Deciding What to Learn: A Rate-Distortion Approach

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

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 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**
  $e_1, \ldots, e_Z \sim \mathbb{P}(\mathcal{E} \in \cdot | H_t)$ {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
  $\hat{z} \sim$ 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. 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 1. *For all $\beta > 0$, target actions $\tilde{A}$, and $t = 0, 1, 2, \ldots$,*

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

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

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

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

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

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

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

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

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

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

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

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

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

□

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.

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.

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

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.

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.

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.

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.

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., 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.

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.