experiments, we use linear hypermodels (Dwaracherla et al., 2020) as a common choice for representing an agent's epistemic uncertainty over the environment $\mathcal{E}$. While several prior works have made use of finite ensembles for representing an agent's posterior beliefs over environment parameters (Osband et al., 2016; Lu & Van Roy, 2017), hypermodels offer a more computationally-tractable approach that demonstrably scales better with a large number of actions. For an independent multi-armed bandit problem with K actions, a linear hypermodel takes as input an index sample $z \sim \mathcal{N}(0, I_K)$ and computes a single posterior sample as $f_\nu(z) = \mu + \sigma z$ where the parameters $\nu = (\mu \in \mathbb{R}^K, \sigma \in \mathbb{R}^K)$ are incrementally updated via gradient descent to minimize a bootstrapped loss function. Due to space constraints, we refer readers to (Dwaracherla et al., 2020) for the precise details of this loss function and further information about hypermodels. It is important to note that both Thompson sampling and BLASTS are agnostic to this modeling choice and are compatible with any approach for representing an agent's uncertainty about the environment. We use a noise variance of 0.1, a prior variance of 1.0, and a batch size of 1024 throughout all experiments while using Adam (Kingma & Ba, 2014) to optimize hypermodel parameters with a learning rate of 0.001.

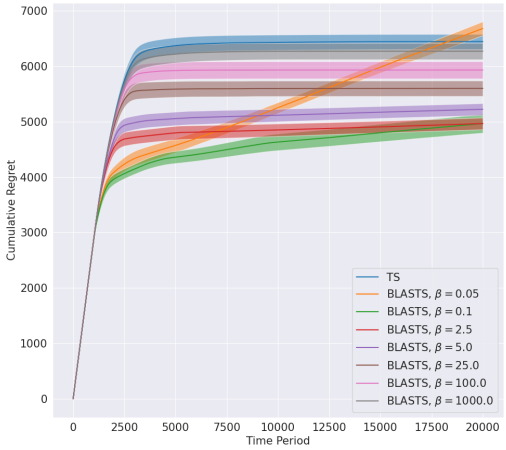
We leverage an existing implementation of the Blahut-Arimoto algorithm for all experiments (James et al., 2018). The number of posterior samples used was fixed to 64 and the maximum number of iterations was set to 100, stopping early if the average distortion between two consecutive iterations fell below a small threshold. In preliminary experiments, we found better numerical stability when running the Blahut-Arimoto algorithm in base 2, rather than base e. To benchmark performance, we plot the (undiscounted) cumulative regret in each time period with shading to represent 95% confidence intervals computed across 10 random seeds.

6.1. Independent Bernoulli & Gaussian Bandits

Our first experiment focuses on a Bernoulli bandit with K independent arms. In each random trial, the environment is represented as a vector $\mathcal{E} \in \mathbb{R}^K$ where $\mathcal{E}_a \sim \text{Uniform}(0, 1), \forall a \in \mathcal{A}$. Accordingly, the reward observed for taking action $a \in \mathcal{A}$ is sampled as a Bernoulli$(\mathcal{E}_a)$. In our second experiment, we pivot to a Gaussian bandit where rewards for action a are drawn from $\mathcal{N}(\mathcal{E}_a, 1)$, again with $\mathcal{E}_a \sim \text{Uniform}(0, 1), \forall a \in \mathcal{A}$. Results for each experiment are shown in Figures 1 and 2, respectively.



(a) 50 arms



(b) 250 arms

Figure 2. Gaussian bandit with independent arms

The first notable observation from both sets of experiments is the control that the β parameter exerts over the performance of BLASTS. As expected, while $\beta \to 0$, BLASTS approaches the performance of a uniform random policy. In contrast, as $\beta \to \infty$, BLASTS gradually recovers the performance of Thompson sampling. Importantly, when obtaining a satisficing solution is viable, there is a suitable range of β values to accommodate different degrees of sub-optimality, many of which converge to such satisficing policies in fewer time periods than what is needed for an optimal policy. In our experiments, we ran BLASTS for a wider range of β values than what is shown and selectively pruned away a subset of values for readability. In all plots, the smallest value of β in our selection that achieves the optimal policy is shown.

A second key finding of the above experiments is the capacity for BLASTS to synthesize an optimal policy more efficiently than Thompson sampling. Recall that the input D to the rate-distortion function $\mathcal{R}(D)$ represents the desired upper bound on expected distortion. In the context of the Blahut-Arimoto algorithm, β represents the desired slope of the recovered solution along the rate-distortion curve. By Corollary 5 of (Blahut, 1972), we know that, given the current history H_t, the distortion D achieved at the point on the rate-distortion curve parameterized by β is given as $D(\beta|H_t) = \mathbb{E}\left[\frac{\tilde{q}_\star(A)\exp(-\beta\mathbb{E}[(\overline{r}(A_\star)-\overline{r}(A))^2|\mathcal{E},H_t])}{\sum_{a'\in\mathcal{A}}\tilde{q}_\star(a')\exp(-\beta\mathbb{E}[(\overline{r}(A_\star)-\overline{r}(a'))^2|\mathcal{E},H_t])}\right]$, where $\tilde{q}_\star$ achieves the infimum

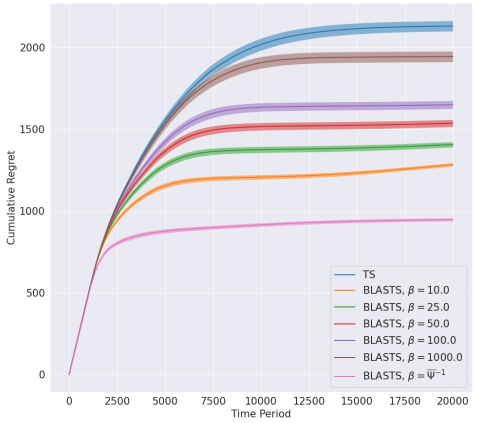
$$\inf_q -\mathbb{E}\left[\log\left(\sum_{a\in\mathcal{A}} q(a)\exp(-\beta\mathbb{E}_t[(\overline{r}(A_\star)-\overline{r}(A))^2|\mathcal{E}]\right)\right].$$

Letting $\Delta > 0$ denote the action gap between the best and second-best arm (Farahmand, 2011; Bellemare et al., 2016), it stands to reason that, for any β obtaining the optimal policy, $\max_t D(\beta|H_t) < \Delta^2$. By Fact 2, it follows that the target actions computed along these same β values serve as easier learning targets (through smaller $\mathbb{I}_t(\tilde{A};\mathcal{E})$) while still converging to the optimal policy.

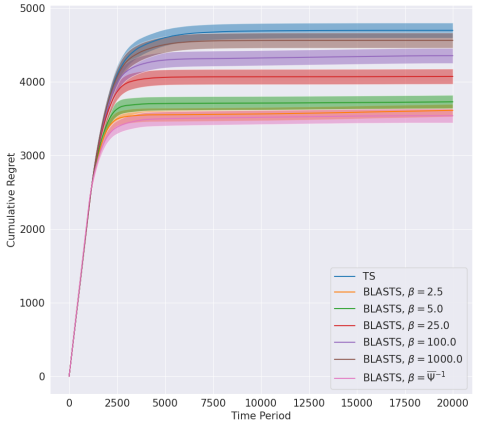
In summary, the results presented here verify that BLASTS is capable of realizing a broad spectrum of policies. Included in this spectrum are satisficing policies that accommodate various problem constraints on time and resources, as well as optimal policies that be identified with greater efficiency than Thompson sampling.

6.2. Balancing Rate-Distortion with the Information Ratio

The previous experiments clearly illustrate the importance of the β hyperparameter in enabling an agent designer to express preferences over behaviors and allowing an agent to realize those preferences through its learned target actions. In the context of the rate-distortion function, β encodes a preference for minimizing rate over minimizing distortion. Some of the β values that ultimately recover satisficing policies, however, do appear to have signs of strong performance in the earlier stages of learning. However, it is clear that despite this initial potential, the fixed value of β is ultimately too small to prioritize regret minimization. It is a natural to wonder if allowing β to vary with time might more efficiently synthesize an optimal policy? One crude strategy for exploring this would be to place β on a manually-tuned schedule, eventually allowing it to increase to a value that emphasizes optimal actions by the end of learning. As a more principled alternative to such a laborious strategy, we consider the relationship between β and the information ratio, inspired by the value of $\beta = \frac{1-\gamma^2}{(1-\gamma)^2\Gamma}$ derived in our analysis.



(a) Bernoulli



(b) Gaussian

Figure 3. BLASTS with adaptive $\beta_t = \overline{\Psi}_t^{-1}$ for independent bandits with 50 arms

The information ratio (Russo & Van Roy, 2014; 2016; 2018a) is a powerful tool for expressing the cost (measured in squared units of regret) per bit of information acquired in each time period. The constant Γ in our analysis acts a uniform upper bound on the information ratio (for our setting) that facilitates our information-theoretic regret bounds. For the more traditional setting of finding optimal policies, the information ratio at time period t is given by $\Psi_t(\pi) = \frac{\Delta_t(\pi)^2}{g_t(\pi)}$ where $\Delta_t(\pi)$ denotes the expected regret with respect to $A_\star$ and $g_t(\pi)$ denotes the information gain $\mathbb{I}_t(A_\star; A_t, O_{t+1})$. While, in theory, an agent wishes to compute a policy $\pi = \min_\pi \Psi(\pi)$ that minimizes the information ratio, practical instantiations of this principle often rely on the fact that $g_t(\pi) \geq \mathbb{E}[v_t(A)]$ where $v_t(A) = \mathbb{V}[\overline{r}(A)|\mathcal{E}]|H_t]$ is the variance of the expected reward for action A conditioned on the agent's current beliefs over the environment $\mathcal{E}$ (Russo & Van Roy, 2014;

2018a). Consequently, in each time period, an agent may aim to compute a policy that minimizes an upper bound $\overline{\Psi}(\pi) = \frac{\Delta_t(\pi)^2}{v_t(\pi)}$. To see an initial connection between β and the information ratio, recall that β is representative of the desired slope along the rate-distortion curve (Blahut, 1972), with units of bits per unit of distortion; since BLASTS operates with a squared-regret distortion, this leaves β as a quantity with units of bits per squared unit of regret. Moreover, once an agent has resolved most of its uncertainty in the environment, small values of the information ratio are indicative of optimal policies where BLASTS should, ideally, take on larger values of β to identify such optimal actions. In light of these connections, we experiment with a version of BLASTS that uses the minimizer of the variance-based information ratio to compute β in each time period. More specifically, let $\overline{\Psi}_t = \min_{\pi \in \Delta(\mathcal{A})} \overline{\Psi}_t(\pi)$ and take $\beta_t = \overline{\Psi}_t^{-1}$; small constant is always added to $\overline{\Psi}_t$ to avoid division by zero. Results for this variant on the independent Bernoulli and Gaussian bandits are shown in Figure 3. While an adaptive β shows marginal gain in the Gaussian bandit, the Bernoulli bandit results show marked improvement in finding an optimal policy.

These results using an adaptive β_t can be translated back to the fixed β setting by considering a distortion function $\hat{d}(\tilde{a}, e) = \overline{\Psi}^{-1} d(\tilde{a}, \mathcal{E})$. Our choice of using expected squared distortion is supported by our theory, however the question of whether more efficient distortion functions exist in practice is an interesting direction for future work.

7. Conclusion

A standard design principle of sequential decision-making is to build agents that learn optimal actions. Recent work has highlighted scenarios wherein problem constraints make the pursuit of optimal actions infeasible, forcing the agent designer to craft a new target for an agent to learn. In this work, we forge a new direction where agents are designed to fabricate their own learning targets whose generic form is now the sole responsibility of the agent designer. We highlight how rate-distortion theory gives rise to a principled form for these learning targets, allowing practitioners to express their preference between the ease of learning and the sub-optimality of the resulting policy. We prove a general regret bound for this setting, contending with the non-stationarity of learning targets, and empirically verify the flexibility of our approach in yielding a broad spectrum of policies with varying degrees of sub-optimality. Importantly, we find that an agent's ability to specify target actions that require fewer bits of information can translate into greater efficiency in finding optimal policies relative to Thompson sampling. Future work may find it fruitful to couple the Blahut-Arimoto algorithm with more powerful strategies for information acquisition (Russo & Van Roy, 2018a).

Acknowledgements

Financial support from Army Research Office (ARO) grant W911NF2010055 is gratefully acknowledged.

References

Abel, D., Arumugam, D., Asadi, K., Jinnai, Y., Littman, M. L., and Wong, L. L. State abstraction as compression in apprenticeship learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3134–3142, 2019.

Agrawal, S. and Goyal, N. Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pp. 39–1, 2012.

Agrawal, S. and Goyal, N. Further optimal regret bounds for Thompson sampling. In *Artificial intelligence and statistics*, pp. 99–107, 2013.

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

Bacon, P.-L., Harb, J., and Precup, D. The option-critic architecture. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.

Barto, A. G. and Mahadevan, S. Recent advances in hierarchical reinforcement learning. *Discrete event dynamic systems*, 13(1-2):41–77, 2003.

Bellemare, M. G., Ostrovski, G., Guez, A., Thomas, P., and Munos, R. Increasing the action gap: New operators for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

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

Berry, D. A., Chen, R. W., Zame, A., Heath, D. C., and Shepp, L. A. Bandit problems with infinitely many arms. *Ann. Statist.*, 25(5):2103–2116, 10 1997. doi: 10.1214/aos/1069362389. URL https://doi.org/10.1214/aos/1069362389.

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

Bonald, T. and Proutiere, A. Two-target algorithms for infinite-armed bandits with bernoulli rewards. In *Advances in Neural Information Processing Systems*, pp. 2184–2192, 2013.

Bubeck, S., Munos, R., Stoltz, G., and Szepesvári, C. X-armed bandits. *Journal of Machine Learning Research*, 12(5), 2011.

Bubeck, S., Cesa-Bianchi, N., et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5 (1):1–122, 2012.

Chapelle, O. and Li, L. An empirical evaluation of Thompson sampling. In *Advances in neural information processing systems*, pp. 2249–2257, 2011.

Cover, T. M. and Thomas, J. A. *Elements of information theory*. John Wiley & Sons, 2012.

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

Csiszár, I. and Tsunády, G. Information geometry and alternating minimization procedures. *Statistics and decisions*, 1:205–237, 1984.

Dayan, P. and Hinton, G. E. Feudal reinforcement learning. In *Advances in Neural Information Processing Systems*, 1993.

Deshpande, Y. and Montanari, A. Linear bandits in high dimension and recommendation systems. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1750–1754. IEEE, 2012.

Dong, S. and Van Roy, B. An information-theoretic analysis for Thompson sampling with many actions. In *Advances in Neural Information Processing Systems*, pp. 4157–4165, 2018.

Dwaracherla, V., Lu, X., Ibrahimi, M., Osband, I., Wen, Z., and Van Roy, B. Hypermodels for exploration. In *International Conference on Learning Representations*, 2020.

Farahmand, A.-m. Action-gap phenomenon in reinforcement learning. *Advances in Neural Information Processing Systems*, 24:172–180, 2011.

Harb, J., Bacon, P.-L., Klissarov, M., and Precup, D. When waiting is not an option: Learning options with a deliberation cost. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

James, R. G., Ellison, C. J., and Crutchfield, J. P. dit: a Python package for discrete information theory. *The Journal of Open Source Software*, 3(25):738, 2018. doi: https://doi.org/10.21105/joss.00738.

Jong, N. K., Hester, T., and Stone, P. The utility of temporal abstraction in reinforcement learning. Citeseer, 2008.

Kaelbling, L. P. Hierarchical learning in stochastic domains: preliminary results. In *Proceedings of the Tenth International Conference on International Conference on Machine Learning*, pp. 167–173, 1993.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kleinberg, R., Slivkins, A., and Upfal, E. Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pp. 681–690, 2008.

Lattimore, T. and Szepesvári, C. An information-theoretic approach to minimax regret in partial monitoring. *arXiv preprint arXiv:1902.00470*, 2019.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Lu, X. and Van Roy, B. Ensemble sampling. In *Advances in neural information processing systems*, pp. 3258–3266, 2017.

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

Nachum, O., Gu, S. S., Lee, H., and Levine, S. Data-efficient hierarchical reinforcement learning. In *Advances in neural information processing systems*, pp. 3303–3313, 2018.

Naja, Z., Alberge, F., and Duhamel, P. 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.

Osband, I., Blundell, C., Pritzel, A., and Van Roy, B. Deep exploration via bootstrapped dqn. In *Advances in neural information processing systems*, pp. 4026–4034, 2016.

Osband, I., Van Roy, B., Russo, D. J., and Wen, Z. Deep exploration via randomized value functions. *Journal of Machine Learning Research*, 20(124):1–62, 2019.

Rusmevichientong, P. and Tsitsiklis, J. N. Linearly parameterized bandits. *Mathematics of Operations Research*, 35 (2):395–411, 2010.

Russo, D. and Van Roy, B. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems*, pp. 1583–1591, 2014.

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

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

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

Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., and Wen, Z. A tutorial on Thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

Ryzhov, I. O., Powell, W. B., and Frazier, P. I. The knowledge gradient algorithm for a general class of online learning problems. *Operations Research*, 60(1):180–195, 2012.

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

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

Singh, S., Lewis, R. L., Sorg, J., Barto, A. G., and Helou, A. On separating agent designer goals from agent goals: Breaking the preferences–parameters confound, 2010.

Sutton, R. S., Precup, D., and Singh, S. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 1999.

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

Vezhnevets, A. S., Osindero, S., Schaul, T., Heess, N., Jaderberg, M., Silver, D., and Kavukcuoglu, K. Feudal networks for hierarchical reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3540–3549, 2017.

Vontobel, P. O., Kavcic, A., Arnold, D. M., and Loeliger, H.-A. A generalization of the Blahut–Arimoto algorithm to finite-state channels. *IEEE Transactions on Information Theory*, 54(5):1887–1918, 2008.

Wang, Y., Audibert, J., and Munos, R. Algorithms for infinitely many-armed bandits. In *NIPS*, 2008.

Wen, Z., Precup, D., Ibrahimi, M., Barreto, A., Van Roy, B., and Singh, S. On efficiency in hierarchical reinforcement learning. *Advances in Neural Information Processing Systems*, 33, 2020.

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

A. Related Work

Our work focuses on principled Bayesian exploration wherein an agent maintains a posterior distribution over its environment (Chapelle & Li, 2011; Agrawal & Goyal, 2012; 2013; Russo & Van Roy, 2016). As complete knowledge of the environment (the vector of mean rewards at each arm, for example) would endow an agent with prescience of optimal actions, efficient exploration amounts to the resolution of an agent's epistemic uncertainty about the environment. A natural approach for resolving such uncertainty is Thompson sampling which employs probability matching in each time period to sample actions according to the probability of being optimal (Thompson, 1933; Agrawal & Goyal, 2012; 2013; Russo & Van Roy, 2016; Russo et al., 2018). Chapelle & Li (2011) kickstarted renewed interest in Thompson sampling through empirical successes in online advertisement and news recommendation applications. While a corresponding regret bound was developed in subsequent work (Agrawal & Goyal, 2012; 2013), our paper follows suit with Russo & Van Roy (2016) who introduced an elegant, information-theoretic analysis of Thompson sampling; their technique has been subsequently studied and extended to a variety of other problem settings (Russo & Van Roy, 2018a;b; Dong & Van Roy, 2018) and applications (Lattimore & Szepesvári, 2019; Osband et al., 2019). In this work, we also leverage the information-theoretic analysis of Russo & Van Roy (2016) while additionally incorporating ideas from rate-distortion theory (Shannon, 1959). Unlike prior work exploring the intersection of sequential decision-making and rate-distortion theory, we are not concerned with state abstraction (Abel et al., 2019) nor are we concerned with an agent exclusively targeting optimal actions through some compressive statistic of the environment (Dong & Van Roy, 2018).

A core novelty of this paper is leveraging the Blahut-Arimoto algorithm (Arimoto, 1972; Blahut, 1972) for the efficient computation of rate-distortion functions. The algorithm was originally developed for the dual problem of computing the channel-capacity function (Arimoto, 1972) and was soon after extended to handle computation of the rate-distortion function as well (Blahut, 1972). An initial study of the algorithm's global convergence properties (for discrete random variables) was done by Arimoto (1972) and further explored by Csiszár (1974); Csiszár & Tsunády (1984). While there have been many variants of the Blahut-Arimoto algorithm introduced over the years (Sayir, 2000; Matz & Duhamel, 2004; Vontobel et al., 2008; Naja et al., 2009; Yu, 2010), we find that the simplicity of the original algorithm is suitable both in theory and in practice.

The goal of finding target actions with a tolerable degree of sub-optimality deviates from the more traditional objective of identifying optimal actions. As previously mentioned, this setting can implicitly arise when faced with a continuous action space (Bubeck et al., 2011; Kleinberg et al., 2008; Rusmevichientong & Tsitsiklis, 2010), a fixed time horizon (Ryzhov et al., 2012; Deshpande & Montanari, 2012), or an infinite-armed bandit problem (Berry et al., 1997; Wang et al., 2008; Bonald & Proutiere, 2013). Russo & Van Roy (2018b) attempt to rectify some shortcomings of these works by introducing a discounted notion of regret that emphasizes initial stages of learning and measures performance shortfall relative to satisficing actions, instead of optimal ones. Moreover, the analysis of their satisficing Thompson sampling algorithm inherits the benefits of flexibility and generality from the analogous information-theoretic results for Thompson sampling (Russo & Van Roy, 2016). In this work, we obviate the need for the manual specification of satisficing actions, instead relying on direct computation of the rate-distortion function to adaptively compute the distribution over satisficing actions in each time period that achieves the rate-distortion limit.

The idea of an agent that learns to designate and achieve its own goals bears close resemblance to hierarchical agents studied in the reinforcement-learning literature (Kaelbling, 1993; Dayan & Hinton, 1993; Sutton et al., 1999; Barto & Mahadevan, 2003). In recent years, the two most-popular paradigms for hierarchical reinforcement learning have been feudal reinforcement learning (Dayan & Hinton, 1993; Nachum et al., 2018) and options (Sutton et al., 1999; Jong et al., 2008; Bacon et al., 2017; Wen et al., 2020). Feudal reinforcement-learning agents are comprised of an internal managerial hierarchy wherein the action space of managers represents sub-goals for workers in the subsequent level of the hierarchy; when workers can be quickly trained to follow the directed sub-goals of their managers (without regard for the optimality of doing so) the top-most manager can more efficiently synthesize an optimal policy. Options provide a coherent abstraction for expressing various temporally-extended behaviors or skills, typically replacing or augmenting the original action space of the agent (Jong et al., 2008). While there is great empirical support for the performance of feudal learning and options when the goal representation or option set is computed and fixed a priori, recent work in learning such components online often relies on laborious tuning and heuristics to achieve success (Vezhnevets et al., 2017; Bacon et al., 2017; Harb et al., 2018). In contrast, the main contribution of this work is to build a principled approach for learning such targets, albeit with a restricted focus to the simpler setting of bandit learning. We leave the exciting question of how the ideas presented here may scale up to tackle the challenges of hierarchical reinforcement learning to future work.

B. Blahut-Arimoto Satisficing Thompson Sampling

Here we present the full BLASTS algorithm with inline comments for clarity.

Algorithm 2 Blahut-Arimoto Satisficing Thompson Sampling (BLASTS)

Input: Lagrange multiplier $\beta \in \mathbb{R}_{\geq 0}$, Blahut-Arimoto iterations $K \in \mathbb{N}$, Posterior samples $Z \in \mathbb{N}$
$H_0 = \{\}$
for $t = 0$ **to** $T - 1$ **do**
  $e_1, \ldots, e_Z \sim \mathbb{P}(\mathcal{E} \in \cdot | H_t)$ {Finite sample from current belief over $\mathcal{E}$}
  $d(a, e|H_t) = \mathbb{E}[(\overline{r}(A_\star) - \overline{r}(a))^2 | \mathcal{E} = e, H_t])$ {Distortion function for target action $\tilde{A}_t$}
  $\tilde{p}_0(a|e_z) = \frac{1}{|\mathcal{A}|}, \forall a \in \mathcal{A}, z \in [Z]$
  for $k = 0$ **to** $K - 1$ **do**
    $\tilde{q}_k(a) = \mathbb{E}_t[\tilde{p}_k(a|\mathcal{E})], \forall a \in \mathcal{A}$ {Run the Blahut-Arimoto algorithm}
    $\tilde{p}_{k+1}(a|e_z) = \frac{\tilde{q}_k(a) \exp(-\beta d(a, e_z|H_t))}{\sum_{a' \in \mathcal{A}} \tilde{q}_k(a') \exp(-\beta d(a, e_z|H_t))}, \forall a \in \mathcal{A}, z \in [Z]$
  end for
  $\hat{z} \sim \text{Uniform}(Z)$ {Select posterior sample uniformly at random}
  $A_t \sim \tilde{p}_K(a|e_{\hat{z}})$ {Probability matching}
  $H_{t+1} = H_t \cup \{(A_t, O_{t+1})\}$
  $R_{t+1} = r(A_t, O_{t+1})$
end for

C. Discounted Regret Analysis

Abstracting away the precise details of BLASTS, we can consider a coarsely-defined algorithm that selects each action A_t as follows: **(1)** identify a target action $\tilde{A}_t$ that minimizes a loss function $\mathcal{L}_\beta(\cdot|H_t)$ and **(2)** sample $A_t \sim \mathbb{P}(\tilde{A}_t = \cdot|H_t)$. Recall that the loss function is defined, for any target action $\tilde{A}$, by

$$\mathcal{L}_\beta(\tilde{A}|H_t) = \mathbb{I}_t(\mathcal{E}; \tilde{A}) + \beta \mathbb{E}_t \left[(\overline{r}(A_\star) - \overline{r}(\tilde{A}))^2 \right].$$

The following result helps establish that the expected loss of any target action decreases as observations accumulate.

Lemma 3. *For all $\beta > 0$, target actions $\tilde{A}$, and $t = 0, 1, 2, \ldots$,*

$$\mathbb{E}_t[\mathcal{L}_\beta(\tilde{A}|H_{t+1})] = \mathcal{L}_\beta(\tilde{A}|H_t) - \mathbb{I}_t(\tilde{A}; (A_t, O_{t+1})).$$

Proof.

Recall that $H_{t+1} = (H_t, A_t, O_{t+1})$. By definition of a target action, we have that $\forall t, H_t \perp \tilde{A}|\mathcal{E}$, which implies $\mathbb{I}_t((A_t, O_{t+1}); \tilde{A}|\mathcal{E}) = 0$. Thus,

$$\mathbb{I}_t(\mathcal{E}; \tilde{A}) = \mathbb{I}_t(\mathcal{E}; \tilde{A}) + \mathbb{I}_t((A_t, O_{t+1}); \tilde{A}|\mathcal{E}) = \mathbb{I}_t(\mathcal{E}, (A_t, O_{t+1}); \tilde{A})$$

by the chain rule of mutual information. Applying the chain rule once again, we have,

$$\mathbb{I}_t(\mathcal{E}; \tilde{A}) = \mathbb{I}_t(\mathcal{E}, (A_t, O_{t+1}); \tilde{A}) = \mathbb{I}_t(\mathcal{E}; \tilde{A}|A_t, O_{t+1}) + \mathbb{I}_t(\tilde{A}; (A_t, O_{t+1})).$$

It follows that

$$\begin{aligned}
\mathbb{E}_t[\mathcal{L}_\beta(\tilde{A}|H_{t+1})] =& \mathbb{E}[\mathcal{L}_\beta(\tilde{A}|H_{t+1})|H_t] \\
=& \mathbb{E}\left[\mathbb{I}_t(\mathcal{E};\tilde{A}|A_t,O_{t+1}) + \beta\mathbb{E}\left[(\overline{r}(A_\star)-\overline{r}(\tilde{A}))^2|H_t,A_t,O_{t+1}\right]\Big|H_t\right] \\
=& \mathbb{E}_t\left[\mathbb{I}_t(\mathcal{E};\tilde{A}|A_t,O_{t+1})\right] + \beta\mathbb{E}_t\left[(\overline{r}(A_\star)-\overline{r}(\tilde{A}))^2\right] \\
=& \mathbb{E}_t\left[\mathbb{I}_t(\mathcal{E};\tilde{A}) - \mathbb{I}_t(\tilde{A};(A_t,O_{t+1}))\right] + \beta\mathbb{E}_t\left[(\overline{r}(A_\star)-\overline{r}(\tilde{A}))^2\right] \\
=& \mathbb{I}_t(\mathcal{E};\tilde{A}) + \beta\mathbb{E}_t\left[(\overline{r}(A_\star)-\overline{r}(\tilde{A}))^2\right] - \mathbb{I}_t(\tilde{A};(A_t,O_{t+1})) \\
=& \mathcal{L}_\beta(\tilde{A}|H_t) - \mathbb{I}_t(\tilde{A};(A_t,O_{t+1})).
\end{aligned}$$

□

As a consequence of the above, the following lemma assures that expected loss decreases as target actions are adapted. It also suggests that there are two sources of decrease in loss: (1) a possible decrease in shifting from target $\tilde{A}_t$ to $\tilde{A}_{t+1}$ and (2) a decrease of $\mathbb{I}_t(\tilde{A}_t;(A_t,O_{t+1}))$ from observing the interaction (A_t,O_{t+1}). The former reflects the agent's improved ability to select a suitable target, and the latter captures information gained about the previous target. We omit the proof as the lemma follows immediately from Lemma 1 and the fact that $\tilde{A}_{t+1}$ minimizes $\mathcal{L}_\beta(\tilde{A}_{t+1}|H_{t+1})$, by definition.

Lemma 4. *For all $\beta > 0$, target actions $\tilde{A}$, and $t = 0, 1, 2, \ldots$,*

$$\mathbb{E}[\mathcal{L}_\beta(\tilde{A}_{t+1}|H_{t+1})|H_t] \leq \mathbb{E}[\mathcal{L}_\beta(\tilde{A}_t|H_{t+1})|H_t] = \mathcal{L}_\beta(\tilde{A}_t|H_t) - \mathbb{I}_t(\tilde{A}_t;(A_t,O_{t+1})).$$

Note that, for all t, loss is non-negative and bounded by mutual information between the optimal action and the environment (since optimal actions incur a distortion of 0):

$$\mathcal{L}_\beta(\tilde{A}_t|H_t) \leq \mathcal{L}_\beta(A_\star|H_t) = \mathbb{I}_t(\mathcal{E};A_\star).$$

We therefore have the following corollary.

Corollary 4. *For all $\beta > 0$ and $\tau = 0, 1, 2, \ldots$,*

$$\mathbb{E}\left[\sum_{t=\tau}^{\infty}\mathbb{I}_t(\tilde{A}_t;(A_t,O_{t+1}))\Big|H_\tau\right] \leq \mathbb{I}_\tau(\mathcal{E};A_\star).$$

We omit the proof of Corollary 1 as it follows directly by applying the preceding inequality to the following generalization that applies to any target action.

Corollary 5. *For all $\beta > 0$, target actions $\tilde{A}$, and $\tau = 0, 1, 2, \ldots$,*

$$\mathbb{E}_\tau\left[\sum_{t=\tau}^{\infty}\mathbb{I}_t(\tilde{A}_t;(A_t,O_{t+1}))\right] \leq \mathcal{L}_\beta(\tilde{A}|H_\tau).$$

Proof.

$$\begin{aligned}
\mathbb{E}_\tau\left[\sum_{t=\tau}^{\infty}\mathbb{I}_t(\tilde{A}_t;(A_t,O_{t+1}))\right] &\leq \mathbb{E}_\tau\left[\sum_{t=\tau}^{\infty}\mathcal{L}_\beta(\tilde{A}_t|H_t)-\mathbb{E}_t\left[\mathcal{L}_\beta(\tilde{A}_{t+1}|H_{t+1})\right]\right]\\
&=\sum_{t=\tau}^{\infty}\mathbb{E}_\tau\left[\mathcal{L}_\beta(\tilde{A}_t|H_t)\right]-\mathbb{E}_\tau\left[\mathbb{E}_t\left[\mathcal{L}_\beta(\tilde{A}_{t+1}|H_{t+1})\right]\right]\\
&=\mathbb{E}_\tau\left[\mathcal{L}_\beta(\tilde{A}_\tau|H_\tau)\right]+\sum_{t=\tau+1}^{\infty}\mathbb{E}_\tau\left[\mathcal{L}_\beta(\tilde{A}_t|H_t)\right]-\sum_{t=\tau}^{\infty}\mathbb{E}_\tau\left[\mathcal{L}_\beta(\tilde{A}_{t+1}|H_{t+1})\right]\\
&=\mathcal{L}_\beta(\tilde{A}_\tau|H_\tau)+\sum_{t=\tau+1}^{\infty}\mathbb{E}_\tau\left[\mathcal{L}_\beta(\tilde{A}_t|H_t)\right]-\sum_{t=\tau+1}^{\infty}\mathbb{E}_\tau\left[\mathcal{L}_\beta(\tilde{A}_t|H_t)\right]\\
&=\mathcal{L}_\beta(\tilde{A}_\tau|H_\tau)\leq\mathcal{L}_\beta(\tilde{A}|H_\tau)
\end{aligned}$$

where the steps follow as Lemma 2, linearity of expectation, the tower property, and the fact that $\tilde{A}_\tau$ is the minimizer of $\mathcal{L}_\beta(\cdot|H_\tau)$, by definition.

□

Let Γ be a constant such that

$$\Gamma \geq \frac{\mathbb{E}_t[\overline{r}(\tilde{A})-\overline{r}(A)]^2}{\mathbb{I}_t(\tilde{A};A,O)},$$

for all histories H_t, target actions $\tilde{A}$, if the executed action A is an independent sample drawn from the marginal distribution of $\tilde{A}$, and O is the resulting observation. Thus, Γ is an upper bound on the information ratio (Russo & Van Roy, 2014; 2016; 2018a) for which existing information-theoretic analyses of worst-case finite-arm bandits and linear bandits provide explicit values of Γ that satisfy this condition.

We can now establish our main results. We omit the proof of Theorem 1 as it is a special case of our subsequent result.

Theorem 4. *If $\beta = \frac{1-\gamma^2}{(1-\gamma)^2\Gamma}$ then, for all $\tau = 0,1,2,\ldots$,*

$$\mathbb{E}_\tau\left[\sum_{t=\tau}^{\infty}\gamma^{t-\tau}(\overline{r}(A_\star)-\overline{r}(A_t))\right] \leq 2\sqrt{\frac{\Gamma\mathbb{I}_\tau(\mathcal{E};A_\star)}{1-\gamma^2}}.$$

In a complex environment with many actions, $\mathbb{I}(\mathcal{E};A_\star)$ can be extremely large, rendering the above result somewhat vacuous under such circumstances. The next result offers a generalization, establishing a regret bound that can depend on the information content of any target action, including of course those that are much simpler than $A_\star$.

Theorem 5. *If $\beta = \frac{1-\gamma^2}{(1-\gamma)^2\Gamma}$ then, for all target actions $\tilde{A}$ and $\tau = 0,1,2,\ldots$,*

$$\mathbb{E}_\tau\left[\sum_{t=\tau}^{\infty}\gamma^{t-\tau}(\overline{r}(A_\star)-\overline{r}(A_t))\right] \leq 2\sqrt{\frac{\Gamma\mathbb{I}(\mathcal{E};\tilde{A}|H_\tau=H_\tau)}{1-\gamma^2}}+\frac{2\epsilon}{1-\gamma},$$

where $\epsilon = \sqrt{\mathbb{E}[(\overline{r}(A_\star)-\overline{r}(\tilde{A})^2|H_\tau]}$.

Proof.

From the inequalities satisfied by Γ, the Cauchy-Schwartz inequality, and Corollary 2, we have

$$\begin{aligned}\mathbb{E}_\tau\left[\sum_{t=\tau}^{\infty}\gamma^{t-\tau}(\overline{r}(\tilde{A}_t)-\overline{r}(A_t))\right] \leq& \mathbb{E}_\tau\left[\sum_{t=\tau}^{\infty}\gamma^{t-\tau}\sqrt{\Gamma\mathbb{I}_\tau(\tilde{A}_t;(A_t,O_{t+1}))}\right]\\ \leq& \sum_{t=\tau}^{\infty}\sqrt{\gamma^{2(t-\tau)}\Gamma}\sqrt{\sum_{t=\tau}^{\infty}\mathbb{E}_\tau\left[\mathbb{I}_\tau(\tilde{A}_t;(A_t,O_{t+1}))\right]}\\ \leq& \sqrt{\Gamma\mathcal{L}_\beta(\tilde{A}|H_\tau)\sum_{t=0}^{\infty}\gamma^{2t}}\\ =& \sqrt{\frac{\Gamma\mathcal{L}_\beta(\tilde{A}|H_\tau)}{1-\gamma^2}}.\end{aligned}$$

Since $\mathcal{L}_\beta(\tilde{A}_t|H_t)\geq 0$,

$$\sqrt{\mathbb{E}_t\left[(\overline{r}(A_\star)-\overline{r}(\tilde{A}_t))^2\right]}\leq(1-\gamma)\sqrt{\frac{\Gamma\mathcal{L}_\beta(\tilde{A}_t|H_t)}{1-\gamma^2}}.$$

Further, applying Jensen's inequality to the left-hand side and using the fact that $\tilde{A}_t$ minimizes $\mathcal{L}_\beta(\tilde{A}_t|H_t)$ on the right-hand side,

$$\mathbb{E}_t\left[\overline{r}(A_\star)-\overline{r}(\tilde{A}_t)\right]\leq(1-\gamma)\sqrt{\frac{\Gamma\mathcal{L}_\beta(\tilde{A}|H_t)}{1-\gamma^2}}.$$

Lemma 1 implies that

$$\mathbb{E}_\tau[\mathcal{L}_\beta(\tilde{A}|H_t)]\leq\mathcal{L}_\beta(\tilde{A}|H_\tau),$$

for all $t\geq\tau$, and therefore, by Jensen's inequality,

$$\mathbb{E}_\tau\left[\overline{r}(A_\star)-\overline{r}(\tilde{A}_t)\right]\leq(1-\gamma)\mathbb{E}_\tau\left[\sqrt{\frac{\Gamma\mathcal{L}_\beta(\tilde{A}|H_t)}{1-\gamma^2}}\right]\leq(1-\gamma)\sqrt{\frac{\Gamma\mathbb{E}_\tau\left[\mathcal{L}_\beta(\tilde{A}|H_t)\right]}{1-\gamma^2}}\leq(1-\gamma)\sqrt{\frac{\Gamma\mathcal{L}_\beta(\tilde{A}|H_\tau)}{1-\gamma^2}}.$$

It follows that

$$\begin{aligned}\mathbb{E}_\tau\left[\sum_{t=\tau}^{\infty}\gamma^{t-\tau}(\overline{r}(A_\star)-\overline{r}(\tilde{A}_t))\right] \leq& \sqrt{\frac{\Gamma\mathcal{L}_\beta(\tilde{A}|H_\tau)}{1-\gamma^2}}\\ \leq& \sqrt{\frac{\Gamma(\mathbb{I}_\tau(\mathcal{E};\tilde{A})+\beta\epsilon^2)}{1-\gamma^2}}\\ \leq& \sqrt{\frac{\Gamma\mathbb{I}_\tau(\mathcal{E};\tilde{A})}{1-\gamma^2}}+\frac{\epsilon}{1-\gamma}.\end{aligned}$$

Applying these same steps, we complete the above bound as

$$\mathbb{E}_\tau\left[\sum_{t=\tau}^{\infty}\gamma^{t-\tau}(\overline{r}(\tilde{A}_t)-\overline{r}(A_t))\right]\leq\sqrt{\frac{\Gamma\mathbb{I}_\tau(\mathcal{E};\tilde{A})}{1-\gamma^2}}+\frac{\epsilon}{1-\gamma}.$$

Putting everything together, we have

$$\begin{aligned}\mathbb{E}_\tau\left[\sum_{t=\tau}^{\infty}\gamma^{t-\tau}(\overline{r}(A_\star)-\overline{r}(A_t))\right] &= \mathbb{E}_\tau\left[\sum_{t=\tau}^{\infty}\gamma^{t-\tau}(\overline{r}(A_\star)-\overline{r}(\tilde{A}_t)+\overline{r}(\tilde{A}_t)-\overline{r}(A_t))\right]\\ &= \mathbb{E}_\tau\left[\sum_{t=\tau}^{\infty}\gamma^{t-\tau}(\overline{r}(A_\star)-\overline{r}(\tilde{A}_t))\right]+\mathbb{E}_\tau\left[\sum_{t=\tau}^{\infty}\gamma^{t-\tau}(\overline{r}(\tilde{A}_t)-\overline{r}(A_t))\right]\\ &\leq 2\sqrt{\frac{\Gamma\mathbb{I}_\tau(\mathcal{E};\tilde{A})}{1-\gamma^2}}+\frac{2\epsilon}{1-\gamma}.\end{aligned}$$

□

D. Undiscounted Regret Analysis

In this section, we derive a variant of Theorem 2 where performance shortfall is measured by the expected cumulative regret across a finite horizon. Consider a fixed time horizon T and observe the analogous result to Corollary 2:

Corollary 6. *For all $\beta > 0$, target actions $\tilde{A}$, and $\tau = 0, 1, 2, \ldots$,*

$$\mathbb{E}_\tau\left[\sum_{t=\tau}^{T+\tau}\mathbb{I}_t(\tilde{A}_t;(A_t,O_{t+1}))\right] \leq \mathcal{L}_\beta(\tilde{A}|H_\tau).$$

Proof.

$$\begin{aligned}\mathbb{E}_\tau\left[\sum_{t=\tau}^{T+\tau}\mathbb{I}_t(\tilde{A}_t;(A_t,O_{t+1}))\right] &\leq \mathbb{E}_\tau\left[\sum_{t=\tau}^{T+\tau}\mathcal{L}_\beta(\tilde{A}_t|H_t)-\mathbb{E}_t\left[\mathcal{L}_\beta(\tilde{A}_{t+1}|H_{t+1})\right]\right]\\ &= \sum_{t=\tau}^{T+\tau}\mathbb{E}_\tau\left[\mathcal{L}_\beta(\tilde{A}_t|H_t)\right]-\mathbb{E}_\tau\left[\mathbb{E}_t\left[\mathcal{L}_\beta(\tilde{A}_{t+1}|H_{t+1})\right]\right]\\ &= \mathbb{E}_\tau\left[\mathcal{L}_\beta(\tilde{A}_\tau|H_\tau)\right]+\sum_{t=\tau+1}^{T+\tau}\mathbb{E}_\tau\left[\mathcal{L}_\beta(\tilde{A}_t|H_t)\right]-\sum_{t=\tau}^{T+\tau}\mathbb{E}_\tau\left[\mathcal{L}_\beta(\tilde{A}_{t+1}|H_{t+1})\right]\\ &= \mathcal{L}_\beta(\tilde{A}_\tau|H_\tau)+\sum_{t=\tau+1}^{T+\tau}\mathbb{E}_\tau\left[\mathcal{L}_\beta(\tilde{A}_t|H_t)\right]-\sum_{t=\tau+1}^{T+\tau+1}\mathbb{E}_\tau\left[\mathcal{L}_\beta(\tilde{A}_t|H_t)\right]\\ &= \mathcal{L}_\beta(\tilde{A}_\tau|H_\tau)-\mathbb{E}_\tau\left[\mathcal{L}_\beta(\tilde{A}_{T+\tau+1}|H_{T+\tau+1})\right]\\ &\leq \mathcal{L}_\beta(\tilde{A}_\tau|H_\tau)\leq\mathcal{L}_\beta(\tilde{A}|H_\tau)\end{aligned}$$

where the steps follow as Lemma 2, linearity of expectation, the tower property, the non-negativity of $\mathcal{L}_\beta(\tilde{A}_t|H_t)\geq 0$, and the fact that $\tilde{A}_\tau$ is the minimizer of $\mathcal{L}_\beta(\cdot|H_\tau)$, by definition.

□

With Corollary 3, we may introduce the undiscounted analog to Theorem 2:

Theorem 6. *If $\beta = \frac{T}{\Gamma}$ then, for all target actions $\tilde{A}$ and $\tau = 0, 1, 2, \ldots$,*

$$\mathbb{E}_\tau\left[\sum_{t=\tau}^{T+\tau}\overline{r}(A_\star)-\overline{r}(A_t)\right] \leq 2\sqrt{\Gamma T\mathbb{I}_\tau(\mathcal{E};\tilde{A})}+2T\epsilon,$$

where $\epsilon = \sqrt{\mathbb{E}[(\overline{r}(A_\star) - \overline{r}(\tilde{A})^2 | H_\tau]}$.

Proof.

From the inequalities satisfied by Γ, the Cauchy-Schwartz inequality, and Corollary 3, we have

$$\begin{aligned}\mathbb{E}_\tau\left[\sum_{t=\tau}^{T+\tau} \overline{r}(\tilde{A}_t) - \overline{r}(A_t)\right] \leq & \sqrt{\Gamma}\mathbb{E}_\tau\left[\sum_{t=\tau}^{T+\tau}\sqrt{\mathbb{I}_\tau(\tilde{A}_t; (A_t, O_{t+1}))}\right] \\ \leq & \sqrt{\Gamma T \sum_{t=\tau}^{T+\tau}\mathbb{E}_\tau\left[\mathbb{I}_\tau(\tilde{A}_t; (A_t, O_{t+1}))\right]} \\ \leq & \sqrt{\Gamma T \mathcal{L}_\beta(\tilde{A}|H_\tau)}\end{aligned}$$

Since $\mathcal{L}_\beta(\tilde{A}_t|H_t) \geq 0$,

$$\sqrt{\mathbb{E}_t\left[(\overline{r}(A_\star) - \overline{r}(\tilde{A}_t))^2\right]} \leq T^{-1}\sqrt{\Gamma T \mathcal{L}_\beta(\tilde{A}_t|H_t)}.$$

Further, applying Jensen's inequality to the left-hand side and using the fact that $\tilde{A}_t$ minimizes $\mathcal{L}_\beta(\tilde{A}_t|H_t)$ on the right-hand side,

$$\mathbb{E}_t\left[\overline{r}(A_\star) - \overline{r}(\tilde{A}_t)\right] \leq T^{-1}\sqrt{\Gamma T \mathcal{L}_\beta(\tilde{A}|H_t)}.$$

Lemma 1 implies that

$$\mathbb{E}_\tau[\mathcal{L}_\beta(\tilde{A}|H_t)] \leq \mathcal{L}_\beta(\tilde{A}|H_\tau),$$

for all $t \geq \tau$, and therefore, by Jensen's inequality,

$$\mathbb{E}_\tau\left[\overline{r}(A_\star) - \overline{r}(\tilde{A}_t)\right] \leq T^{-1}\mathbb{E}_\tau\left[\sqrt{\Gamma T \mathcal{L}_\beta(\tilde{A}|H_t)}\right] \leq T^{-1}\sqrt{\Gamma T \mathbb{E}_\tau\left[\mathcal{L}_\beta(\tilde{A}|H_t)\right]} \leq T^{-1}\sqrt{\Gamma T \mathcal{L}_\beta(\tilde{A}|H_\tau)}.$$

It follows that

$$\begin{aligned}\mathbb{E}_\tau\left[\sum_{t=\tau}^{T+\tau} \overline{r}(A_\star) - \overline{r}(\tilde{A}_t)\right] \leq & \sqrt{\Gamma T \mathcal{L}_\beta(\tilde{A}|H_\tau)} \\ \leq & \sqrt{\Gamma T(\mathbb{I}_\tau(\mathcal{E}; \tilde{A}) + \beta\epsilon^2)} \\ \leq & \sqrt{\Gamma T \mathbb{I}_\tau(\mathcal{E}; \tilde{A})} + T\epsilon.\end{aligned}$$

Applying these same steps, we complete the above bound as

$$\mathbb{E}_\tau\left[\sum_{t=\tau}^{T+\tau} \overline{r}(\tilde{A}_t) - \overline{r}(A_t)\right] \leq \sqrt{\Gamma T \mathbb{I}_\tau(\mathcal{E}; \tilde{A})} + T\epsilon.$$

Putting everything together, we have

$$\begin{aligned}\mathbb{E}_\tau\left[\sum_{t=\tau}^{T+\tau}\overline{r}(A_\star)-\overline{r}(A_t)\right] &= \mathbb{E}_\tau\left[\sum_{t=\tau}^{T+\tau}\overline{r}(A_\star)-\overline{r}(\tilde{A}_t)+\overline{r}(\tilde{A}_t)-\overline{r}(A_t)\right]\\ &= \mathbb{E}_\tau\left[\sum_{t=\tau}^{T+\tau}\overline{r}(A_\star)-\overline{r}(\tilde{A}_t)\right]+\mathbb{E}_\tau\left[\sum_{t=\tau}^{T+\tau}\overline{r}(\tilde{A}_t)-\overline{r}(A_t)\right]\\ &\leq 2\sqrt{\Gamma T\mathbb{I}_\tau(\mathcal{E};\tilde{A})}+2T\epsilon.\end{aligned}$$

□