Exploration Unbound

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

Wanqiao Xu*
wanqiaoxu@stanford.edu
Department of Management Science & Engineering
Stanford University

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

Abstract

A sequential decision-making agent balances between exploring to gain new knowledge about an environment and exploiting current knowledge to maximize immediate reward. For environments studied in the traditional literature, optimal decisions gravitate over time toward exploitation as the agent accumulates sufficient knowledge and the benefits of further exploration vanish. What if, however, the environment offers an unlimited amount of useful knowledge and there is large benefit to further exploration no matter how much the agent has learned? We offer a simple, quintessential example of such a complex environment. In this environment, rewards are unbounded and an agent can always increase the rate at which rewards accumulate by exploring to learn more. Consequently, an optimal agent forever maintains a propensity to explore.

1 Introduction

Consider an unscrupulous geometry teacher and a persistent student eager to learn the digits of the mathematical constant $\pi \approx 3.1415926535$. On each day, the student approaches the teacher and may provide any arbitrarily-long sequence of digits. For any k-digit sequence they provide, the teacher simply checks to see if these exactly match the first k digits of π or not. If there is such an exact k-digit match, the teacher awards the student $r(k)$ dollars; otherwise, for anything less than a perfect match, the teacher charges the student $c(k)$ dollars. The teacher sets $r(k)$ and $c(k)$ to be increasing functions of the number of digits and, therefore, guessing longer digit sequences offers more potential upside but also increased risk from erroneous guesses.

Suppose rewards are bounded and the student wishes to maximize expected discounted return. Then, if on any day, it is optimal to exploit current knowledge by choosing k digits of π the student has learned thus far, it will be optimal to choose those same k digits of π on every subsequent day. This is because the theory of sequential decision-making problems establishes that, when rewards are bounded, there is an optimal policy that maps each possible state of knowledge to a single action. In our example, the student's state of knowledge remains unchanged by the exploitative choice. This leads to indefinite daily repetition of the same knowledge state and same optimal action.

On the other hand, if rewards are unbounded, there may always be substantial value in exploring to learn more. In this circumstance, the aforementioned behavior, where on some date the student exploits and continues to do so indefinitely, can fall far short of optimal. Instead, an optimal policy may have to randomize

*Equal contribution

between exploration and exploitation, without ever tapering the intensity of the former. This randomization must strike a delicate balance as exploration can be costly and exploitation is required to recoup those costs.

The preceding example represents but one instance of a much broader class of complex environments, for which policies that taper exploration forgo value. In this paper, we will introduce and analyze a representative instance to offer insights into optimal behavior. Beyond this paper, these complex environments encapsulate an increasingly ubiquitous exploration challenge that modern decision-making systems confront at a much grander scale, comparable to that of the World Wide Web (Shi et al., 2017; Toyama et al., 2021; Stiennon et al., 2020; Ouyang et al., 2022; Yao et al., 2022; Dwaracherla et al., 2024). Consider a fixed user prompt submitted to a large language model (LLM). To any current response produced by the LLM, the versatility of natural language affords opportunities to explore a vast space of alternative responses and identify more valuable ones. Clearly, pure exploration would prove very costly because users would largely receive less-desirable, experimental responses. At the same time, our results suggest that tapering exploration would be sub-optimal as repeating any current response could be improved with further exploration and learning.

The paper proceeds as follows: we formulate a general class of bandit-learning problems in Section 2, introduce in Section 3 our complex bandit environment which admits perpetual improvement that demands eternal exploration, and conclude with discussion as well as future outlook in Section 4. For ease of readability, all technical proofs have been relegated to the appendix.

2 Problem Formulation

We consider a **bandit environment** (Lattimore & Szepesvári, 2020) with a (possibly infinite) action set $\mathcal{A}$ and stochastic reward process $\{R_t\}_{t\in\mathbb{Z}_{>0}}$. Here, $\{R_t\}_{t\in\mathbb{Z}_{>0}}$ is an exchangeable sequence of random vectors, each with one component per action; each vector R_t assigns a scalar reward $R_{t,a}$ to each action a. At each time $t \in \mathbb{Z}_{\geq 0}$, an action A_t is executed and generates a scalar reward R_{t+1,A_t}. If there exists an $R \in \mathbb{R}$ such that $\mathbb{E}[R_{t,a}] < R$ for each time t and action a, we say that rewards are bounded. Otherwise, we say that rewards are unbounded.

By de Finetti's Theorem (de Finetti, 1937), exchangeability immediately implies the existence of a random variable θ, representing the unknown bandit environment, conditioned upon which the sequence $\{R_t\}_{t\in\mathbb{Z}_{>0}}$ is iid. A common example may clarify our formulation; consider the Bernoulli bandit with unknown success probabilities $\theta \in [0,1]^{\mathcal{A}}$ and observe that, conditioned on θ, the rewards $\{R_t\}_{t\in\mathbb{Z}_{>0}}$ are iid. Hence, $\{R_t\}_{t\in\mathbb{Z}_{>0}}$ is exchangeable. The initial distribution of θ expresses prior beliefs. At time t, there is a history $H_t = (A_0, R_{1,A_0}, A_2, \ldots, A_{t-1}, R_{t,A_{t-1}})$ of actions and realized rewards, and posterior beliefs are expressed by the distribution of θ conditioned on H_t.

A (stationary) **policy** π is a mapping from histories to action probabilities. In particular, for any history h and action a, $\pi(a \mid h)$ is the probability assigned to executing action a upon observing history h. Hence, if an agent applies a policy π, each action A_t is sampled from $\pi(\cdot \mid H_t)$. To frame a notion of optimality across policies, we first define the expected finite-horizon discounted return of a policy π:

$$V_\pi^{\gamma,T} = \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t R_{t+1,A_t}\right].$$

Here, $\gamma \in [0,1]$ is a discount factor and $T \in \mathbb{Z}_{>0}$ is a time horizon. Note that this expectation integrates over uncertainty not only in θ but also in rewards conditioned on θ.

A policy π is said to be **discounted-overtaking optimal** if, for all policies π',

$$\liminf_{T\to\infty} (V_\pi^{\gamma,T} - V_{\pi'}^{\gamma,T}) \geq 0.$$

If $\gamma = 1$, this objective corresponds to the standard notion of overtaking optimality (see Section 5.4.2 of (Puterman, 1994)). If rewards are bounded, a policy is discounted-overtaking optimal if and only if it maximizes

the expected discounted return $V_\pi^\gamma \equiv V_\pi^{\gamma,\infty}$, which is finite. On the other hand, if rewards are unbounded, V_π^γ can be infinite for many policies. The more general notion of discounted-overtaking optimality affords more nuanced comparisons among policies that attain infinite discounted return. Intuitively, if there is a set of policies that attain infinite expected discounted return, only those that more quickly accumulate discounted rewards can be discounted-overtaking optimal.

While discounted-overtaking optimality offers an unambiguous criteria for a best policy, a discounted-overtaking optimal policy may not always exist. To facilitate comparing policies without knowledge of a discounted-overtaking optimal policy, we define the expected finite-horizon regret of a policy π relative to a reference policy μ:

$$\text{Regret}_{\pi,\mu}^T = \mathbb{E}_\pi\left[\sum_{t=0}^{T-1} R_{t+1,A_t}\right] - \mathbb{E}_\mu\left[\sum_{t=0}^{T-1} R_{t+1,A_t}\right].$$

3 Necessity of Randomized Exploration

In this section, we define and analyze a complex environment that will be our main object of study for the remainder of the paper.

Example 1. *We define a bandit with an action set* $\mathcal{A} = \bigcup_{k=0}^{\infty} \mathbb{Z}_{>0}^k$. *Hence, each action* $a \in \mathcal{A}$ *is a positive-integer-valued tuple* $(a_1, \ldots, a_k)$ *of some arbitrary length* $k \geq 0$. *The stochastic process of reward vectors* $\{R_{t+1}\}_{t\in\mathbb{Z}_{\geq 0}}$ *is parameterized by fixed, known scalars* $\alpha > 1$ *and* $\tau \in (1, \frac{\gamma(\alpha-1)}{2(1-\gamma)})$, *and an unknown positive-integer sequence* $a^* = (a_1^*, a_2^*, a_3^*, \ldots)$ *such that, for each* $a \in \mathbb{Z}_{>0}^k$,

$$R_{t+1,a} = \begin{cases} \alpha^k & \text{if } a = a_{1:k}^* \\ -\frac{\alpha+1}{\tau-1}\alpha^{k-1} & \text{otherwise.} \end{cases}$$

Note that the action set includes the length 0 *vector, which we will denote by* $\emptyset \equiv a_{1:0}^*$ *and yields reward* $R_{t+1,\emptyset} = 1$.

The reward process $\{R_{t+1}\}_{t\in\mathbb{Z}_{\geq 0}}$ *is exchangeable since* $R_1 = R_2 = \cdots$. *As* $\{R_{t+1}\}_{t\in\mathbb{Z}_{\geq 0}}$ *is determined by* a^*, *we define its prior distribution in terms of a prior distribution of* a^*. *In particular, for each* $k \in \mathbb{Z}_{>0}$, *let the distribution* $p_k(\cdot|a_{1:k-1}^*)$ *of* a_k^* *conditioned on* $a_{1:k-1}^*$ *be geometric with mean* τ. *Then, let the prior probability assigned to* $a_{1:K}^*$ *be* $\prod_{k=1}^K p_k(a_k^*|a_{1:k-1}^*)$.

We say that an agent is *exploiting* at timestep t if its chosen action A_t is a previously selected action known to achieve highest reward, given history H_t. Otherwise, we say that an agent is *exploring*. Example 1 offers a representative instance of how the presence of infinitely-many actions and unbounded rewards demands an infinite amount of exploration in order to synthesize optimal behavior. An initial thought may be to purely explore in perpetuity and continually uncover higher tiers of reward. Unfortunately, the cost structure associated with incorrect actions (that do not match the goal sequence a^*) is calibrated such that this policy is provably not optimal.

Theorem 1. *In Example 1, an agent that always explores is never discounted-overtaking optimal.*

Upon further reflection, it is perhaps not terribly surprising that a strategy of pure exploration is sub-optimal in Example 1, as is often the case for many sequential decision-making problems studied in the literature. However, unlike these latter commonly-studied problems, we may also obtain an analogous theoretical result establishing that any policy which ceases exploration in any time period is also provably not-optimal. This represents a more substantial departure from the traditional literature, where the presence of bounded rewards (even with infinitely-many actions) guarantees the existence of an optimal policy which eventually exploits with probability 1; we defer a review of this prior work to Section 4.

Theorem 2. *In Example 1, an agent that stops exploring is never discounted-overtaking optimal.*

Naturally, if neither exploring nor exploiting with probability 1 yields an optimal strategy, an agent's only recourse is to randomize. Our next theoretical result formalizes this intuition.

Theorem 3. *In Example 1, for all $T < \infty$, there exists a policy π_T^* that is exploiting with non-zero probability $p_T^* \in (0, 1)$ and exploring with non-zero probability $1 - p_T^*$ at each timestep, such that, for all policy π,*

$$\text{Regret}_{\pi, \pi_T^*}^T \leq 0.$$

The precise probability with which an agent optimally balances its preference for exploration versus exploitation in each time period, p_T^*, is sensitive to the overall time horizon over which it aims to maximize reward. We offer the following conjecture to clarify how this probability behaves asymptotically as the agent engages with the full brunt of infinite exploration in a complex environment.

Conjecture 1. *In Example 1, as $T \to \infty$, $p_T^* \to \frac{\alpha+1}{\alpha+\tau}$.*

Conjecture 1 posits that, as T increases, the optimal policy randomizes between exploitation and exploration with exploiting probability p_T^* approaching a non-zero threshold. Intuitively, even as T approaches infinity, the agent should never taper off its exploration probability. To see why this conjecture holds, we provide a potential roadmap in Appendix B. It is worth noting that there is a concrete instantiation of Example 1 for which we can offer a more precise, elegant characterization of the result in Conjecture 1. Specifically, when $\alpha = 2$; $\tau = 4$; and $\gamma \geq 0.85$, $p_T^* \to \frac{\alpha+1}{\alpha+\tau} = 0.5$ and an optimal agent must perseverate with equiprobable exploration and exploitation for all time.

4 Discussion

Two key facets of the complex environment studied in this work are the presence of infinitely-many actions and unbounded rewards. In this section, we begin with an overview of prior work, which largely focuses on the former condition in the absence of the latter, as well as a small handful of papers from outside the machine-learning literature which consider both conditions together. We conclude with a discussion of how one might begin to approach the design of practical agents for such complex environments and offer a simple computational experiment to corroborate our proposal.

4.1 Prior Work

Multi-armed bandit problems with infinitely-many actions available to the agent have been a topic of interest in the literature for decades. The earliest work by Mallows & Robbins (1964) demonstrates that reward distributions with bounded moments allows for the characterization of an optimal, non-stationary policy maximizing average reward. A similar non-stationary strategy is analyzed for regret minimization by Yakowitz & Lowe (1991), who also rely on bounded higher moments of the reward distributions at each arm. Banks & Sundaram (1992) extend the classic machinery of Gittins' indices (Gittins, 1974; 1979; Gittins & Jones, 1979) to bandit problems with a countably-infinite number of independent arms, each of which is assumed to have an associated distribution that yields uniformly bounded expected rewards. Identical structural assumptions are made explicitly, by Lai & Yakowitz (1995); Wang et al. (2008); Carpentier & Valko (2015); Aziz et al. (2018); Kalvit & Zeevi (2020); De Heide et al. (2021); Lai et al. (2022); Wang et al. (2022); Russo & Van Roy (2022), as well as implicitly, by Herschkorn et al. (1996); Berry et al. (1997); Chen & Lin (2004; 2005); Hung (2012); Bonald & Proutiere (2013); Gong & Sellke (2023), where the latter all study the infinite-armed Bernoulli bandit. A related setting to the infinite-armed Bernoulli bandit is the so-called many-armed bandit setting where the number of actions is finite but considered large relative to the problem horizon (Teytaud et al., 2007; Zhu & Nowak, 2020). Notably, all of the aforementioned papers do not entertain unbounded rewards alongside an infinitely-large action space.

Agrawal (1995) studies bandit problems whose action space forms a subset of the real line $\mathbb{R}$ under the assumptions of sub-Gaussian rewards and uniformly locally-Lipschitz continuous mean rewards; crucially,

the latter assumption allows for a carefully-constructed collection of actions which adequately cover the action space to obtain an approximation of the expected reward function suitable for identifying near-optimal actions. Improvements of these results for the one-dimensional case as well as extensions to vector-valued action spaces of arbitrary dimension are studied by Kleinberg (2004); Auer et al. (2007); Kleinberg et al. (2008); Cope (2009); Bubeck et al. (2011). Aside from the boundedness of expected rewards implied by the sub-Gaussianity assumptions in some of these works, the more critical distinction is (again) in the use of uniform local-Lipschitz continuity, which serves as the key inductive bias to facilitate effective learning when there are infinitely-many actions. In contrast, the complex environment studied in this work presents a latent curricular structure that enables efficient learning despite the unboundedness of rewards.

Unbounded rewards represent an important piece of the complex environment studied in this work, ensuring ample opportunity to dramatically improve the value of any current best decision known to an agent. Decision-making problems with unbounded rewards have long been a subject of study in philosophy (Arntzenius et al., 2004; Goodsell, 2023), though, to the best of the authors' knowledge, seem to not be a topic of study in the traditional bandit literature. A typical discussion point of such external papers to RL are the paradoxes that emerge among potentially optimal behaviors defined to maximize expected utility. Meanwhile, this work focuses on alternative performance criteria for which optimal behavior can be clearly defined, though may not be guaranteed to exist.

4.2 Towards Practical Agent Design

While we offer theoretical results which underscore the importance of randomization for a discounted-overtaking optimal policy in a complex environment, it is perhaps not immediately apparent how to go about the practical implementation of a computational agent that could learn or approximate this optimal behavior from interaction data. We anticipate that the concept of a *learning target* (Lu et al., 2023) is essential to the design and practical implementation of such an agent. When, at any given time, optimal behavior is so complicated that it requires too much information to learn, it behooves the agent to have a mechanism for prioritizing some other modest corpus of information that, while capable of facilitating behavioral improvement, is itself insufficient to enable near-optimal performance; broadly speaking, a learning target is such a mechanism. As the agent gains competency through its prolonged interaction with the environment, one might envision that this learning target could adapt in kind to reflect updated knowledge and reorient exploration towards new, feasible discoveries.

A recent line of work (Arumugam & Van Roy, 2021a;b; 2022) has studied the design, analysis, and implementation of decision-making agents endowed with the ability to compute such learning targets and autonomously decide what to learn. We strongly suspect that deciding what to learn and striking a desired trade-off between information requirements and performance is a critical capability for an agent coping with complex environments where eternal exploration is the only path to optimal behavior. As a preliminary empirical illustration of our hypothesis, we offer a computational result based on the π-guessing example of Section 1. For the sake of computational feasibility, we prune down the original π-guessing game and reduce the task to learning the first two digits of π, yielding a finite action set $\mathcal{A} = \{0, \ldots, 99\}$. We refer interested readers to Section A of the appendix for more granular details of the computational experiment.

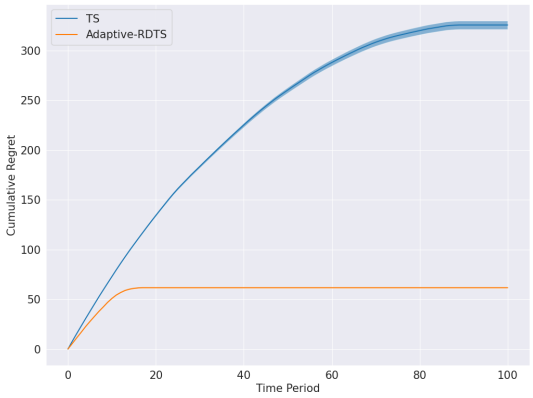


Figure 1: Cumulative regret curve comparing Thompson Sampling and Rate-Distortion Thompson Sampling agents for learning the first two digits of π.

A classic approach like Thompson Sampling (Thompson, 1933; Russo et al., 2018) (TS) invests exploratory effort only in those actions with non-zero probability of being optimal, thereby forgoing the opportunity to leverage the inherent curricular structure of the environment outline in Example 1. Meanwhile, an adaptive variant of the Rate-Distortion Thompson Sampling (RDTS) agent introduced by Arumugam et al. (2024) is able to compute a sequence of learning targets for scaffolding exploration around the identification of successive digits. Consequently, in the worst case, TS demands at most 90 time periods to identify the first two digits of π whereas RDTS requires at most 20. Importantly, while the associated cumulative regret curves shown in Figure 1 pertains to an environment lacking the key features of infinitely-many actions and unbounded rewards, it does illustrate how a learning target can be operationalized to modulate exploration in a manner resemblant of what would be needed for efficient learning in Example 1 and complex environments more broadly.

5 Conclusion

In this paper, we have engaged with a broadened treatment of the exploration challenge in sequential decision-making problems, departing from the traditional setting where optimal behavior tapers exploration over time. Instead, as an agent always has substantial opportunity for improvement, optimal behavior demands eternal exploration. A single, quintessential environment studied throughout this paper exemplifies this nuanced exploration problem through the combination of an infinitely-large action space and unbounded rewards. Our theoretical analysis clarifies the manner in which strategies of pure exploration and pure exploitation fall short of optimal performance, thereby necessitating the use of a randomization to preserve an agent's propensity to explore the world.

We posit that our work offers a simple microcosm for the formal study of an exploration problem that manifests at a much grander scale across several real-world applications (Shi et al., 2017; Toyama et al., 2021; Stiennon et al., 2020; Ouyang et al., 2022; Yao et al., 2022; Dwaracherla et al., 2024). Large language models (LLMs), across each individual user prompt, must contend with exploring the entire space of natural language responses to identify the best one (Stiennon et al., 2020; Ouyang et al., 2022; Dwaracherla et al., 2024). To impose an upper bound on rewards in LLMs (usually corresponding to human preference or utility scores) is to imply a known upper limit on human satisfaction across the entire space of natural language prompts. Instead, one might consider the possibility that such an upper limit on response utility does not exist and so, to any current best-known response, there always remains an opportunity to refine and improve. Our analysis suggests a natural exploration strategy for such settings where untested candidate responses are emitted in careful proportion to currently preferred responses over time. Similar scenarios are perhaps likely to emerge in environments of comparable scale, including those for learning desired behaviors on mobile devices (Toyama et al., 2021) or the Internet itself (Shi et al., 2017; Yao et al., 2022). Additionally, extending beyond the considerations of any single task, the lifelong or continual reinforcement learning setting has been a recent source of great interest to the community (Ring, 1994; Khetarpal et al., 2022; Kumar et al., 2023; Abel et al., 2024). The inherent non-stationarity of the environment in continual reinforcement learning also beckons for a strategy of prolonged and enduring exploration, for which future theoretical analyses and agent-design principles may take inspiration from this work.

References

David Abel, André Barreto, Benjamin Van Roy, Doina Precup, Hado P van Hasselt, and Satinder Singh. A Definition of Continual Reinforcement Learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Rajeev Agrawal. The Continuum-Armed Bandit Problem. *SIAM Journal on Control and Optimization*, 33 (6):1926–1951, 1995.

Suguru Arimoto. An Algorithm for Computing the Capacity of Arbitrary Discrete Memoryless Channels. *IEEE Transactions on Information Theory*, 18(1):14–20, 1972.

Frank Arntzenius, Adam Elga, and John Hawthorne. Bayesianism, Infinite Decisions, and Binding. *Mind*, 113(450):251–283, 2004.

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

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

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

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

Peter Auer, Ronald Ortner, and Csaba Szepesvári. Improved Rates for the Stochastic Continuum-Armed Bandit Problem. In *International Conference on Computational Learning Theory*, pp. 454–468. Springer, 2007.

Maryam Aziz, Jesse Anderton, Emilie Kaufmann, and Javed Aslam. Pure Exploration in Infinitely-Armed Bandit Models with Fixed-Confidence. In *Algorithmic Learning Theory*, pp. 3–24. PMLR, 2018.

Jeffrey S Banks and Rangarajan K Sundaram. Denumerable-Armed Bandits. *Econometrica: Journal of the Econometric Society*, pp. 1071–1096, 1992.

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

Donald A Berry, Robert W Chen, Alan Zame, David C Heath, and Larry A Shepp. Bandit Problems with Infinitely Many Arms. *The Annals of Statistics*, 25(5):2103–2116, 1997.

Richard Blahut. Computation of Channel Capacity and Rate-Distortion Functions. *IEEE Transactions on Information Theory*, 18(4):460–473, 1972.

Thomas Bonald and Alexandre Proutiere. Two-Target Algorithms for Infinite-Armed Bandits with Bernoulli Rewards. *Advances in Neural Information Processing Systems*, 26, 2013.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12(5), 2011.

Alexandra Carpentier and Michal Valko. Simple Regret for Infinitely Many Armed Bandits. In *International Conference on Machine Learning*, pp. 1133–1141. PMLR, 2015.

Kung-Yu Chen and Chien-Tai Lin. A Note on Strategies for Bandit Problems with Infinitely Many Arms. *Metrika*, 59:193–203, 2004.

Kung-Yu Chen and Chien-Tai Lin. A Note on Infinite-Armed Bernoulli Bandit Problems with Generalized Beta Prior Distributions. *Statistical Papers*, 46(1):129–140, 2005.

Mung Chiang and Stephen Boyd. Geometric Programming Duals of Channel Capacity and Rate Distortion. *IEEE Transactions on Information Theory*, 50(2):245–258, 2004.

Eric W Cope. Regret and Convergence Bounds for a Class of Continuum-Armed Bandit Problems. *IEEE Transactions on Automatic Control*, 54(6):1243–1253, 2009.

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

Imre Csiszár. On an Extremum Problem of Information Theory. *Studia Scientiarum Mathematicarum Hungarica*, 9, 1974.

Bruno de Finetti. Foresight: Its Logical Laws, its Subjective Sources. In *Breakthroughs in Statistics: Foundations and Basic Theory*, pp. 134–174. Springer, 1937.

Rianne De Heide, James Cheshire, Pierre Ménard, and Alexandra Carpentier. Bandits with Many Optimal Arms. *Advances in Neural Information Processing Systems*, 34:22457–22469, 2021.

Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded Modeling Language for Convex Optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Vikranth Dwaracherla, Seyed Mohammad Asghari, Botao Hao, and Benjamin Van Roy. Efficient Exploration for LLMs. *arXiv preprint arXiv:2402.00396*, 2024.

John Gittins. A Dynamic Allocation Index for the Sequential Design of Experiments. *Progress in Statistics*, pp. 241–266, 1974.

John Gittins. Bandit Processes and Dynamic Allocation Indices. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 41(2):148–164, 1979.

John C Gittins and David M Jones. A Dynamic Allocation Index for the Discounted Multiarmed Bandit Problem. *Biometrika*, 66(3):561–565, 1979.

Evelyn Xiao-Yue Gong and Mark Sellke. Asymptotically Optimal Quantile Pure Exploration for Infinite-Armed Bandits. *Advances in Neural Information Processing Systems*, 36, 2023.

Zachary Goodsell. Decision Theory Unbound. *Noûs*, 2023.

Stephen J Herschkorn, Erol Pekoez, and Sheldon M Ross. Policies Without Memory for the Infinite-Armed Bernoulli Bandit Under the Average-Reward Criterion. *Probability in the Engineering and Informational Sciences*, 10(1):21–28, 1996.

Ying-Chao Hung. Optimal Bayesian Strategies for the Infinite-Armed Bernoulli Bandit. *Journal of Statistical Planning and Inference*, 142(1):86–94, 2012.

Anand Kalvit and Assaf Zeevi. From Finite to Countable-Armed Bandits. *Advances in Neural Information Processing Systems*, 33:8259–8269, 2020.

Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards Continual Reinforcement Learning: A Review and Perspectives. *Journal of Artificial Intelligence Research*, 75:1401–1476, 2022.

Robert Kleinberg. Nearly Tight Bounds for the Continuum-Armed Bandit Problem. *Advances in Neural Information Processing Systems*, 17, 2004.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-Armed Bandits in Metric Spaces. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, pp. 681–690, 2008.

Saurabh Kumar, Henrik Marklund, Ashish Rao, Yifan Zhu, Hong Jun Jeon, Yueyang Liu, and Benjamin Van Roy. Continual Learning as Computationally Constrained Reinforcement Learning. *arXiv preprint arXiv:2307.04345*, 2023.

Tze-Leung Lai and Sidney Yakowitz. Machine Learning and Nonparametric Bandit Theory. *IEEE Transactions on Automatic Control*, 40(7):1199–1209, 1995.

Tze Leung Lai, Michael Benjamin Sklar, and Huanzhong Xu. Bandit and Covariate Processes, with Finite or Non-Denumerable Set of Arms. *Stochastic Processes and their Applications*, 150:1222–1237, 2022.

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

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

CL Mallows and Herbert Robbins. Some Problems of Optimal Sampling Strategy. *Journal of Mathematical Analysis and Applications*, 8(1):90–103, 1964.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training Language Models to Follow Instructions with Human Feedback. *arXiv preprint arXiv:2203.02155*, 2022.

Martin L Puterman. Markov Decision Processes: Discrete Stochastic Dynamic Programming, 1994.

Mark Bishop Ring. *Continual Learning in Reinforcement Environments*. PhD thesis, The University of Texas at Austin, 1994.

Daniel Russo and Benjamin Van Roy. Satisficing in Time-Sensitive Bandit Learning. *Mathematics of Operations Research*, 2022.

Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A Tutorial on Thompson Sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.

Claude E Shannon. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3): 379–423, 1948.

Claude E. Shannon. Coding Theorems for a Discrete Source with a Fidelity Criterion. *IRE Nat. Conv. Rec., March 1959*, 4:142–163, 1959.

Tianlin Shi, Andrej Karpathy, Linxi Fan, Jonathan Hernandez, and Percy Liang. World of Bits: An Open-Domain Platform for Web-Based Agents. In *International Conference on Machine Learning*, pp. 3135–3144. PMLR, 2017.

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

Olivier Teytaud, Sylvain Gelly, and Michele Sebag. Anytime Many-Armed Bandits. In *CAP07*, 2007.

William R Thompson. On the Likelihood That One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25(3/4):285–294, 1933.

Daniel Toyama, Philippe Hamel, Anita Gergely, Gheorghe Comanici, Amelia Glaese, Zafarali Ahmed, Tyler Jackson, Shibl Mourad, and Doina Precup. AndroidEnv: A Reinforcement Learning Platform for Android. *arXiv preprint arXiv:2105.13231*, 2021.

Yifei Wang, Tavor Baharav, Yanjun Han, Jiantao Jiao, and David Tse. Beyond the Best: Distribution Functional Estimation in Infinite-Armed Bandits. *Advances in Neural Information Processing Systems*, 35:9262–9273, 2022.

Yizao Wang, Jean-Yves Audibert, and Rémi Munos. Algorithms for Infinitely Many-Armed Bandits. *Advances in Neural Information Processing Systems*, 21, 2008.

Sid Yakowitz and Wing Lowe. Nonparametric Bandit Methods. *Annals of Operations Research*, 28:297–312, 1991.

Shunyu Yao, Howard Chen, John Yang, and Karthik R Narasimhan. WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents. In *Advances in Neural Information Processing Systems*, 2022.

Yinglun Zhu and Robert Nowak. On Regret with Multiple Best Arms. *Advances in Neural Information Processing Systems*, 33:9050–9060, 2020.

A Computational Experiment

In this section, we provide additional details on the concluding computational experiment of Section 4. The environment can be seen as a restricted version of Example 1 where the action set $\mathcal{A} = \{0, 1, \ldots, 99\}$ consists of all two-digit sequences and the agent aims to learn the first two digits of π with $\alpha = 2$.

Given the agent's current (posterior) beliefs about the underlying environment $\mathbb{P}(\theta \in \cdot \mid H_t)$, a Thompson Sampling agent proceeds via the probability-matching principle to select an action A_t such that $\mathbb{P}(A_t = a \mid H_t) = \mathbb{P}(A^\star = a \mid H_t)$, where $A^\star \in \arg\max_{a \in \mathcal{A}} \mathbb{E}\left[R_{1,a} \mid \theta\right]$. Broadly speaking, one might consider an alternative learning target $\chi \in \mathcal{A}$ such that an agent may employ a variant of Thompson Sampling by probability matching with respect to χ: $\mathbb{P}(A_t = a \mid H_t) = \mathbb{P}(\chi = a \mid H_t)$.

A line of work (Arumugam & Van Roy, 2021a;b; Arumugam et al., 2024) studies how to compute such a learning target via information theory (Shannon, 1948; Cover & Thomas, 2012). More specifically, when targeting an optimal action $A^\star$, the mutual information between the environment and target $\mathbb{I}_t(\theta; A^\star)$ given the current (random) history H_t quantifies the amount of information an agent must obtain through prudent exploration in order to identify optimal behavior. In the context of this work, a complex environment is one for which this amount of information is near-infinite or intractably large $\mathbb{I}_t(\theta; A^\star) \uparrow \infty$ across all time periods. Consequently, an agent may instead find it fruitful to orient exploration around an alternative target χ that is easier to learn, in the sense that $\mathbb{I}_t(\theta; \chi) \leq \mathbb{I}_t(\theta; A^\star)$. Of course, as the agent still aims to be productive with respect to the task at hand and optimize reward, this should be done carefully so as to incur bounded expected regret in each time period: $\mathbb{E}_t\left[R_{t,A^\star} - R_{t,\chi}\right] \leq D$, for some threshold $D \in \mathbb{R}_{\geq 0}$.

Striking a desired balance between information and utility is a hallmark characteristic of lossy compression problems studied by the information theory community within the sub-area of rate-distortion theory (Shannon, 1959; Berger, 1971). The fundamental limit for the lossy compression faced by an agent is given by the rate-distortion function

$$\mathcal{R}_t(D) = \inf_{\widetilde{A}} \mathbb{I}_t(\theta; \widetilde{A}) \text{ such that } \mathbb{E}_t\left[\left(R_{t,A^\star} - R_{t,\widetilde{A}}\right)^2\right] \leq D.$$

The Rate-Distortion Thompson Sampling (RDTS) algorithm of Arumugam et al. (2024) provides a theoretical analysis outlining the benefits of computing and probability matching with respect to the target $\widetilde{A}_t$ that achieves the rate-distortion limit in each time period. While their study pertains to a fixed distortion threshold $D \in \mathbb{R}_{\geq 0}$, our computational experiments employ an adaptive version where the dynamic threshold D_t is chosen to ensure the agent focuses its efforts on identifying the digits of π in sequence, rather than pursuing a guess for all digits at once like Thompson Sampling (corresponding to $D_t = 0$ for all time periods). While previous work (Arumugam & Van Roy, 2021a) has avoided dealing directly with the rate-distortion function computationally by appealing to the classic Blahut-Arimoto algorithm (Blahut, 1972; Arimoto, 1972), our experiment leverages the fact that the rate-distortion function constitutes a convex optimization problem (Csiszár, 1974; Chiang & Boyd, 2004) which can be solved in each time period via CVXPY (Diamond & Boyd, 2016).

B Analysis

In this section, we provide all proofs for theoretical results presented in the main paper.

B.1 Proof of Theorem 2

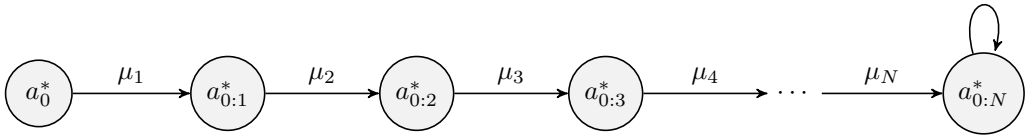
Proof. For ease of exposition, we refer to each component in an action tuple $(a_1, \ldots, a_k)$ as one *digit*. The analogy is made for the illustrative π-guessing game in Section 1. A history $H_t = (A_0, R_1, A_1, \ldots, A_t, R_{t+1})$ can be completely characterized by an agent state S_t made up of digits tried and failed so far as well as

a_k^* known so far. Selecting digits that the agent has tried and failed before is clearly suboptimal, as those digits incur a cost yet offer no new information. Therefore, it suffices to restrict our attention to (stationary) policies with states S_t that do not try digits already attempted and failed.

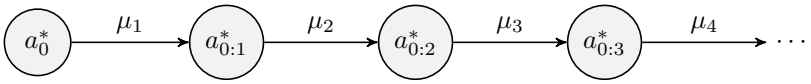
By the temporal symmetric structure of Example 1, if $a^*_{1:k} \in H_t$ and $a^*_{1:k+1} \notin H_t$, a (stationary) policy at time t must sample from two actions: 1) exploit $a^*_{1:k}$ or 2) explore the next digit.

We consider two classes of policies:

1. π_N: for each $N \in \mathbb{Z}_{>0}$, π_N explores to identify $a^*_{0:N}$ sequentially, then exploits $a^*_{0:N}$.



2. π^{explore}: π^{explore} always explores sequentially.



We recursively define stopping times for the exploration of each digit under the always exploring agent π^{explore}. Let $\mu_0 = 1$ and $\mu_k = \min\{t > 0 : A_{t+\sum_{j=0}^{k-1}\mu_j} = a^*_{1:k}, A_i \sim \pi^{\text{explore}}\}$. In other words, μ_k denotes the time it takes for π^{explore} to discover the k-th digit given that it knows the first $k-1$ digits. The agent's prior implies that μ_k is i.i.d. geometric with mean τ. An useful quantity is the expected discount factor at μ_k:

$$\mathbb{E}\left[\gamma^{\mu_k}\right] = \sum_{j=1}^{\infty} \mathbb{P}(\mu_k = j)\gamma^j = \sum_{j=1}^{\infty} \left(1 - \frac{1}{\tau}\right)^{j-1} \frac{1}{\tau}\gamma^j = \frac{\gamma}{(1-\gamma)\tau + \gamma}.$$

We consider two cases: $\gamma < 1$ and $\gamma = 1$. When $\gamma < 1$, the expected reward at time $\sum_{j=0}^{k}\mu_j + 1$ is

$$\mathbb{E}\left[\gamma^{\sum_{j=0}^{k}\mu_j - 1}\alpha^k\right] = \frac{\alpha}{(1-\gamma)\tau + \gamma}\left(\frac{\alpha\gamma}{(1-\gamma)\tau + \gamma}\right)^{k-1},$$

and the expected cost accumulated while exploring for a_k^* satisfies

$$\mathbb{E}\left[\sum_{i=\sum_{j=0}^{k-1}\mu_j + 1}^{\sum_{j=0}^{k}\mu_j - 1} \gamma^{i-1}\frac{\alpha+1}{\tau-1}\alpha^{k-1}\right] = \frac{\alpha+1}{(1-\gamma)\tau + \gamma}\left(\frac{\alpha\gamma}{(1-\gamma)\tau + \gamma}\right)^{k-1} = \frac{\alpha+1}{\alpha}\mathbb{E}\left[\gamma^{\sum_{j=0}^{k}\mu_j - 1}\alpha^k\right].$$

Assuming that T is large enough such that $T \gg \sum_{j=0}^{N} \mu_j$ for fixed N almost surely, we calculate the returns over a horizon T.

$$\begin{aligned}
V_{\pi_N}^{\gamma,T} &= \mathbb{E}\left[\sum_{k=0}^{N-1} \gamma^{\sum_{j=0}^{k}\mu_j - 1}\alpha^k - \sum_{k=1}^{N}\frac{\alpha+1}{\alpha}\gamma^{\sum_{j=0}^{k}\mu_j-1}\alpha^k + \sum_{i=\sum_{j=0}^{N}\mu_j}^{T}\gamma^{i-1}\alpha^N\right] \\
&= 1 - \sum_{k=1}^{N-1}\frac{1}{(1-\gamma)\tau+\gamma}\left(\frac{\alpha\gamma}{(1-\gamma)\tau+\gamma}\right)^{k-1} - \frac{\alpha+1}{(1-\gamma)\tau+\gamma}\left(\frac{\alpha\gamma}{(1-\gamma)\tau+\gamma}\right)^{N-1} \\
&\quad + \alpha^N\frac{\left(\frac{\gamma}{(1-\gamma)\tau+\gamma}\right)^N - \gamma^{T+1}}{\gamma(1-\gamma)} \\
&= 1 - \frac{1}{(1-\gamma)\tau+(1-\alpha)\gamma} + \left(\frac{1}{(1-\gamma)\tau+(1-\alpha)\gamma} + \frac{1}{1-\gamma}\right)\left(\frac{\gamma}{(1-\gamma)\tau+\gamma}\right)^N - \frac{\alpha^N}{1-\gamma}\gamma^T.
\end{aligned}$$

For $N_1 < N_2$, we have that

$$\begin{aligned}
V_{\pi_{N_2}}^{\gamma,T} - V_{\pi_{N_1}}^{\gamma,T} &= \left(\frac{1}{(1-\gamma)\tau+(1-\alpha)\gamma} + \frac{1}{1-\gamma}\right)\left[\left(\frac{\alpha\gamma}{(1-\gamma)\tau+\gamma}\right)^{N_2} - \left(\frac{\alpha\gamma}{(1-\gamma)\tau+\gamma}\right)^{N_1}\right] \\
&\quad - \frac{\alpha^{N_2}-\alpha^{N_1}}{1-\gamma}\gamma^T \\
&\xrightarrow{T\to\infty} \left(\frac{1}{(1-\gamma)\tau+(1-\alpha)\gamma} + \frac{1}{1-\gamma}\right)\left[\left(\frac{\alpha\gamma}{(1-\gamma)\tau+\gamma}\right)^{N_2} - \left(\frac{\alpha\gamma}{(1-\gamma)\tau+\gamma}\right)^{N_1}\right] \\
&> 0.
\end{aligned}$$

This implies that a policy can always improve by exploring more digits before committing. Therefore, a policy that stops exploring is never discounted-overtaking optimal.

When $\gamma = 1$, the expected cost accumulated while exploring for a_k^* satisfies:

$$\mathbb{E}\left[\sum_{i=\sum_{j=0}^{k-1}\mu_j+1}^{\sum_{j=0}^{k}\mu_j-1}\frac{\alpha+1}{\tau-1}\alpha^{k-1}\right] = (\alpha+1)\alpha^{k-1}.$$

Assuming that T is large enough such that $T \gg \sum_{j=0}^{N} \mu_j$ for fixed N almost surely, we calculate the returns over a horizon T:

$$\begin{aligned}
V_{\pi_N}^{1,T} &= \mathbb{E}\left[\sum_{k=0}^{N-1}\alpha^k - \sum_{k=1}^{N}\sum_{i=\sum_{j=0}^{k-1}\mu_j+1}^{\sum_{j=0}^{k}\mu_j-1}\frac{\alpha+1}{\tau-1}\alpha^{k-1} + \sum_{i=\sum_{j=0}^{N}\mu_j}^{T}\alpha^N\right] \\
&= \mathbb{E}\left[\sum_{k=0}^{N-1}\alpha^k - \sum_{k=1}^{N}(\alpha+1)\alpha^{k-1} + (T - \sum_{j=0}^{N}\mu_j + 1)\alpha^N\right] \\
&= -\alpha\frac{\alpha^N-1}{\alpha-1} + (T-N\tau)\alpha^N.
\end{aligned}$$

For $N_1 < N_2$, we have that

$$\liminf_{T\to\infty}(V_{\pi_{N_2}}^{1,T} - V_{\pi_{N_1}}^{1,T}) > 0.$$

This again implies that a policy can always improve by exploring more digits before committing. Therefore, a policy that stops exploring is never discounted-overtaking optimal. Moreover,

$$\text{Regret}^T_{\pi_{N_2},\pi_{N_1}} \geq \mathcal{O}(T).$$

□

B.2 Proof of Theorem 1

Proof. We consider two cases: $\gamma = 1$ and $\gamma < 1$. When $\gamma < 1$, the return of the exploring agent over a finite horizon T can be computed as

$$\begin{aligned}
V^{\gamma,T}_{\pi^{\text{explore}}} &= \mathbb{E}\Bigg[\sum_{N=1}^{\infty}\left(\sum_{k=0}^{N}\gamma^{\sum_{j=0}^{k}\mu_j-1}\alpha^k - \sum_{k=1}^{N}\frac{\alpha+1}{\alpha}\gamma^{\sum_{j=0}^{k}\mu_j-1}\alpha^k - \sum_{i=\sum_{j=0}^{N}\mu_j+1}^{T}\gamma^{i-1}\frac{\alpha+1}{\tau-1}\alpha^N\right)\\
&\qquad \cdot \mathbf{1}\{\sum_{j=0}^{N}\mu_j \leq T < \sum_{j=0}^{N+1}\mu_j\}\Bigg]\\
&\leq \mathbb{E}\left[\sum_{N=1}^{\infty}\left(\sum_{k=0}^{N}\gamma^{\sum_{j=0}^{k}\mu_j-1}\alpha^k - \sum_{k=1}^{N}\frac{\alpha+1}{\alpha}\gamma^{\sum_{j=0}^{k}\mu_j-1}\alpha^k\right)\cdot\mathbf{1}\{\sum_{j=0}^{N}\mu_j \leq T < \sum_{j=0}^{N+1}\mu_j\}\right]\\
&= \mathbb{E}\left[\sum_{N=1}^{\infty}\left(1 - \frac{1}{(1-\gamma)\tau+\gamma}\sum_{k=1}^{N}\left(\frac{\alpha\gamma}{(1-\gamma)\tau+\gamma}\right)^{k-1}\right)\cdot\mathbf{1}\{\sum_{j=0}^{N}\mu_j \leq T < \sum_{j=0}^{N+1}\mu_j\}\right].
\end{aligned}$$

Comparing π_1 and π^{explore},

$$\begin{aligned}
V^{1,T}_{\pi_1} - V^{1,T}_{\pi^{\text{explore}}} &= \sum_{N=1}^{\infty}\sum_{k=2}^{N}\left(\frac{\alpha\gamma}{(1-\gamma)\tau+\gamma}\right)^{k-1}\cdot\mathbf{1}\{\sum_{j=0}^{N}\mu_j \leq T < \sum_{j=0}^{N+1}\mu_j\} + \frac{1}{1-\gamma}\left(\frac{\alpha\gamma}{(1-\gamma)\tau+\gamma}\right)\\
&\quad - \frac{\gamma^T}{1-\gamma}\alpha\\
&\xrightarrow{T\to\infty}\left(\frac{\alpha\gamma}{(1-\gamma)\tau+\gamma}\right)\frac{1}{1-\frac{\alpha\gamma}{(1-\gamma)\tau+\gamma}} + \frac{1}{1-\gamma}\left(\frac{\alpha\gamma}{(1-\gamma)\tau+\gamma}\right)\\
&= \left(\frac{\alpha\gamma}{(1-\gamma)\tau+\gamma}\right)\left(\frac{(1-\gamma)\tau+\gamma}{(1-\gamma)\tau+(1-\alpha)\gamma} + \frac{1}{1-\gamma}\right).
\end{aligned}$$

Since $\tau < \frac{\gamma(\alpha-1)}{2(1-\gamma)}$, one can check that

$$\frac{(1-\gamma)\tau+\gamma}{(1-\gamma)\tau+(1-\alpha)\gamma} + \frac{1}{1-\gamma} > 0$$

and thus $\liminf_{T\to\infty}(V^{1,T}_{\pi_1} - V^{1,T}_{\pi^{\text{explore}}}) > 0$. This implies that the exploring agent using π^{explore} is never discounted-overtaking optimal.

When $\gamma = 1$, the return of the exploring agent over a finite horizon T can be computed as

$$
\begin{aligned}
V^{1,T}_{\pi^{\text{explore}}} &= \mathbb{E}\left[\sum_{N=1}^{\infty}\left(\sum_{k=0}^{N}\alpha^k - \sum_{k=1}^{N}\sum_{i=\sum_{j=0}^{k-1}\mu_j+1}^{\sum_{j=0}^{k}\mu_j-1}\frac{\alpha+1}{\tau-1}\alpha^{k-1} - (T-\sum_{j=0}^{N}\mu_j)\frac{\alpha+1}{\tau-1}\alpha^N\right)\cdot\mathbf{1}\{\sum_{j=0}^{N}\mu_j \le T < \sum_{j=0}^{N+1}\mu_j\}\right] \\
&\le \mathbb{E}\left[\sum_{N=1}^{\infty}\left(\sum_{k=0}^{N}\alpha^k - \sum_{k=1}^{N}\sum_{i=\sum_{j=0}^{k-1}\mu_j+1}^{\sum_{j=0}^{k}\mu_j-1}\frac{\alpha+1}{\tau-1}\alpha^{k-1}\right)\cdot\mathbf{1}\{\sum_{j=0}^{N}\mu_j \le T < \sum_{j=0}^{N+1}\mu_j\}\right] \\
&= \mathbb{E}\left[\sum_{N=1}^{\infty}\left(\alpha^N - \alpha\frac{\alpha^N-1}{\alpha-1}\right)\cdot\mathbf{1}\{\sum_{j=0}^{N}\mu_j \le T < \sum_{j=0}^{N+1}\mu_j\}\right] \\
&\le \mathbb{E}\left[\sum_{N=1}^{\infty}\left(\alpha^N - \alpha^N\right)\cdot\mathbf{1}\{\sum_{j=0}^{N}\mu_j \le T < \sum_{j=0}^{N+1}\mu_j\}\right] = 0.
\end{aligned}
$$

Thus, comparing π_1 and π^{explore},

$$\liminf_{T\to\infty}(V^{1,T}_{\pi_1} - V^{1,T}_{\pi^{\text{explore}}}) > 0.$$

This implies that the exploring agent using π^{explore} is never discounted-overtaking optimal. Moreover,

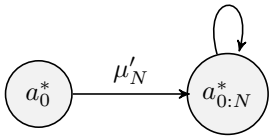
$$\text{Regret}^{T}_{\pi_1,\pi^{\text{explore}}} \ge \mathcal{O}(T).$$

□

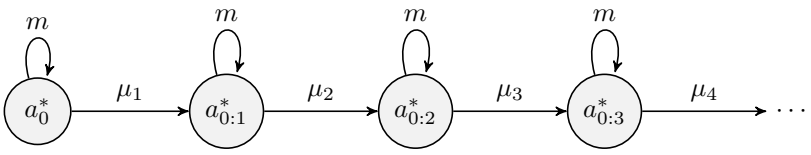
B.3 Proof of Theorem 3

Proof. We consider three additional classes of policies to those mentioned in the proof of Theorem 2:

1. Stochastic policies π^p: for $p \in [0,1)$, π^p exploits the best known action with probability p and explores the next digit with probability $1-p$ at each time. Note that $\pi^0 = \pi^{\text{explore}}$.
2. Non-curricular policies π'_N: for each $N \in \mathbb{Z}_{>0}$, π'_N explores to identify $a^*_{1:N}$ directly, then exploits $a^*_{1:N}$.



3. Non-stationary policies π^{NS}_m: for $m \in \mathbb{R}_{\ge 0}$, after each $a^*_{1:k}$ is discovered, π^{NS}_m exploits $a^*_{1:k}$ for m times before exploring the next digit.



First we note a one-to-one mapping between the set of stochastic policies $\{\pi^p\}_{p\in[0,1)}$ and non-stationary policies $\{\pi^{\text{NS}}_m\}_{m\in\mathbb{R}_{\ge 0}}$. Indeed, when $p = \frac{m+1}{m+\tau}$, each $a^*_{1:k}$ is selected $m+1$ times on average.

For non-curricular policies, under the agent's prior belief, the optimal strategy of guessing is to guess all the combinations of $(a_1, \ldots, a_N)$ where $\sum_{i=1}^N a_i = N$, then those where $\sum_{i=1}^N a_i = N+1$, and so on. We take π'_N to be such a policy and define a stopping time $\mu'_N = \min\{t > 0 | A_t = a^*_{1:N}, A_i \sim \pi'_N\}$. We calculate the returns for π'_N and π^p as follows.

$$
\begin{aligned}
V^{1,T}_{\pi^{\mathrm{NS}}_m} = \sum_{n=1}^{\infty} \mathbb{E}\Bigg[& \mathbf{1}\{\sum_{j=0}^{n} \mu_j + nm \le T\} \left((m+1)\alpha^{n-1} - (\mu_n - 1)\frac{\alpha+1}{\tau-1}\alpha^{n-1} \right) \\
& + \mathbf{1}\{\sum_{j=1}^{n} \mu_j + nm \le T \le \sum_{j=1}^{n} \mu_j + (n+1)m\} \left(T - \sum_{j=1}^{n} \mu_j - nm + 1 \right) \alpha^n \\
& + \mathbf{1}\{\sum_{j=1}^{n} \mu_j + (n+1)m < T < \sum_{j=1}^{n+1} \mu_j + (n+1)m\} \left((m+1)\alpha^n - (T - \sum_{j=1}^{n} \mu_j - (n+1)m)\frac{\alpha+1}{\tau-1}\alpha^n \right) \Bigg].
\end{aligned} \tag{1}
$$

$$
\begin{aligned}
V^{1,T}_{\pi'_N} &= \mathbb{E}\left[-\frac{\alpha+1}{\tau-1}\alpha^{N-1}\mu'_N + (T - \mu'_N + 1)\alpha^N) \right] \\
&= -\mathbb{E}\left[\mu'_N\right]\frac{\alpha+1}{\tau-1}\alpha^{N-1} + \mathbb{E}\left[T - \mu'_N + 1\right]\alpha^N.
\end{aligned}
$$

Note that if $m > T$, $V^{1,T}_{\pi^{\mathrm{NS}}_m} = 0$. Hence, it suffices to consider $m \in [0, T]$. Since $V^{1,T}_{\pi^{\mathrm{NS}}_m}$ is continuous in m on $[0, T]$, by the extreme value theorem, there exists an $m^* \in [0, T]$ such that $V^{1,T}_{\pi^{\mathrm{NS}}_{m^*}} = \max_{m \in [0,T]} V^{1,T}_{\pi^{\mathrm{NS}}_m}$. Taking $p^*_T = \frac{m^*}{m^*+1}$, we have

$$
\mathrm{Regret}_{\pi, \pi^*_T} \le 0.
$$

Finally, we prove that an agent is better off following the curriculum, i.e., guessing one digit at a time in order. Recall that

$$
V^{1,T}_{\pi_N} = -\alpha\frac{\alpha^N - 1}{\alpha - 1} + (T - N\tau)\alpha^N.
$$

For $j \in \mathbb{Z}_{\ge 0}$, define $s_j = \sum_{i=1}^{j} \binom{N+i-2}{N-1}$, then

$$
\begin{aligned}
\mathbb{E}\left[\mu'_N\right] &= \sum_{n=N}^{\infty} \sum_{\ell=s_{n-N}+1}^{s_{n-N+1}} \ell \left(1 - \frac{1}{\lambda}\right)^{n-N} \left(\frac{1}{\lambda}\right)^N \\
&= \sum_{n=N}^{\infty} \frac{(s_{n-N} + s_{n-N+1} + 1)(s_{n-N+1} - s_{n-N})}{2} \left(1 - \frac{1}{\lambda}\right)^{n-N} \left(\frac{1}{\lambda}\right)^N \\
&= \frac{1}{2} \sum_{n=N}^{\infty} \left[2 \sum_{i=1}^{n-N} \binom{N+i-2}{N-1} + \binom{n-1}{N-1} + 1 \right] \binom{n-1}{N-1} \left(1 - \frac{1}{\lambda}\right)^{n-N} \left(\frac{1}{\lambda}\right)^N \\
&\gg N\tau + 1.
\end{aligned}
$$

Thus, $\mathbb{E}[T - \mu'_N + 1] < T - N\tau$ and

$$
\mathbb{E}[\mu'_N]\frac{\alpha+1}{\tau-1}\alpha^{N-1} > (N\tau + 1)\frac{\alpha+1}{\tau-1}\alpha^{N-1} > \alpha\frac{\alpha^N - 1}{\alpha - 1}
$$

for sufficiently large N. Thus, there exists an $N_0 \in \mathbb{Z}_{>0}$ such that for all $N > N_0$, $V^{1,T}_{\pi'_N} < V^{1,T}_{\pi_N}$.

□

B.4 Proof Roadmap for Conjecture 1

Proof roadmap. Recall the expression for $V^{1,T}_{\pi^{\mathrm{NS}}_m}$ in Equation (1). Consider the first summand in the expected value. We let

$$f_n(m;T) = \mathbb{E}\left[\mathbf{1}\{\sum_{j=0}^{n}\mu_j + nm \le T\}\left((m+1)\alpha^{n-1} - (\mu_n - 1)\frac{\alpha+1}{\tau-1}\alpha^{n-1}\right)\right].$$

To decouple the dependence between the indicator random variable $\mathbf{1}\{\sum_{j=0}^{n}\mu_j + nm \le T\}$ and the multiplier $\left((m+1)\alpha^{n-1} - (\mu_n - 1)\frac{\alpha+1}{\tau-1}\alpha^{n-1}\right)$, we define an independent copy $\tilde{\mu}_n$ of μ_n and analyze

$$\begin{aligned}
\tilde{f}_n(m;T) &= \mathbb{E}\left[\mathbf{1}\{\sum_{j=0}^{n-1}\mu_j + \tilde{\mu}_n + nm \le T\}\left((m+1)\alpha^{n-1} - (\mu_n - 1)\frac{\alpha+1}{\tau-1}\alpha^{n-1}\right)\right] \\
&= \mathbb{P}\left(\sum_{j=0}^{n-1}\mu_j + \tilde{\mu}_n + nm \le T\right)\mathbb{E}\left[(m+1)\alpha^{n-1} - (\mu_n - 1)\frac{\alpha+1}{\tau-1}\alpha^{n-1}\right] \\
&= \mathbb{P}\left(\sum_{j=0}^{n-1}\mu_j + \tilde{\mu}_n + nm \le T\right)(m-\alpha)\alpha^{n-1}.
\end{aligned}$$

We may then proceed to lower bound $\mathbb{P}\left(\sum_{j=0}^{n-1}\mu_j + \tilde{\mu}_n + nm \le T\right)$ using Kolmogorov's inequality. An upper bound on $\tilde{f}_n(m;T)$ can be obtained by taking each $\mu_j = 1$ in the indicator. Finally, we account for the difference $f_n(m;T) - \tilde{f}_n(m;T)$. A similar analysis can be carried out for the second and third summand in Equation (1). Optimizing the upper and lower bounds of $V^{1,T}_{\pi^{\mathrm{NS}}_m}$ over m gives two sequences $\overline{m}^*_T$ and $\widehat{m}^*_T$, respectively, which both converge to α. Thus, the corresponding exploitation probabilities should converge to $\frac{\alpha+1}{\alpha+\tau}$. □