Chained Information-Theoretic Bounds and Tight Regret Rate for Linear Bandit Problems

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper studies the Bayesian regret of a variant of the Thompson Sampling algorithm for bandit problems. It builds upon the information-theoretic framework of Russo and Van Roy (2015) and, more specifically, on the rate-distortion analysis from Dong and Van Roy (2020), where they proved a bound with regret rate of $O(d\sqrt{T\log(T)})$ for the d-dimensional linear bandit setting. We focus on bandit problems with a metric action space, and, using a chaining argument, we establish new bounds that depend on the action space's metric entropy for a Thompson Sampling variant. Under suitable continuity assumption of the rewards, our bound offers a tight rate of $O(d\sqrt{T})$ for d-dimensional linear bandit problems.

1 Introduction

Bandit problems are decision problems in which an agent interacts sequentially with an unknown environment by choosing actions and earning rewards in return. The agent's goal is to maximize its expected cumulative reward, the expected sum of rewards it will earn throughout its interaction with the environment. This necessitates a delicate balance between exploring different actions to gather information for potential future rewards and exploiting known actions to receive immediate gains. The theoretical study of the performance of an algorithm in a bandit problem is done by analyzing the *expected regret*, which is defined as the difference between the cumulative reward of the algorithm and the hypothetical cumulative reward that an oracle would obtain by choosing the optimal action at each time step. An effective method for achieving small regret is the Thomson Sampling (TS) algorithm (Thompson, 1933), which, despite its simplicity, has shown remarkable performance (Russo et al., 2018; Russo and Van Roy, 2017; Chapelle and Li, 2011).

Studying the Thompson Sampling regret, Russo and Van Roy (2015) introduced the concept of information ratio. This statistic captures the trade-off between the information gained by the algorithm about the environment and the immediate regret. They used this concept to provide a general upper bound for finite action spaces $\mathcal{A}$ that depends on the entropy of the optimal action $\mathrm{H}(A^\star)$, the time horizon T (the total number of times that the agent interacts with the environment), and a problem-dependent upper bound on the information-ratio Γ, namely $\sqrt{\Gamma \cdot T \cdot \mathrm{H}(A^\star)}$. For finite environment parameter spaces, under a Lipschitz continuity assumption of the expected reward and using Lipschitz maximal inequality argument, Dong and Van Roy (2020) were able to control the regret of the TS algorithm via a "compressed statistic" Θ_ε of the environment parameters Θ, with a bound of the form $\varepsilon \cdot T + \sqrt{\Gamma \cdot T \cdot \mathrm{H}(\Theta_\varepsilon)}$. In particular, they derived a near-optimal regret rate of $O(d\sqrt{T \log T})$ for d-dimensional linear bandit problems.

Submitted to the ICML 2024 Workshop on Foundations of Reinforcement Learning and Control. Do not distribute.

In this paper, building on the work of Dong and Van Roy (2020), we explored using the chaining technique for bandit problems where the rewards exhibit some subgaussian continuity property with respect to the action space. We introduced the *Two Steps Thompson Sampling* (2-TS), a variant of the original algorithm where the history is updated every time step. For this algorithm, we derive a bound that captures the continuity property of the reward process and depends on the metric entropy of the action space. Notably, our bound does not require a finite environment or action space and holds for continuous action spaces. For the class of linear bandit problems, we obtained a bound in $O(d\sqrt{T})$ matching the best possible regret $\Omega(d\sqrt{T})$ from Dani et al. (2008).

The rest of the paper is organized as follows. Section 2 presents the bandit problem setup, defines the Bayesian expected regret, and introduces the Two Steps Thompson Sampling algorithm and the specific notations. Section 3 explains the idea of the bounding technique, and defines the required tools and assumptions we will be using. Section 4 states and proves our main Theorem. Section 5 applies our Theorem to the important case of linear bandit problems and derives several specific bounds before giving s a bound for linear bandit problems with a ball-structured action space. Finally, Section 6 discusses our results, possible extensions, and future work.

2 Problem setup

We consider a sequential decision problem, where at each time step (or round) $t \in \{1, \dots, T\}$, an agent interacts with an environment by selecting an action A_t from an action set $\mathcal{A}$ and, based on that action, receives a real-valued reward $R_t \in \mathbb{R}$. The pair of the selected action and the received reward is collected in a history $H^{t+1} = H^t \cup H_{t+1}$, where $H_{t+1} = \{A_t, R_t\}$, that will be accessible to the agent in the next round. The procedure repeats until the last round $t = T$.

Following the Bayesian framework, we consider the environment to be characterized by some parameters $\theta \in \mathcal{O}$, unknown to the agent, sampled from a known prior distribution $\mathbb{P}_\Theta$. This prior, together with the reward distribution $\mathbb{P}_{R|A,\Theta}$ fully describes the bandit problem. As the reward distribution depends on the selected action and the environment parameters, it may be written as $R_t = R(A_t, \Theta)$ for some possibly random function $R : \mathcal{A} \times \mathcal{O} \to \mathbb{R}$.

The agent's goal is to take actions that maximize the total collected reward. More specifically, the agent seeks to learn a policy $\varphi = \{\varphi_t : \mathcal{H}^t \to \mathcal{A}\}_{t=1}^T$ that, for each time $t \in \{1, \dots, T\}$, selects an action A_t based on the history H^t such that it maximizes the *expected cumulative reward* $R_T(\varphi) \coloneqq \mathbb{E}\left[\sum_{t=1}^T R(\varphi_t(H^t), \Theta)\right]$.

2.1 The Bayesian expected regret

The Bayesian expected regret quantifies the difference between the expected cumulative reward achieved by the agent following a policy φ and the optimal expected cumulative reward that could be obtained by an *omniscient* agent having access to the true reward function and selecting the action yielding the highest expected reward.

Definition 1 (Optimal cumulative reward) *The* optimal cumulative reward *of a bandit problem is defined as*

$$R_T^\star \coloneqq \sup_\psi \mathbb{E}\left[\sum_{t=1}^T R(\psi(\Theta), \Theta)\right],$$

where the supremum is taken over all decision rules $\psi : \mathcal{O} \to \mathcal{A}$ such that the expectation above is defined.

We denote a policy that achieves the supremum of Definition 1 as $\psi^\star$ and we refer to the action it selects as the *optimal action* $A^\star \coloneqq \psi^\star(\Theta)$. We make the following technical assumption on the action set to ensure such a policy exists.

Assumption 1 (Compact action set) *The set of actions $\mathcal{A}$ is compact.*

The difference between the optimal expected cumulative reward and expected cumulative reward of a policy φ is called the Bayesian expected regret of φ, denoted $\mathrm{REG}_T(\varphi)$.

Definition 2 (Bayesian expected regret) *The* Bayesian expected regret *of a policy φ in a bandit problem is defined as*

$$\mathrm{REG}_T(\varphi) \coloneqq R_T^\star - R_T(\varphi).$$

2.2 Thompson Sampling algorithm and the Two Steps variant

One of the most popular and most studied algorithms for solving bandit problems is the *Thompson Sampling* (TS) algorithm Russo et al. (2018); Russo and Van Roy (2017); Chapelle and Li (2011); Dong and Van Roy (2020). TS works by sampling a Bayesian estimate of the environment parameters from the posterior distribution and taking the optimal action for the sampled estimate. Specifically, at each time step $t \in \{1, \ldots, T\}$, the agent draws a Bayesian estimate $\hat{\Theta}_t$ based on the past collected history H^t, takes the corresponding optimal action $\hat{A}_t = \psi^\star(\hat{\Theta}_t)$, receives a reward R_t, and updates the history $H^{t+1} = \{H^t, \hat{A}_t, R_t\}$.

In this work, we consider a variation of TS, which we refer to as *Two Steps Thompson Sampling* (2-TS). The critical difference between this algorithm and the TS algorithm is that the history is updated every two time steps[1]. Intuitively, the algorithm will behave the same but wait to collect two rewards before updating its history. This modification in the history update is motivated by theoretical needs. Specifically, the chaining technique requires controlling the differences between consecutive regret approximations. In our analysis, those differences are bounded via the information gained upon observing the rewards corresponding to two approximate actions. The pseudocode for Two Steps Thompson Sampling is given in Algorithm 1.

Algorithm 1 Two Steps Thompson Sampling algorithm

```
1:  Input: environment parameters prior P_Θ.
2:  for t = 1 to T do
3:      Sample a parameter estimation Θ̂_t ~ P_{Θ|H^t}.
4:      Take the corresponding optimal action Â_t = ψ*(Θ̂_t).
5:      Collect the reward R_t = R(Â_t, Θ).
6:      if t is even then
7:          Update the history H^{t+1} = {H^t, Â_t, R_t, Â_{t-1}, R_{t-1}}.
8:      else
9:          Keep the history H^{t+1} = H^t.
10:     end if
11: end for
```

2.3 Notation specific to bandit problems

Since the σ-algebras of the history H^t are often used in the conditioning of the expectations and probabilities coming up in the analysis, similarly to Russo and Van Roy (2015); Dong and Van Roy (2020); Neu et al. (2022); Gouverneur et al. (2023), we define the operators $\mathbb{E}_t[\cdot] \coloneqq \mathbb{E}[\cdot|H^t]$ and $\mathbb{P}_t[\cdot] \coloneqq \mathbb{P}[\cdot|H^t]$, whose outcomes are $\sigma(\mathcal{H}^t)$-measurable random variables and $\mathcal{H} = \mathcal{A} \times \mathbb{R}$.

Analogously, we define $\mathrm{I}_t(A^\star; R_t) \coloneqq \mathbb{E}_t[\mathrm{D}_{\mathrm{KL}}(\mathbb{P}_{R_t|H^t, A^\star} \| \mathbb{P}_{R_t|H^t})]$ as the *disintegrated* conditional mutual information between the optimal action $A^\star$ and the reward R_t, *given the history* H^t, see (Negrea et al., 2020, Definition 1.1), which is itself also a $\sigma(\mathcal{H}^t)$-measurable random variable.

When it is clear from the context that the random rewards depend on the environment parameters Θ, we will often use the notation $R(A_t)$ as a shorthand for $R(A_t, \Theta)$ to simplify the expressions.

[1]We implicitly assume that, for Two Steps Thompson Sampling, the total number of steps T is an even number.

3 Chain-link Information Ratio and Chaining Technique

In bandit problems where the rewards of nearby actions exhibit some continuity property, we aim to exploit this dependence using a chaining argument. More specifically, our idea is to approach the *Two Step Thompson Sampling* algorithm by a chain of increasingly accurate approximations, which we refer to as *"approximate learning"*.

Inspired by Dong and Van Roy (2020), our construction relies on the existence of a sequence of finer and finer quantizations $\{A_k^\star\}_{k=k_0}^\infty$ of the optimal action $A^\star$ and a corresponding carefully crafted action sampling function $f_t^k : \mathcal{A} \to \mathcal{A}$ for each round $t \in \{1, \ldots, T\}$. These quantization and sampling functions are designed to satisfy the following three requirements simultaneously:

(i) The quantizations $A_k^\star$ are less informative than $A^\star$, that is, $\mathrm{H}(A_k^\star) \leq \mathrm{H}(A^\star)$ for all $k \geq k_0$.

(ii) At each round $t \in \{1, \ldots, T\}$, the *Two Step Thompson Sampling* regret can be written as an infinite sum of the difference between the approximate learning regrets:

$$\mathbb{E}_t\Big[R(A^\star) - R(\hat{A}_t)\Big] = \sum_{k=k_0+1}^{\infty} \mathbb{E}_t\Big[\Big(R(f_t^k(A_k^\star)) - R(f_t^k(\hat{A}_{t,k}))\Big) - \Big(R(f_t^{k-1}(A_{k-1}^\star)) - R(f_t^{k-1}(\hat{A}_{t,k-1}))\Big)\Big].$$

(iii) For each time step $t \in \{1, \ldots, T\}$, and for every $k > k_0$, the regret difference between the k^{th}-consecutive *"approximate learning"* can be bounded using the information gained about the quantization $A_k^\star$ while, at the same time, it reveals no more information about the quantization $A_k^\star$ than Two Step Thompson Sampling.

3.1 Nets and quantizations

When designing the quantization $A_k^\star \in \mathcal{A}_k$ of the optimal action, we face two conflicting goals: on the one hand, we want the quantization to be as little informative about $A^\star$ as possible while, on the other hand, we want to ensure that $\mathcal{A}_k$ converges quickly to a good approximation of $\mathcal{A}$. This dual objective naturally leads to considering ε-nets.

Definition 3 (ε-net and covering number) *A set $\mathcal{N}$ is called an ε-*net *for $(\mathcal{A}, \rho)$ if, for every $a \in \mathcal{A}$, there is a $\pi(a) \in \mathcal{N}$ such that $\rho(a, \pi(a)) \leq \varepsilon$. The smallest cardinality of an ε-net for $(\mathcal{A}, \rho)$ is called the* covering number*, that is*

$$\mathcal{N}(\mathcal{A}, \rho, \varepsilon) \triangleq \inf\big\{|\mathcal{N}| : \mathcal{N} \text{ is an } \varepsilon\text{-net of } (\mathcal{A}, \rho)\big\}.$$

The covering number $\mathcal{N}(\mathcal{A}, \rho, \varepsilon)$ can be understood as a measure of the complexity of the action set $\mathcal{A}$ at the resolution ε. Equipped with this new concept, a possible k^{th}-quantization $A_k^\star$ is the quantization of the optimal action $A^\star$ at the scale 2^{-k}.

Definition 4 (k^{th}-quantization) *Let $\mathcal{A}_k$ be a 2^{-k}-net for $(\mathcal{A}, \rho)$ with an associated mapping $\pi_k : \mathcal{A} \to \mathcal{A}_k$, such that the mappings π_k are restricted to those of the form $\pi_k = \pi_k' \circ \pi_{k+1}$, where $\pi_k' : \mathcal{A}_{k+1} \to \mathcal{A}_k$. We define $A_k^\star = \pi_k(A^\star)$ as the k^{th}-quantization of the optimal action $A^\star$ with respect to $(\mathcal{A}, \rho)$. Similarly, the quantization $\hat{A}_{t,k} = \pi_k(\hat{A}_t)$ is the k^{th}-quantization of the sampled action $\hat{A}_t$.*

Note that $A_k^\star$ is completely determined by $A_{k+1}^\star$ via the mapping $\pi_k' : \mathcal{A}_{k+1} \to \mathcal{A}_k$. In the following, we set k_0 to be the largest integer such that $2^{-k_0} \geq \mathrm{diam}(\mathcal{A})$.

3.2 Existence of the *"approximate learning"*

The sequence of quantizations $\{A_k^\star\}_{k=k_0}^\infty$ given in Definition 4 satisfy Requirement (i) since there is a deterministic mapping between $A^\star$ and $A_k^\star$ (Yury Polyanskiy, 2022, Theorem 1.4 (f)). We claim that for each time step $t \in \{1, \ldots, T\}$, and for each $k > k_0$, there exists a random function $f_t^k : \mathcal{A}_k \to \mathcal{A}_k$ that satisfies Requirements (ii) and (iii).

Proposition 1 *Let $\{A_k^\star\}_{k=k_0}^\infty$ be defined as in Definition 4. For each time step $t \in \{1, \dots, T\}$, there exists a sequence of random functions $\{f_t^k\}_{k=k_0}^\infty$ that for each $k > k_0$, satisfies the following:*

(i) $\mathbb{E}_t\left[R(f_t^{k_0}(A_{k_0}^\star)) - R(f_t^{k_0}(\hat{A}_{t,k_0}))\right] = 0,$

(ii) $\lim_{k\to\infty} \mathbb{E}_t\left[R(f_t^k(A_k^\star)) - R(f_t^k(\hat{A}_{t,k}))\right] = \mathbb{E}_t\left[R(A^\star) - R(\hat{A}_t)\right]$, and

(iii) $\mathrm{I}_t\big(A_k^\star; R(f_t^k(\hat{A}_{t,k})), R(f_t^{k-1}(\hat{A}_{t,k-1})) \le \mathrm{I}_t(A_k^\star; R(\hat{A}_t), R(\hat{A}_t')\big)$*, a.s.*

where in (iii) the sampled actions $\hat{A}_t$ and $\hat{A}_t'$ are identically distributed.

Proof 1 *The proof follows closely the proof of (Dong and Van Roy, 2020, Proposition 2) and is given in Appendix B.1.*

3.3 Subgaussian process, smooth rewards, and chain-link information ratio

The motivation for using a chaining technique is to derive a regret bound that could effectively capture the dependence between the rewards of nearby actions. We conceptualize this dependence, considering that the rewards are subgaussian with respect to the actions.

Definition 5 (Subgaussian process) *A stochastic process $\{R_a\}_{a\in\mathcal{A}}$ on the metric space $(\mathcal{A}, \rho)$ is called* subgaussian *if for all $a, b \in \mathcal{A}$ and all $\lambda \in \mathbb{R}$*

$$\log \mathbb{E}\left[e^{\lambda(R_a - R_b)}\right] \le \frac{\lambda^2 \rho(a,b)^2}{2}.$$

Technically, for a process $\{R_a\}_{a\in\ cA}$ to be subgaussian it is also required that $\mathbb{E}[R_a] = 0$ for all $a \in \mathcal{A}$, see, for example (van Handel, 2016, Definition 5.20). However, we do not require this restriction moving forward. One way to interpret the subgaussian process property is to understand it as an "in-probability continuity" requirement. Actually, Definition 5, up to constant terms, can be equivalently written as

$$\mathbb{P}[|R_a - R_b| \ge t] \le 2\exp\left(-\frac{t^2}{2\rho(a,b)^2}\right)$$

for all $t \ge 0$ and all all $a, b \in \mathcal{A}$.

Lastly, we can impose the following mild technical assumption to ensure that the difference of regret between consecutive *approximate learning* vanishes asymptotically, we can impose the following mild technical assumption.

Definition 6 (Separable process) *A stochastic process $\{R_a\}_{a\in\mathcal{A}}$ is called* separable *if there is a countable set $\mathcal{A}' \subseteq \mathcal{A}$ such that, for all $a \in \mathcal{A}$*

$$R_a \in \lim_{\substack{a'\to a \\ a'\in\mathcal{A}'}} R_{a'} \ \text{ a.s.}$$

We refer to rewards satisfying both definition 5 and 6 as *smooth rewards* on the metric space $(\mathcal{A}, \rho)$.

Definition 7 (Smooth rewards) *We say that the rewards are* smooth *on the metric space $(\mathcal{A}, \rho)$, if for all environment parameters $\theta \in \mathcal{O}$, the random rewards $\{R(a,\theta)\}_{a\in\mathcal{A}}$ form a separable subgaussian process on $(\mathcal{A}, \rho)$.*

Some typical rewards for linear bandits satisfy them. For example, let $R_a := \langle a, \Theta\rangle + W_a$, where actions and parameters are in $\overline{\mathbf{B}_d(0,1)}$, and where W_a can either be some arbitrarily distributed noise independent of the action, $W_a = W$, or can be a subgaussian process w.r.t. $(\mathcal{A}, ||\cdot||_2)$ e.g. (Wainwright, 2019, Chapter 5). Indeed, in this case

$$\log \mathbb{E}\left[\exp\left(\lambda\left(R_a - R_b\right)\right)\right] \le \frac{\lambda^2\|a-b\|^2}{8} \quad \text{for all } \lambda \in \mathbb{R} \text{ and all } a, b \in \mathcal{A}. \tag{1}$$

By Cauchy–Schwarz $R_a - R_b = \langle a-b, \Theta \rangle \le \|a-b\|\|\Theta\|$. Thus, $\|\Theta\| \in [0,1]$ is a subgaussian random variable with parameter $1/2$, and therefore (1) follows and Definition 5 holds with $\rho(a,b) = \|a-b\|/2$. Finally, Definition 6 holds since R_a is continuous on a (van Handel, 2016, Remark 5.23).

To control the difference of regret between successive *approximate learning*, it is helpful to introduce the concept of *chain-link information ratio*. It is a direct adaptation of our chaining technique to the *information ratio* introduced by Russo and Van Roy (2015) and later used by Dong and Van Roy (2020).

Definition 8 (Chain-link information ratio) *For each time step $t \in \{1, \dots, T\}$, and for each $k > k_0$, we define the* chain-link information ratio *as*

$$\Gamma_{t,k} := \frac{\mathbb{E}_t\Big[\Big(R(f_t^k(A_k^\star)) - R(f_t^k(\hat{A}_{t,k}))\Big) - \Big(R(f_t^{k-1}(A_{k-1}^\star)) - R(f_t^{k-1}(\hat{A}_{t,k-1}))\Big)\Big]^2}{\mathrm{I}_t(f_t^k(A_k^\star), f_t^{k-1}(A_{k-1}^\star); R(f_t^k(\hat{A}_{t,k})), R(f_t^{k-1}(\hat{A}_{t,k-1})))}$$

where $A_k^\star, A_{k-1}^\star$ and $\hat{A}_{t,k}, \hat{A}_{t,k}$ are the k^{th} and $(k-1)^{\text{th}}$ quantizations of $A^\star$ and $\hat{A}_t$ respectively and where the random functions f_t^k and f_t^{k-1} satisfy the conditions of Proposition 1,

There is no particular interpretation of the chain-link information ratio. The purpose of its introduction is to unify elegantly specific results via problem-dependent upper bounds on $\Gamma_{t,k}$ similarly to what is done in prior works for the information ratio (Russo and Van Roy, 2015; Dong and Van Roy, 2020) and the lifted information ratio (Neu et al., 2022; Gouverneur et al., 2023).

4 Main result

In this section, we leverage the previously introduced concepts to derive a general chained bound on the Two Steps Thompson Sampling regret for bandit problems with smooth rewards. We obtain a bound that depends on the complexity of the action space. Remarkably, through the use of Lemma 1 (see in Appendix A), our results hold for continuous action spaces. We note that Lemma 1 could be applied to Dong and Van Roy (2020) as a generalization of their (Dong and Van Roy, 2020, Lemma 1), thus extending their results to infinite and continuous environment spaces.

Theorem 1 (Chained bound) *For bandit problems with smooth rewards on the metric space $(\mathcal{A}, \rho)$, the* 2-TS *expected cumulative regret after T steps is bounded as*

$$\mathrm{REG}_T^{\text{2-TS}} \le \sum_{k=k_0+1}^{\infty} \sqrt{2 \cdot \bar{\Gamma}_k \cdot T \cdot \mathrm{H}(A_k^\star)},$$

where $A_k^\star$ is the k^{th}-quantization about the optimal action $A^\star$ with respect to $(\mathcal{A}, \rho)$ and where for each $k > k_0$, and $\bar{\Gamma}_k$ is a upper bound on $\mathbb{E}[\Gamma_{t,k}]$.

Proof 2 *We start by rewriting the expected regret of 2-TS as a sum of consecutive regret differences between two consecutive "approximate learning":*

$$\begin{aligned}\mathrm{REG}_T^{\text{2-TS}} &= \sum_{t=1}^{T} \mathbb{E}[R(A^\star) - R(\hat{A}_t)] \overset{(a)}{=} 2 \sum_{1 \le t \le T, t \text{ odd}} \mathbb{E}\Big[\mathbb{E}_t[R(A^\star) - R(\hat{A}_t)]\Big] \\ &\overset{(b)}{=} 2 \sum_{1 \le t \le T, t \text{ odd}} \mathbb{E}\Big[\sum_{k=k_0+1}^{\infty} \mathbb{E}_t\Big[\Big(R(f_t^k(A_k^\star)) - R(f_t^k(\hat{A}_{t,k}))\Big) \\ &\qquad - \Big(R(f_t^{k-1}(A_{k-1}^\star)) - R(f_t^{k-1}(\hat{A}_{t,k-1}))\Big)\Big]\Big]\end{aligned}$$

where (a) holds since the history of the 2-TS is being updated every two time steps, and (b) follows from the definition of approximate learning.

We then bound the regret differences between two consecutive "approximate learning" via the information gained upon applying the rewards corresponding to two approximate actions. Relating *the latter to the information gained by the 2-TS and applying the chain rule yields the claimed result. Indeed we have the following sequence of inequalities*

$$\begin{aligned}
\mathrm{REG}_T^{\text{2-TS}} &\overset{(c)}{\leq} 2 \sum_{1\leq t\leq T, t \text{ odd}} \sum_{k=k_0+1}^{\infty} \mathbb{E}\Big[\sqrt{\Gamma_{t,k} \cdot \mathrm{I}_t(A_k^\star, A_{k-1}^\star; R(f_t^k(\hat{A}_{t,k})), R(f_t^{k-1}(\hat{A}_{t,k-1})))}\Big] \\
&\overset{(d)}{\leq} 2 \sum_{1\leq t\leq T, t \text{ odd}} \sum_{k=k_0+1}^{\infty} \sqrt{\mathbb{E}[\Gamma_{t,k}] \cdot \mathrm{I}(A_k^\star; R(\hat{A}_t), R(\hat{A}_{t+1})|H^t)} \\
&\overset{(e)}{\leq} 2 \sum_{k=k_0+1}^{\infty} \sqrt{\frac{T}{2} \cdot \bar{\Gamma}_k \cdot \sum_{1\leq t\leq T, t \text{ odd}} \mathrm{I}(A_k^\star; R(\hat{A}_t), R(\hat{A}_{t+1}))|H^t)} \\
&\overset{(f)}{\leq} \sum_{k=k_0+1}^{\infty} \sqrt{2 \cdot \bar{\Gamma}_k \cdot T \cdot \sum_{1\leq t\leq T, t \text{ odd}} \mathrm{I}(A_k^\star; \hat{A}_t, R(\hat{A}_t), \hat{A}_{t+1}, R(\hat{A}_{t+1}))|H^t)} \\
&\overset{(g)}{=} \sum_{k=k_0+1}^{\infty} \sqrt{2 \cdot \bar{\Gamma}_k \cdot T \cdot \mathrm{I}(A_k^\star; H^T)} \\
&\overset{(h)}{\leq} \sum_{k=k_0+1}^{\infty} \sqrt{2 \cdot \bar{\Gamma}_k \cdot T \cdot \mathrm{H}(A_k^\star)}
\end{aligned}$$

where (c) follows from the definition of $\Gamma_{t,k}$ and the data-processing inequality; (d) follows from consecutively using the fact that $A_{k-1}^\star$ is completely determined by $A_k^\star$, then using Proposition 1 (iii), and finally applying Jensen's inequality (e) follows from the definition of $\bar{\Gamma}_k$ and the application of the Cauchy-Schwartz inequality; (f) results from the "more data, more information" property (Yury Polyanskiy, 2022, Proposition 2.3.5); (g) follows from the chain rule for mutual information; and (h) comes from (Yury Polyanskiy, 2022, Proposition 2.4.4) and the fact that $\mathcal{A}_k$ is a finite set.

In the next section, we present the application of Theorem 1 to derive explicit regret bounds for particular settings of bandit problems with structure and show that our bound offers a tight regret rate for the linear bandit problem.

5 Applications to linear bandit problems

In *linear bandits* problems, each action is parameterized by a feature vector, and the associated expected reward can be written as the inner product between the feature vector and the environment parameter. Mathematically, a d-dimensional linear bandit problem is a bandit problem with $\mathcal{A}, \mathcal{O} \subset \mathbb{R}^d$ and such that for all $a \in \mathcal{A}$ and all $\theta \in \mathcal{O}$ we have

$$\mathbb{E}[R(a,\theta)] = \langle a, \theta \rangle,$$

where the expectation is taken over the randomness of the reward function.

Using a similar analysis as Russo and Van Roy (2015), we can bound the chain-link information ratio in linear bandits via the dimension of the action space. The proof is given in Appendix B.2.

Proposition 2 *For d-dimensional linear bandit problems with smooth rewards on the metric space $(\mathcal{A}, \rho)$, for each $t \in \{1, \ldots, T\}$, and each $k > k_0$, we have that*

$$\Gamma_{t,k} \leq 2 \cdot (6 \cdot 2^{-k})^2 \cdot d,$$

where $\Gamma_{t,k}$ is the k^{th}-chain-link information ratio.

Combining Proposition 2 and Theorem 1 leads to the following bound on the *2-TS* regret for linear bandit problems with smooth rewards.

Theorem 2 (Smooth linear bandit) *For d-dimensional linear bandit problems with smooth rewards* *on the metric space* $(\mathcal{A}, \rho)$*, the* 2-TS *expected cumulative regret after* T *steps is bounded by*

$$\mathrm{REG}_T^{\text{2-TS}} \leq 12 \sum_{k=k_0+1}^{\infty} 2^{-k} \sqrt{d \cdot T \cdot \mathrm{H}(A_k^\star)},$$

where $A_k^\star$ *is the* k^{th} *quantization of the optimal action* $A^\star$ *with respect to the metric space* $(\mathcal{A}, \rho)$*, as defined in Definition 4.*

From Theorem 2, we can derive a bound that depends on the *entropy integral*. The proof follows the steps from (van Handel, 2016, Corollary 5.25) and is given in Appendix B.3.

Corollary 1 (Entropy integral) *For a linear bandit of dimension* d*, with smooth rewards on the metric space* $(\mathcal{A}, \rho)$*, the* 2-TS *expected cumulative regret after* T *steps is bounded as*

$$\mathrm{REG}_T^{\text{2-TS}} \leq 24\sqrt{d \cdot T} \int_0^{\infty} \sqrt{\log(|\mathcal{N}(\mathcal{A}, \rho, \varepsilon)|)} d\varepsilon,$$

where $\mathcal{N}(\mathcal{A}, \rho, \varepsilon)$ *is the* ε*-net of smallest cardinality for* $(\mathcal{A}, \rho)$*.*

For linear bandit problems where the possible actions lie in the unit ball, with the help of a covering argument, we come to the following result. The proof is given in Appendix B.4.

Proposition 3 *For* d*-dimensional linear bandits with smooth rewards with respect to* $(\mathcal{A}, ||.||_2)$ *and a ball-structured action space* $\mathcal{A} \subseteq \overline{\mathbf{B}_d(0,1)}$*, where* $\overline{\mathbf{B}_d(0,1)}$ *is the* d*-dimensional closed Euclidean unit ball, the* 2-TS *expected cumulative regret is bounded as*

$$\mathrm{REG}^{\text{2-TS}} \leq 7 \cdot d\sqrt{T}.$$

The remarkable property of the above bound is that it is the first information-theoretic bound on the regret of an algorithm for linear bandits problem that only depends on the dimension d and the square root of the total number of steps T. It improves on the bound $O(d\sqrt{T\log(T)})$ from Dong and Van Roy (2020, Theorem 2) and matches the minimax lower bound $\Omega(d\sqrt{T})$ proven by Dani et al. (2008, Theorem 3) thus suggesting that Two Steps Thompson Sampling is optimal in this context.

6 Conclusion

In this paper, we studied bandit problems with rewards that exhibit some continuity property with respect to the action space. We have introduced a variation of the Thompson Sampling algorithm, named the Two-step Thompson Sampling. The only difference between this algorithm and the original Thompson Sampling algorithm is that the history is updated every two steps. In Theorem 1, we have demonstrated using a chaining argument that the Two Steps Thompson Sampling cumulative expected regret is bounded from above by a measure of the complexity of the action space. For d-dimensional linear bandit problems where the rewards form a subgaussian process with respect to the action space, we obtain a tight regret rate $O(d\sqrt{T})$ that improves upon the best information-theoretic bounds and matches with the minimax lower bound $\Omega(d\sqrt{T})$ (Dani et al., 2008). An interesting future direction is whether we can relate the regret of TS to the 2-TS regret and obtain an optimal regret rate of $O(d\sqrt{T})$ for the original algorithm. Given the new insights that our analysis provides, we conjecture that it should be possible. Future work also includes extending our results to generalized linear bandits and logistic bandit problems.

References

D. Russo and B. Van Roy, "An Information-Theoretic Analysis of Thompson Sampling," Jun. 2015, number: arXiv:1403.5341 arXiv:1403.5341 [cs]. [Online]. Available: http://arxiv.org/abs/1403.5341

S. Dong and B. Van Roy, "An Information-Theoretic Analysis for Thompson Sampling with Many Actions," Jul. 2020, arXiv:1805.11845 [cs, math, stat]. [Online]. Available: http://arxiv.org/abs/1805.11845

W. R. Thompson, “On the likelihood that one unknown probability exceeds another in view of the evidence of two samples,” *Biometrika*, vol. 25, no. 3-4, pp. 285–294, 1933.

D. J. Russo, B. Van Roy, A. Kazerouni, I. Osband, Z. Wen *et al.*, “A tutorial on thompson sampling,” *Foundations and Trends® in Machine Learning*, vol. 11, no. 1, pp. 1–96, 2018.

D. Russo and B. Van Roy, “Learning to Optimize via Information-Directed Sampling,” Jul. 2017, arXiv:1403.5556 [cs]. [Online]. Available: http://arxiv.org/abs/1403.5556

O. Chapelle and L. Li, “An empirical evaluation of Thompson sampling,” *Advances in neural information processing systems*, vol. 24, 2011.

V. Dani, T. P. Hayes, and S. M. Kakade, “Stochastic Linear Optimization under Bandit Feedback,” *21st Annual Conference on Learning Theory*, vol. 21st Annual Conference on Learning Theory, pp. 355–366, 2008.

G. Neu, J. Olkhovskaya, M. Papini, and L. Schwartz, “Lifting the information ratio: An information-theoretic analysis of thompson sampling for contextual bandits,” *arXiv preprint arXiv:2205.13924*, 2022.

A. Gouverneur, B. Rodríguez-Gálvez, T. J. Oechtering, and M. Skoglund, “Thompson Sampling Regret Bounds for Contextual Bandits with sub-Gaussian rewards,” Apr. 2023, arXiv:2304.13593 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2304.13593

J. Negrea, M. Haghifam, G. K. Dziugaite, A. Khisti, and D. M. Roy, “Information-Theoretic Generalization Bounds for SGLD via Data-Dependent Estimates,” *arXiv:1911.02151 [cs, math, stat]*, Jan. 2020, arXiv: 1911.02151. [Online]. Available: http://arxiv.org/abs/1911.02151

Y. W. Yury Polyanskiy, *Information Theory - From Coding to Learning*, 1st ed. Cambridge University Press, Oct. 2022.

R. van Handel, *Probability in High Dimension*. Princeton University, Dec. 2016, vol. APC 550 Lecture Notes.

M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge university press, 2019, vol. 48.

R. M. Gray, *Entropy and Information Theory - First Edition, Corrected*. Information Systems Laboratory Electrical Engineering Department Stanford University: Springer-Verlag, 2013.

A Additional Lemmata

Lemma 1 *Consider a space $\mathcal{A}$, two functions $f : \mathcal{A} \to \mathbb{R}_+$ and $g : \mathcal{A} \to \mathbb{R}_+$, and a probability distribution $\mathbb{Q}$ on $\mathcal{A}$. Then, there exists a pair $(a_1, a_2) \in \mathcal{A}^2$ and a $q \in [0, 1]$ such that*

$$qf(a_1) + (1-q)f(a_2) \leq \int_{a\in\mathcal{A}} f(a)\mathrm{d}\mathbb{Q}(a) \quad \text{and} \quad qg(a_1) + (1-q)g(a_2) \leq \int_{a\in\mathcal{A}} g(a)\mathrm{d}\mathbb{Q}(a).$$

Proof 3 *The proof is inspired by the one from Dong and Van Roy (2020, Lemma 2). However, it contains vital modifications that allow this version of the lemma to work for general spaces $\mathcal{A}$ that are not necessarily finite.*

Let $\bar{F} = \int_{a\in\mathcal{A}} f(a)\mathrm{d}\mathbb{Q}(a)$ and $\bar{G} = \int_{a\in\mathcal{A}} g(a)\mathrm{d}\mathbb{Q}(a)$. Now, consider the spaces $\mathcal{A}_f \coloneqq \{a \in \mathcal{A} : f(a) \leq \bar{F}\}$ and $\mathcal{A}_g \coloneqq \{a \in \mathcal{A} : g(a) \leq \bar{G}\}$. If $\mathcal{A}_f \cap \mathcal{A}_g \neq \emptyset$, then taking both a_1 and a_2 from $\mathcal{A}_f \cap \mathcal{A}_g$ trivially satisfies the conditions for all $q \in [0, 1]$. Therefore, let us assume that the sets are disjoint for the rest of the proof.

Consider some $a_1 \in \mathcal{A}_f = \mathcal{A}_g^c$ and some $a_2 \in \mathcal{A}_g = \mathcal{A}_f^c$. The required condition from the lemma can be rewritten as

$$q \geq \frac{f(a_2) - \bar{F}}{f(a_2) - f(a_1)} \quad \text{and} \quad q \leq \frac{\bar{G} - g(a_2)}{g(a_1) - g(a_2)},$$

where the first inequality took into account that $f(a_1) < f(a_2)$ by the definition of the sets $\mathcal{A}_f$ and $\mathcal{A}_g = \mathcal{A}_f^c$. This inequality can, in turn, be written as

$$\frac{f(a_2) - \bar{F}}{f(a_2) - f(a_1)} \leq \frac{\bar{G} - g(a_2)}{g(a_1) - g(a_2)}$$

which is equivalent to

$$f(a_2)g(a_1) - \bar{F}\big(g(a_1) - g(a_2)\big) \leq \bar{G}\big(f(a_2) - f(a_1)\big) + f(a_1)g(a_2).$$

At this point, we have all the ingredients to prove the statement by contradiction. Assume that there is no pair $(a_1, a_2) \in \mathcal{A}_f \times \mathcal{A}_g$ such that the condition holds, then it must be that

$$f(a_2)g(a_1) - \bar{F}\big(g(a_1) - g(a_2)\big) > \bar{G}\big(f(a_2) - f(a_1)\big) + f(a_1)g(a_2)$$

for every pair $(a_1, a_2) \in \mathcal{A}_f \times \mathcal{A}_g$. Therefore, we can integrate over all such pairs, and the inequality should still hold, namely

$$\begin{aligned} &\int_{\mathcal{A}_f}\int_{\mathcal{A}_g} \Big[f(a_2)g(a_1) - \bar{F}\big(g(a_1) - g(a_2)\big)\Big]\mathrm{d}\mathbb{Q}(a_1)\mathrm{d}\mathbb{Q}(a_2) \\ &\qquad > \int_{\mathcal{A}_f}\int_{\mathcal{A}_g} \Big[\bar{G}\big(f(a_2) - f(a_1)\big) + f(a_1)g(a_2)\Big]\mathrm{d}\mathbb{Q}(a_1)\mathrm{d}\mathbb{Q}(a_2). \end{aligned} \tag{2}$$

We need to introduce some notation to show that (2) cannot happen. Let $F^- \coloneqq \int_{\mathcal{A}_f} f(a)\mathrm{d}\mathbb{Q}(a)$ and $F^+ \coloneqq \int_{\mathcal{A}_g} f(a)\mathrm{d}\mathbb{Q}(a)$ and note that $F^+ + F^- = \bar{F}$. Similarly, $G^- \coloneqq \int_{\mathcal{A}_g} g(a)\mathrm{d}\mathbb{Q}(a)$ and $F^+ \coloneqq \int_{\mathcal{A}_f} g(a)\mathrm{d}\mathbb{Q}(a)$ and $G^+ + G^- = \bar{G}$. Using this notation, we can use Fubini's theorem in (2) and rewrite it as

$$F^+G^+ - (F^+ + F^-)(G^+ - G^-) > (G^+ + G^-)(F^+ - F^-) + F^-G^-,$$

which can be simplified to

$$F^-G^- > F^+G^+$$

and which is impossible by the definition of F^-, F^+, G^+ and G^-, completing the contradiction and therefore the proof.

Lemma 2 ((van Handel, 2016, Lemma 5.13)) *Let $\overline{\mathbf{B}_d(0,1)}$ denote the d-dimensional closed Euclidean unit ball. We have $|\mathcal{N}(\overline{\mathbf{B}_d(0,1)}, ||\cdot||_2, \varepsilon) = 1$ for $\varepsilon \geq 1$ and*

$$\left(\frac{1}{\varepsilon}\right)^d \leq |\mathcal{N}(\overline{\mathbf{B}_d(0,1)}, ||\cdot||_2, \varepsilon)| \leq \left(1 + \frac{2}{\varepsilon}\right)^d \qquad \text{for } 0 < \varepsilon < 1.$$

B Proofs

B.1 Proof of Proposition 1

For each time step $t \in \{1, \ldots, T\}$, we will construct the sequence of function $\{f_t^k\}_{k=k_0}^{\infty}$ by induction and instead of constructing a sequence that satisfies directly (iii), we will design it such that for each $k > k_0$, it satisfies simultaneously the two following equations:

$$\mathrm{I}_t\big(A_k^\star; R(f_t^k(\hat{A}_{t,k})), R(f_t^{k-1}(\hat{A}_{t,k-1}))\big) \leq \mathrm{I}_t\big(A_k^\star; R(\hat{A}_t), R(f_t^{k-1}(\hat{A}_{t,k-1}))\big) \text{ and} \tag{3}$$

$$\mathrm{I}_t\big(A_{k+1}^\star; R(\hat{A}_t), R(f_t^k(\hat{A}_{t,k}))\big) \leq \mathrm{I}_t\big(A_{k+1}^\star; R(\hat{A}_t), R(\hat{A}_t')\big), \tag{4}$$

thus ensuring that f_t^k satisfies (iii).

First, we start by showing that there exists a function $f_t^{k_0}$ that satisfies requirement (i) and equation (4). By definition of k_0, we have that the cardinality of $\mathcal{A}_{k_0}$ is 1, that is $\mathcal{A}_{k_0} = \{a_0\}$ for some $a_0 \in \mathcal{A}$ and, as $A_{k_0}^\star \in \mathcal{A}_{k_0}$ and $\hat{A}_{t,k_0} \in \mathcal{A}_{k_0}$, we have $A_{k_0}^\star = \hat{A}_{t,k_0} = a_0$, thus satisfying requirement (i). Setting the random function $f_t^{k_0}$ to have the same conditional probability distribution as $\mathbb{P}_{A^\star|H^t}$ ensures equation (4) is satisfied.

Now, we assume that for each $k \in \{k_0, \ldots, K-1\}$, we have constructed a function f_t^k that satisfied (3) and (4). We then want to show that we can construct a random function f_t^K that also satisfies (3) and (4).

First, for each $a_{K,i} \in \mathcal{A}_K$ with $i \in \{1, \ldots, |\mathcal{A}_K|\}$, we define $\mathcal{A}_{K,i} = \{a \in \mathcal{A} : \pi_K(a) = a_{K,i}\}$ as the set of actions in $\mathcal{A}$ that are mapped to $a_{K,i}$ by the mapping π_K associated to $\mathcal{A}_K$, that is formally. In this way, for each $a_{K,i} \in \mathcal{A}_K$, we can write

$$\begin{aligned}
&\mathrm{I}_t\big(A_K^\star; R(\hat{A}_t), R(f_t^{K-1}(\hat{A}_{t,K}))|\hat{A}_t \in \mathcal{A}_{K,i}\big) \\
&= \sum_{a \in \mathcal{A}_{K,i}} \mathbb{P}_t[\hat{A}_t = a|\hat{A}_t \in \mathcal{A}_{K,i}]\mathrm{I}_t\big(A_K^\star; R(a), R(f_t^{K-1}(\hat{A}_{t,K}))|\hat{A}_t \in \mathcal{A}_{K,i}\big) \\
&= \sum_{a \in \mathcal{A}_{K,i}} \mathbb{P}_t[\hat{A}_t = a|\hat{A}_t \in \mathcal{A}_{K,i}]\mathrm{I}_t\big(A_K^\star; R(a), R(f_t^{K-1}(\hat{A}_{t,K}))\big)
\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{I}_t\big(A_{K+1}^\star; R(\hat{A}_t), R(\hat{A}_t')|\hat{A}_t \in \mathcal{A}_{K,i}\big) \\
&= \sum_{a \in \mathcal{A}_{K,i}} \mathbb{P}_t[\hat{A}_t = a|\hat{A}_t \in \mathcal{A}_{K,i}]\mathrm{I}_t\big(A_{K+1}^\star; R(a), R(\hat{A}_t')|\hat{A}_t \in \mathcal{A}_{K,i}\big) \\
&= \sum_{a \in \mathcal{A}_{K,i}} \mathbb{P}_t[\hat{A}_t = a|\hat{A}_t \in \mathcal{A}_{K,i}]\mathrm{I}_t\big(A_{K+1}^\star; R(a), R(\hat{A}_t')\big),
\end{aligned}$$

where we used the fact that $A_K^\star$ and $A_{K+1}^\star$ are independent of $\hat{A}_t$ when conditioned on H^t.

Applying Lemma 1, for each step $t \in \{1, \ldots, T\}$ and each $a_{K,i} \in \mathcal{A}_K$, there exist two actions $a_{K,i}^{t,1}, a_{K,i}^{t,2} \in \mathcal{A}_{K,i}$ and a value $p_{K,i}^t \in [0,1]$, such that:

$$\begin{aligned}
\mathrm{I}_t\big(A_K^\star;&R(\hat{A}_t), R(f_t^{K-1}(\hat{A}_{t,K}))|\hat{A}_t \in \mathcal{A}_{K,i}\big) \\
&\geq p_{K,i}^t \mathrm{I}_t(A_K^\star; R(a_{K,i}^{t,1}), R(f_t^{K-1}(\hat{A}_{t,K}))\big) + (1 - p_{K,i}^t)\mathrm{I}_t(A_K^\star; R(a_{K,i}^{t,2}), R(f_t^{K-1}(\hat{A}_{t,K}))\big)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{I}_t\big(A_{K+1}^\star;&R(\hat{A}_t), R(\hat{A}_t')|\hat{A}_t \in \mathcal{A}_{K,i}\big) \\
&\geq p_{K,i}^t \mathrm{I}_t(A_K^\star; R(a_{K,i}^{t,1}), R(\hat{A}_t')\big) + (1 - p_{K,i}^t)\mathrm{I}_t(A_K^\star; R(a_{K,i}^{t,2}), R(\hat{A}_t')\big).
\end{aligned}$$

For $a \in \mathcal{A}_{K,i}$, we define the random function $f_t^K(a)$ such that it outputs $a_{K,i}^{t,1} \in \mathcal{A}_{K,i}$ with probability $p_{K,i}^t$ and $a_{K,i}^{t,2} \in \mathcal{A}_{K,i}$ with probability $1 - p_{K,i}^t$. We observe that for $a \in \mathcal{A}_{K,i}$,

$\pi_K(a) = \pi_k(f_t^K(a)) = a_{K,i}$ as both a and $f_t^K(a)$ belong to $\mathcal{A}_{K,i}$. Then, the distance $\rho(a, f_t^k(a))$ is bounded by 2^{-K}. We repeat this procedure for all $a_{K,i} \in \mathcal{A}_K$ and their corresponding $\mathcal{A}_{K,i}$ to define $f_t^K(a)$ for all $a \in \mathcal{A}$ and it holds by that, for all $a \in \mathcal{A}$, $\rho(f_t^K(a), a) \leq 2^{-K}$.

We can verify that

$$\begin{aligned}
&\mathrm{I}_t\big(A_K^\star; R(f_t^K(\hat{A}_{t,K})), R(f_t^{K-1}(\hat{A}_{t,K-1}))\big) \\
&= \sum_{a_{K,i}\in\mathcal{A}_K} \sum_{j=1,2} \mathbb{P}_t[f_t^K(\hat{A}_{t,K})) = a_{K,i}^{t,j} | \hat{A}_t \in \mathcal{A}_{K,i}] \cdot \mathbb{P}_t[\hat{A}_t \in \mathcal{A}_{K,i}] \cdot \mathrm{I}_t\big(A_K^\star; R(a_{K,i}^{t,j}), R(f_t^{K-1}(\hat{A}_{t,K-1}))\big) \\
&= \sum_{a_{K,i}\in\mathcal{A}_K} \mathbb{P}_t[\hat{A}_t \in \mathcal{A}_{K,i}](p_{K,i}^t \cdot \mathrm{I}_t\big(A_K^\star; R(a_{K,i}^{t,1}), R(f_t^{K-1}(\hat{A}_{t,K-1}))\big) \\
&\quad + (1 - p_{K,i}^t) \cdot \mathrm{I}_t\big(A_K^\star; R(a_{K,i}^{t,2}), R(f_t^{K-1}(\hat{A}_{t,K-1}))\big) \\
&\leq \sum_{a_{K,i}\in\mathcal{A}_K} \mathbb{P}_t[\hat{A}_t \in \mathcal{A}_{K,i}] \mathrm{I}_t\big(A_K^\star; R(\hat{A}_t), R(f_t^{K-1}(\hat{A}_{t,K-1})) | \hat{A}_t \in \mathcal{A}_{K,i}\big) \\
&= \mathrm{I}_t\big(A_K^\star; R(\hat{A}_t), R(f_t^{K-1}(\hat{A}_{t,K-1}))\big)
\end{aligned}$$

and similarly that

$$\begin{aligned}
&\mathrm{I}_t\big(A_{K+1}^\star; R(f_t^K(\hat{A}_{t,K})), R(\hat{A}_t')\big) \\
&= \sum_{a_{K,i}\in\mathcal{A}_K} \sum_{j=1,2} \mathbb{P}_t[f_t^K(\hat{A}_{t,K})) = a_{K,i}^{t,j} | \hat{A}_t \in \mathcal{A}_{K,i}] \cdot \mathbb{P}_t[\hat{A}_t \in \mathcal{A}_{K,i}] \cdot \mathrm{I}_t\big(A_{K+1}^\star; R(a_{K,i}^{t,j}), R(\hat{A}_t')\big) \\
&= \sum_{a_{K,i}\in\mathcal{A}_K} \mathbb{P}_t[\hat{A}_t \in \mathcal{A}_{K,i}](p_{K,i}^t \cdot \mathrm{I}_t\big(A_{K+1}^\star R(a_{K,i}^{t,1}), R(\hat{A}_t')\big) + (1 - p_{K,i}^t) \cdot \mathrm{I}_t\big(A_{K+1}^\star; R(a_{K,i}^{t,2}), R(\hat{A}_t')\big) \\
&\leq \sum_{a_{K,i}\in\mathcal{A}_K} \mathbb{P}_t[\hat{A}_t \in \mathcal{A}_{K,i}] \mathrm{I}_t\big(A_{K+1}^\star; R(\hat{A}_t), R(\hat{A}_t') | \hat{A}_t \in \mathcal{A}_{K,i}\big) \\
&= \mathrm{I}_t\big(A_{K+1}^\star; R(\hat{A}_t), R(\hat{A}_t')\big)
\end{aligned}$$

where the inequalities follow from the construction of f_t^K. Thus f_t^K satisfies requirement (iii). As the result holds already for $k = k_0$, by induction, we extend this result for all $k \geq k_0$.

We note that by construction, for each step $t \in \{1, \ldots, T\}$ and for each $k \geq k_0$, we have that

$$\rho(f_t^k(A_k^\star), A^\star) \leq \rho(f_t^k(A_k^\star), A_k^\star) + \rho(A_k^\star, A^\star) \leq 2 \cdot 2^{-k}, \tag{5}$$

$$\rho(f_t^k(\hat{A}_{t,k}), \hat{A}_t) \leq \rho(f_t^k(\hat{A}_{t,k}), \hat{A}_{t,k}) + \rho(\hat{A}_{t,k}, \hat{A}_t) \leq 2 \cdot 2^{-k}, \tag{6}$$

where we use the triangle inequality together with the definition of f_t^k and of $A_k^\star$ and $\hat{A}_{t,k}$.

Lastly, we have to verify that at each period $t \in \{1, \ldots, T\}$, the regret of the "*approximate learning*" asymptotically converges to the regret of Two Steps Thompson Sampling regret for finer approximations.

Using the fact that by construction of f_t^k, we have for all $a \in \mathcal{A}_k$ that $\pi_k(f_t^k(a)) = a$ and that by definition $A_k^\star = \pi_k(A^\star)$, we can write:

$$\begin{aligned}
\mathbb{E}_t[R(f_t^k(A_k^\star)) - R(A^\star)] &= \mathbb{E}_t[R(f_t^k(A_k^\star)) - R(A_k^\star)] + \mathbb{E}_t[R(A_k^\star) - R(A^\star)] \\
&= \mathbb{E}_t[R(f_t^k(A_k^\star)) - R(\pi_k(f_t^k(A_k^\star)))] + \mathbb{E}_t[R(\pi_k(A^\star)) - R(A^\star)] \\
&\leq 2 \cdot \mathbb{E}_t[\sup_{a\in\mathcal{A}} R(\pi_k(a)) - R(a)].
\end{aligned}$$

Since the process is separable, there is a countable set $\mathcal{A}' \subseteq \mathcal{A}$ such that $\sup_{a\in\mathcal{A}} R(a) = \sup_{a\in\mathcal{A}'} R(a)$ almost surely. Recall from Definition 4 that $\mathcal{A}_k$ is the 2^{-k} net with mapping $\pi_k : \mathcal{A} \to \mathcal{A}_k$ such that $\mathcal{A}_k \subseteq \mathcal{A}_{k+1}$ for all k. Then, by the monotone convergence theorem

$$\mathbb{E}\left[\sup_{a\in\mathcal{A}} R(a)\right] = \mathbb{E}\left[\sup_{a\in\mathcal{A}'} R(a)\right] = \sup_{k\geq k_0} \mathbb{E}\left[\sup_{a\in\mathcal{A}_k} R(a)\right] = \lim_{k\to\infty} \mathbb{E}\left[\sup_{a\in\mathcal{A}} R(\pi_k(a))\right],$$

which implies

$$\lim_{k\to\infty}\mathbb{E}_t[R(f_t^k(A_k^\star))] = \mathbb{E}_t[R(A^\star)].$$

A similar analysis can be applied to $\mathbb{E}_t[R(f_t^k(\hat{A}_{t,k})) - R(\hat{A}_t)]$ and leads to

$$\lim_{k\to\infty}\mathbb{E}_t[R(f_t^k(A_k^\star)) - R(f_t^k(\hat{A}_{t,k}))] = \mathbb{E}_t[R(A^\star) - R(\hat{A}_t)].$$

B.2 Proof of Proposition 2

We start the proof by recalling the definition of $\Gamma_{t,k}$ as

$$\Gamma_{t,k} = \frac{\mathbb{E}_t\Big[\Big(R(f_t^k(A_k^\star)) - R(f_t^k(\hat{A}_{t,k}))\Big) - \Big(R(f_t^{k-1}(A_{k-1}^\star)) - R(f_t^{k-1}(\hat{A}_{t,k-1}))\Big)\Big]^2}{\mathrm{I}_t(f_t^k(A_k^\star), f_t^{k-1}(A_{k-1}^\star); R(f_t^k(\hat{A}_{t,k})), R(f_t^{k-1}(\hat{A}_{t,k-1})))}$$

where $A_k^\star$ and $\hat{A}_{t,k}$ are the k^{th}-quantizations respectively of the optimal action $A^\star$ and the sampled action $\hat{A}_t$. We recall from the proof of Proposition 1 that the definition of $f_t^k(A)$ implies that for all $a_{k,m}\in\mathcal{A}_k$ there exist a pair of actions $a_{k,m}^{t,1}, a_{k,m}^{t,2}\in\mathcal{A}_{k,m}$ such that

$$\mathbb{P}_t[f_t^k(A) = a_{k,m}^{t,1}|A\in\mathcal{A}_{k,m}] = p_{k,m}^t, \quad \mathbb{P}_t[f_t^k(A) = a_{k,m}^{t,2}|A\in\mathcal{A}_{k,m}] = 1 - p_{k,m}^t.$$

For the sake of brevity, we define the notation

$$\begin{aligned}\mathbb{Q}_t[a_{k-1,m}, a_{k,l}, i, i'] :=& \mathbb{P}_t[f_t^{k-1}(A_{k-1}^\star) = a_{k-1,m}^{t,i}|A_{k-1}^\star\in\mathcal{A}_{k,m}]\\ &\cdot\mathbb{P}_t[f_t^k(A_k^\star) = a_{k,l}^{t,i'}|A_k^\star\in\mathcal{A}_{k,l}]\\ &\cdot\mathbb{P}_t[A_k^\star\in\mathcal{A}_{k,l}, A_{k-1}^\star\in\mathcal{A}_{k-1,m}]\end{aligned}$$

and use the notation $\{(a_{k-1,\delta_n}, a_{k,\gamma_n}, i_{\mu_n}, i'_{\nu_n})\}_{n=1}^{N_k}$ to represent the sequence of all quadruplets $\{a_{k-1}, a_k, i, i'\}$ such that $a_{k-1}\in\mathcal{A}_{k-1}, a_k\in\mathcal{A}_k, i\in\{1,2\}, i'\in\{1,2\}$ and $\pi_{k-1}(a_k) = a_{k-1}$, where N_k is the number of such quadruplets.

We will first focus on

$$\mathbb{E}_t\left[\big(R(f_t^k(A_k^\star)) - R(f_t^{k-1}(A_{k-1}^\star))\big) - \big(R(f_t^k(\hat{A}_{t,k})) - R(f_t^{k-1}(\hat{A}_{t,k-1}))\big)\right]$$

and note that we can relate it to the trace of a random matrix. Indeed, using the previously introduced notations, we can write this expectation as

$$\begin{aligned}&\sum_{n=1}^{N_k}\mathbb{Q}_t[a_{k-1,\delta_n}, a_{k,\gamma_n}, i_{\mu_n}, i'_{\nu_n}]\\ &\cdot\big(\mathbb{E}_t[R(a_{k,\gamma_n}^{t,i_{\mu_n}}) - R(a_{k-1,\delta_n}^{t,i'_{\nu_n}})|f_t^k(A_k^\star) = a_{k,\gamma_n}^{t,i_{\mu_n}}, f_t^{k-1}(A_{k-1}^\star) = a_{k-1,\delta_n}^{t,i'_{\nu_n}}] - \mathbb{E}_t[R(a_{k,\gamma_n}^{t,i_{\mu_n}}) - R(a_{k-1,\delta_n}^{t,i'_{\nu_n}})]\big).\end{aligned}$$

Therefore, for any round $t\in\{1,\ldots,T\}$, conditioned on the history $\hat{H}^t$, we can define a random matrix $M^{k,t}\in\mathbb{R}^{N_k\times N_k}$ by specifying the entry $M_{p,q}^{k,t}$ to be equal to

$$\begin{aligned}&\sqrt{\mathbb{Q}_t[a_{k-1,\delta_p}, a_{k,\gamma_p}, i_{\mu_p}, i'_{\nu_p}]}\sqrt{\mathbb{Q}_t[a_{k-1,\delta_q}, a_{k,\gamma_q}, i_{\mu_q}, i'_{\nu_q}]}\\ &\big(\mathbb{E}_t[R(a_{k,\gamma_q}^{t,i_{\mu_q}}) - R(a_{k-1,\delta_q}^{t,i'_{\nu_q}})|f_t^k(A_k^\star) = a_{k,\gamma_p}^{t,i_{\mu_p}}, f_t^{k-1}(A_{k-1}^\star) = a_{k-1,\delta_p}^{t,i'_{\nu_p}}] - \mathbb{E}_t[R(a_{k,\gamma_q}^{t,i_{\mu_q}}) - R(a_{k-1,\delta_q}^{t,i'_{\nu_q}})]\big)\end{aligned}$$

for all $p,q = 1,\ldots,N_k$. In this way, the trace of the matrix $M^{k,t}$ is equal to the desired expectation, namely

$$\mathrm{Tr}(M^{k,t}) = \mathbb{E}_t\left[(R(f_t^k(A_k^\star)) - R(f_t^{k-1}(A_{k-1}^\star))) - (R(f_t^k(\hat{A}_{t,k})) - R(f_t^{k-1}(\hat{A}_{t,k-1})))\right].$$

Here, we can note that $R(f_t^k(A_k^\star)) - R(f_t^{k-1}(A_{k-1}^\star))$ is $(6 \cdot 2^{-k})^2$-sub-Gaussian. Indeed, by construction, of $f_t^k(A_k^\star)$ and $f_t^{k-1}(A_{k-1}^\star)$, we had showed in (5) that $\rho(f_t^k(A_k^\star), A^\star) \leq 2 \cdot 2^{-k}$ and $\rho(f_t^{k-1}(A_{k-1}^\star), A^\star) \leq 2 \cdot 2^{-(k-1)}$. Then, by using the triangle inequality, we have that

$$\rho(f_t^k(A_k^\star), f_t^{k-1}(A_{k-1}^\star)) \leq \rho(f_t^k(A_k^\star), A^\star) + \rho(A^\star, f_t^{k-1}(A_{k-1}^\star)) \leq 2 \cdot 2^{-k} + 2 \cdot 2^{-(k-1)} = 6 \cdot 2^{-k}.$$

Similarly, we can show that $R(f_t^k(\hat{A}_{t,k})) - R(f_t^{k-1}(\hat{A}_{t,k-1}))$ is also $(6 \cdot 2^{-k})^2$-sub-Gaussian.

In the same fashion as in (Russo and Van Roy, 2015, Proposition 5), we relate the mutual information

$$\mathrm{I}_t(f_t^k(A_k^\star), f_t^{k-1}(A_{k-1}^\star); R(f_t^k(\hat{A}_{t,k})), R(f_t^{k-1}(\hat{A}_{t,k-1})))$$

to the squared Frobenius norm of $M^{k,t}$ as:

$$\begin{aligned}
&\mathrm{I}_t(f_t^k(A_k^\star), f_t^{k-1}(A_{k-1}^\star); R(f_t^k(\hat{A}_{t,k})), R(f_t^{k-1}(\hat{A}_{t,k-1}))) \\
&\geq \mathrm{I}_t(f_t^k(A_k^\star), f_t^{k-1}(A_{k-1}^\star); R(f_t^k(\hat{A}_{t,k})) - R(f_t^{k-1}(\hat{A}_{t,k-1}))) \\
&= \sum_{p=1}^{N^k} \sum_{q=1}^{N^k} \mathbb{Q}_t[a_{k-1,\delta_p}, a_{k,\gamma_p}, i_{\mu_p}, i'_{\nu_p}] \mathbb{Q}_t[a_{k-1,\delta_q}, a_{k,\gamma_q}, i_{\mu_q}, i'_{\nu_q}] \\
&\cdot \mathrm{D}_{\mathrm{KL}}(\mathbb{P}_{R(a_{k,\gamma_q}^{t,i_{\mu_q}}) - R(a_{k-1,\delta_q}^{t,i'_{\nu_q}}) | \hat{H}^t, f_t^k(A_k^\star) = a_{k,\gamma_p}^{t,i_{\mu_p}}, f_t^{k-1}(A_{k-1}^\star) = a_{k-1,\delta_p}^{t,i'_{\nu_p}}} || \mathbb{P}_{R(a_{k,\gamma_q}^{t,i_{\mu_q}}) - R(a_{k-1,\delta_q}^{t,i'_{\nu_q}}) | \hat{H}^t}) \\
&\geq \sum_{p=1}^{N^k} \sum_{q=1}^{N^k} \mathbb{Q}_t[a_{k-1,\delta_p}, a_{k,\gamma_p}, i_{\mu_p}, i'_{\nu_p}] \mathbb{Q}_t[a_{k-1,\delta_q}, a_{k,\gamma_q}, i_{\mu_q}, i'_{\nu_q}] \cdot \frac{1}{2 \cdot (6 \cdot 2^{-k})^2} \\
&\cdot \big(\mathbb{E}_t[R(a_{k,\gamma_q}^{t,i_{\mu_q}}) - R(a_{k-1,\delta_q}^{t,i'_{\nu_q}}) | f_t^k(A_k^\star) = a_{k,\gamma_p}^{t,i_{\mu_p}}, f_t^{k-1}(A_{k-1}^\star) = a_{k-1,\delta_p}^{t,i'_{\nu_p}}] - \mathbb{E}_t[R(a_{k,\gamma_q}^{t,i_{\mu_q}}) - R(a_{k-1,\delta_q}^{t,i'_{\nu_q}})]\big)^2 \\
&= \frac{1}{2(6 \cdot 2^{-k})^2} ||M^{k,t}||_F^2
\end{aligned}$$

where the last inequality is obtained again using the Donsker–Varadhan inequality (Gray, 2013, Theorem 5.2.1) as in (Russo and Van Roy, 2015, Lemma 3).

Combining the last two equations and using the inequality $\mathrm{trace}(M) \leq \sqrt{\mathrm{rank}(M)} ||M||_F$ (Russo and Van Roy, 2015, Fact 10), it comes that

$$\Gamma_{t,k} \leq 2(6 \cdot 2^{-k})^2 \frac{\mathrm{Trace}(M^{k,t})^2}{||M^{k,t}||_F^2} \leq 2(6 \cdot 2^{-k})^2 \cdot \mathrm{rank}(M^{k,t}) \text{ a.s..}$$

We conclude the proof by showing that the rank of the matrix $M^{k,t}$ is upper bounded by d.

For the sake of brevity, we define $\Theta_t := \mathbb{E}_t[\Theta]$ and for $n = 1, \ldots, N_k$, we define $\mathbb{Q}_{n,t} = \mathbb{Q}_t[a_{k-1,\delta_n}, a_{k,\gamma_n}, i_{\mu_n}, i'_{\nu_n}]$ and $\Theta_{n,t} = \mathbb{E}_t[\Theta | f_t^k(A_k^\star) = a_{k,\gamma_n}^{t,i_{\mu_n}}, f_t^{k-1}(A_{k-1}^\star) = a_{k-1,\delta_n}^{t,i'_{\nu_n}}]$.

We then have

$$\mathbb{E}_t\left[R(a_{k,\gamma_q}^{t,i_{\mu_q}}) - R(a_{k-1,\delta_q}^{t,i'_{\nu_q}})\right] = \mathbb{E}_t\left[\langle a_{k,\gamma_q}^{t,i_{\mu_q}}, \Theta\rangle - \langle a_{k-1,\delta_q}^{t,i'_{\nu_q}}, \Theta\rangle\right] = \langle a_{k,\gamma_q}^{t,i_{\mu_q}} - a_{k-1,\delta_q}^{t,i'_{\nu_q}}, \Theta_t\rangle$$

and

$$\begin{aligned}
&\mathbb{E}_t\left[R(a_{k,\gamma_q}^{t,i_{\mu_q}}) - R(a_{k-1,\delta_q}^{t,i'_{\nu_q}}) | f_t^k(A_k^\star) = a_{k,\gamma_p}^{t,i_{\mu_p}}, f_t^{k-1}(A_{k-1}^\star) = a_{k-1,\delta_p}^{t,i'_{\nu_p}}\right] \\
&= \mathbb{E}_t\left[\langle a_{k,\gamma_q}^{t,i_{\mu_q}}, \Theta\rangle - \langle a_{k-1,\delta_q}^{t,i'_{\nu_q}}, \Theta\rangle | f_t^k(A_k^\star) = a_{k,\gamma_p}^{t,i_{\mu_p}}, f_t^{k-1}(A_{k-1}^\star) = a_{k-1,\delta_p}^{t,i'_{\nu_p}}\right] \\
&= \langle a_{k,\gamma_q}^{t,i_{\mu_q}} - a_{k-1,\delta_q}^{t,i'_{\nu_q}}, \Theta_{p,t}\rangle
\end{aligned}$$

Since the inner product is linear, we can rewrite each entry $M_{p,q}^{k,t}$ of the matrix $M^{k,t}$ as

$$\sqrt{\mathbb{Q}_{p,t}\mathbb{Q}_{q,t}} \langle a_{k,\gamma_q}^{t,i_{\mu_q}} - a_{k-1,\delta_q}^{t,i'_{\nu_q}}, \Theta_{p,t} - \Theta_t\rangle.$$

Equivalently, the matrix $M^{k,t}$ can be written as

$$\begin{bmatrix} \sqrt{\mathbb{Q}_{1,t}}(\Theta_{1,t} - \Theta_t) \\ \vdots \\ \sqrt{\mathbb{Q}_{N_k,t}}(\Theta_{N_k,t} - \Theta_t) \end{bmatrix} \begin{bmatrix} \sqrt{\mathbb{Q}_{1,t}}(a^{t,i_{\mu_1}}_{k,\gamma_1} - a^{t,i'_{\nu_1}}_{k-1,\delta_1}) & \cdots & \sqrt{\mathbb{Q}_{N_k,t}}(a^{t,i_{\mu_{N_k}}}_{k,\gamma_{N_k}} - a^{t,i'_{\nu_{N_k}}}_{k-1,\delta_{N_k}}) \end{bmatrix}.$$

This rewriting highlights that $M^{k,t}$ can be written as the product of a N_k by d matrix and a d by N_k matrix and therefore has a rank lower or equal than $\min(d, N_k)$.

For completeness, we can write that the chain-link information ratio is upper bounded by $\Gamma_{t,k} \leq 2 \cdot \rho_k^2 \cdot d$ where ρ_k is an upper bound on $\rho(f_t^k(A_k^\star), f_t^{k-1}(A_{k-1}^\star))$. This remark will be of use in the proof of Proposition 3.

B.3 Proof of Corollary 1

Bounding the entropy of $A_k^\star$ by the cardinality of set $\mathcal{A}_k$, we have that

$$\sum_{k=k_0+1}^{\infty} 2^{-k}\sqrt{\mathrm{H}(A_k^\star)} \leq \sum_{k=k_0+1}^{\infty} 2^{-k}\sqrt{\log(|\mathcal{N}(\mathcal{A}, \rho, 2^{-k})|)}.$$

By definition of the ε-net, $|\mathcal{N}(\mathcal{A}, \rho, \varepsilon)|$ is decreasing in ε. It then comes that

$$\begin{aligned}
\sum_{k=k_0+1}^{\infty} 2^{-k}\sqrt{\log(|\mathcal{N}(\mathcal{A}, \rho, 2^{-k})|)} &= 2\sum_{k=k_0+1}^{\infty}\int_{2^{-k}}^{2^{-k-1}} \sqrt{\log(|\mathcal{N}(\mathcal{A}, \rho, 2^{-k})|)}\, d\varepsilon \\
&\leq 2\sum_{k=k_0+1}^{\infty}\int_{2^{-k}}^{2^{-k-1}} \sqrt{\log(|\mathcal{N}(\mathcal{A}, \rho, \varepsilon)|)}\, d\varepsilon \\
&= 2\int_{0}^{\mathrm{diam}(\mathcal{A})} \sqrt{\log(|\mathcal{N}(\mathcal{A}, \rho, \varepsilon)|)}\, d\varepsilon. \\
&= 2\int_{0}^{\infty} \sqrt{\log(|\mathcal{N}(\mathcal{A}, \rho, \varepsilon)|)}\, d\varepsilon,
\end{aligned}$$

where the last equality comes from the fact that $\mathcal{N}(\mathcal{A}, \rho, \varepsilon)$ is a singleton for every $\varepsilon > \mathrm{diam}(\mathcal{A})$.

Using this fact together with Theorem 1 yields the desired result.

B.4 Proof of Proposition 3

At the end of the proof of Proposition 2, we have shown that the chain-link information ratio was in general bounded by $\Gamma_{t,k} \leq 2 \cdot \rho_k^2 \cdot d$ where ρ_k is an upper bound on $\rho(f_t^k(A_k^\star), f_t^{k-1}(A_{k-1}^\star))$ and proved that by definition of the quantization, $A_k^\star$ and the sampling functions f_t^k, it holds that

$$\rho(f_t^k(A_k^\star), f_t^{k-1}(A_{k-1}^\star)) \leq 2 \cdot 2^{-k} + 2 \cdot 2^{-(k-1)}.$$

We can reflect that the choice of using 2^{-k}-nets to define our sequence of quantizations $\{A_k^\star\}_{k=k_0+1}^{\infty}$ was arbitrary. In general, we could have considered a α^{-k}-net for some $\alpha > 1$. Adapting the bound on ρ_k and to that reflection leads to the following bound:

$$\rho(f_t^k(A_k^\star), f_t^{k-1}(A_{k-1}^\star)) \leq 2 \cdot \alpha^{-k} + 2 \cdot \alpha^{-(k-1)}.$$

Combining this result with Theorem 1, we get that

$$\mathrm{REG}_T^{\text{2-TS}} \leq 2\sum_{k=k_0+1}^{\infty} \sqrt{2 \cdot \rho_k^2 \cdot d \cdot T \cdot \log(|\mathcal{N}(\mathcal{A}, \rho, \alpha^{-k})|)},$$

where we upper bounded the entropy of $A_k^\star$ by the logarithm of the cardinality of the set $\mathcal{A}_k$.

Applying Lemma 2 to the upper bound the cardinality of the smallest α^{-k}-net $\mathcal{N}(\mathcal{A}, \rho, \alpha^{-k})$ and rearranging the terms, we get the following bound:

$$\mathrm{REG}_T^{\text{2-TS}} \leq 2 \cdot d \cdot \sqrt{T} \sum_{k=k_0+1}^{\infty} \sqrt{2 \cdot \rho_k^2 \cdot \log(2 \cdot \alpha^k + 1)}.$$

Now, we note that for linear bandit problems, we can define the first quantization set $\mathcal{A}_{k_0}$ to be the center of the ball, that is $\mathcal{A}_{k_0} = \{0_d\}$ where 0_d is the d-dimensional zero and chose $f_t^{k_0}(0_d) = 0_d$. It is easy to verify that this choice satisfies Proposition 1 (i) as $A_{k_0}^\star = \hat{A}_{t,k_0} = 0_d$ and $f_t^{k_0}(A_{k_0}^\star) = f_t^{k_0}(\hat{A}_{t,k_0}) = 0_d$, as well as fulfills 4 as $R(f_t^{k_0}(\hat{A}_{t,k_0})) = R(0_d)$ does not depend on the Θ and therefore is independent of $A^\star$ and $A_{k_0+1}^\star$.

Observing that in the unit ball, by definition the radius is 1, we first note that $\mathcal{A}_{k_0}$ is a (α^0)-net for $\mathcal{A}$, implying $k_0 = 0$ and secondly that $\rho(f_t^{k_0+1}(A_{k_0+1}^\star), f_t^{k_0}(A_{k_0}^\star)) = \rho(f_t^{k_0+1}(A_{k_0+1}^\star), 0_d) \leq 1$ and therefore we can use $\rho_{k_0+1} = 1$ which is a better upper bound than $2 \cdot \alpha^{-(k_0+1)} + 2 \cdot \alpha^{-k_0} = 2 \cdot (1 + \alpha^{-1})$.

Applying those results, we obtain the following bound:

$$\mathrm{REG}_T^{\text{2-TS}} \leq d\sqrt{T} \cdot 2 \cdot \left(\sqrt{2 \cdot \log(2\alpha + 1)} + \sum_{k=2}^{\infty} (2 \cdot \alpha^{-k} + 2 \cdot \alpha^{-(k-1)}) \sqrt{2 \cdot \log(2\alpha^k + 1)} \right).$$

For instance, choosing $\alpha = 20$, we have that

$$2 \cdot \left(\sqrt{2 \cdot \log(2\alpha + 1)} + \sum_{k=2}^{\infty} (2 \cdot \alpha^{-k} + 2 \cdot \alpha^{-(k-1)}) \sqrt{2 \cdot \log(2\alpha^k + 1)} \right) \approx 6.27.$$

Finally, rounding up this value leads to the claimed result.