

---

# AN INFORMATION-THEORETIC ANALYSIS OF THOMPSON SAMPLING FOR LOGISTIC BANDITS

---

A PREPRINT

**Amaury Gouverneur, Borja Rodríguez-Gálvez, Tobias J. Oechtering, and Mikael Skoglund**  
KTH Royal Institute of Technology  
Stockholm, Sweden

## ABSTRACT

We study the performance of the Thompson Sampling algorithm for logistic bandit problems. In this setting, an agent receives binary rewards with probabilities determined by a logistic function,  $\exp(\beta\langle a, \theta \rangle) / (1 + \exp(\beta\langle a, \theta \rangle))$ , with slope parameter  $\beta > 0$ , and where both the action  $a \in \mathcal{A}$  and parameter  $\theta \in \mathcal{O}$  lie within the  $d$ -dimensional unit ball. Adopting the information-theoretic framework introduced by Russo and Van Roy (2016), we analyze the information ratio, a statistic that quantifies the trade-off between the immediate regret incurred and the information gained about the optimal action. We improve upon previous results by establishing that the information ratio is bounded by  $\frac{9}{2}d\alpha^{-2}$ , where  $\alpha$  is a *minimax measure* of the alignment between the action space  $\mathcal{A}$  and the parameter space  $\mathcal{O}$ , and is independent of  $\beta$ . Using this result, we derive a bound of order  $O(d/\alpha\sqrt{T\log(\beta T/d)})$  on the Bayesian expected regret of Thompson Sampling incurred after  $T$  time steps. To our knowledge, this is the first regret bound for logistic bandits that depends only logarithmically on  $\beta$  while being independent of the number of actions. In particular, when the action space contains the parameter space, the bound on the expected regret is of order  $\tilde{O}(d\sqrt{T})$ .

## 1 Introduction

This paper studies the logistic bandit problem, where an agent sequentially interacts with an unknown environment with parameter  $\Theta \in \mathcal{O}$ . At each time step, the agent selects an action  $A_t \in \mathcal{A}$  and receives binary rewards  $R_t \in \{0, 1\}$  with probabilities determined by the logistic function  $\exp(\beta\langle A_t, \Theta \rangle) / (1 + \exp(\beta\langle A_t, \Theta \rangle))$  with slope parameter  $\beta > 0$ . In this setting, both the action space and the parameter space lie within the  $d$ -dimensional unit ball. The goal of the agent is to maximize its total reward, or equivalently to *minimize its regret*, that is the difference between the optimal cumulative reward and the cumulative reward achieved by the agent. This setting is used to model various scenarios, such as click-through rate prediction, spam email detection, and personalized advertisement systems, where, in the latter case, content is suggested to users who provide binary feedback (Chapelle and Li, 2011; Russo and Van Roy, 2018).

The performance, or regret, of algorithms for logistic bandits has been extensively studied, with significant contributions including analyses of Upper-Confidence-Bound (UCB) algorithms (Filippi et al., 2010; Li et al., 2017; Faury et al., 2020) as well as the study of Thompson Sampling (TS) (Russo and Van Roy, 2014a; Dong et al., 2019; Abeille and Lazaric, 2017). However, nearly all existing regret bounds for logistic bandits exhibit an exponential dependence on the parameter  $\beta$  (see Table 1). This dependence is unsatisfactory as, in practice, the distinction between near-optimal and sub-optimal actions gets more pronounced as  $\beta$  increases and it can be faster to find optimal actions. This lead McMahan and Streeter (2012) to make an open call for better bounds.

In this work, we focus on the Thompson Sampling algorithm (Thompson, 1933), which, despite its simplicity, has proven to be effective for a wide range of problems (Russo et al., 2018; Chapelle and Li, 2011). To analyze the TS expected regret, Russo and Van Roy (2016) introduced the concept of the information ratio, defined as the ratio between the squared expected instant regret, and the information gained about the optimal action. Building on this framework, Dong and Van Roy (2018) derived a near-optimal regret rate of  $O(d\sqrt{T\log T})$  for  $d$ -dimensional linear bandit problems. However, applying this analysis to the logistic setting yields regret bounds that grow exponentially with the parameter

$\beta$  (see Appendix E). Using numerical simulations, they conjectured that the TS information ratio for logistic bandits depends only on the problem’s dimension  $d$ .

Algorithm	Regret Upper Bound	Note
Thompson Sampling (Russo and Van Roy, 2014a)	$O(e^\beta \cdot d \cdot T^{1/2} \cdot \log(T)^{3/2})$	Bayesian bound
GLM-TSL (Abeille and Lazaric, 2017)	$O(e^\beta \cdot d^{3/2} \cdot \log(d)^{1/2} \cdot T^{1/2} \log(T)^{3/2})$	Frequentist bound
Thompson Sampling (Dong et al., 2019)	$O(e^\beta \cdot d \cdot T^{1/2} \log(T/d)^{1/2})$	Bayesian bound
GLM-TSL (Kveton et al., 2020)	$O(d \cdot T^{1/2} \cdot \log(T) + e^\beta \beta^6 d^2 \log(T)^2)$	Frequentist bound
Logistic-UCB-2 (Fauray et al., 2020)	$O(d \cdot T^{1/2} \cdot \log(T) + e^\beta \cdot d^2 \cdot \log(T)^2)$	Frequentist bound
Thompson Sampling (this paper)	$O(d/\alpha \cdot T^{1/2} \cdot \log(T\beta/d)^{1/2})$	Bayesian bound, $\alpha$ is independent of $\beta$

Table 1: Comparison of various regret guarantees for the logistic bandit problem.

Recently, Neu et al. (2022) derived a regret bound of the order  $O(\sqrt{dT|\mathcal{A}|\log(\beta T)})$  on the logistic bandit problem, but their result relies on a worst-case information ratio bound scaling with the cardinality of the action space  $|\mathcal{A}|$  and their regret bound becomes vacuous for problems with continuous or infinite action space even though Thompson Sampling is known to perform well under these settings (Russo and Van Roy, 2014a). Studying the TS regret for logistic bandits, Dong et al. (2019) introduced two statistics to characterize the sets  $\mathcal{A}$  and  $\mathcal{O}$ , the *minimax alignment constant*<sup>1</sup>  $\alpha = \min_{\theta \in \mathcal{O}} \max_{a \in \mathcal{A}} \langle a, \theta \rangle$  and the *fragility dimension*  $\eta$ , which is the cardinality of the largest subset of parameters such that their corresponding optimal action is misaligned with any other parameter from the subset. Using those statistics, they showed that for  $\beta < 2$ , the TS information ratio is bounded by  $100 \max(d, \eta) \alpha^{-2}$ . They also suggested, through numerical computations, that this bound holds for larger values of  $\beta$ . However, their work has two key limitations. First, they did not provide a rigorous proof for generalizing their bound to larger values of  $\beta$ . Second, and more critically, their regret analysis is incorrect as it relies on the rate-distortion bound from Dong and Van Roy (2018), which is incompatible with a bound on the TS information ratio. Indeed, the regret analysis in Dong and Van Roy (2018) specifically requires a bound on the one-step compressed TS information ratio, which is a fundamentally different quantity from the TS information ratio studied in the work of Dong et al. (2019). We elaborate on these gaps in more detail in Appendix F.

In this paper, we address these issues and obtain a regret bound that scales only logarithmically with the slope of the logistic function, while remaining independent of the cardinality of the action space. Our key contributions are as follows:

- We prove an information-theoretic regret bound of order  $O(\sqrt{TT(\mathcal{H}(\Theta_\epsilon) + \beta\epsilon T)})$  that holds for infinite and continuous action and parameter spaces. The bound relies on the entropy of the parameter quantized at scale  $\epsilon > 0$ , and on the average expected TS information ratio,  $\Gamma$ .
- We present a new analysis showing that, for all  $\beta > 0$ , the TS information ratio for logistic bandits is bounded by  $\frac{9}{2} d \alpha^{-2}$ , improving upon the previous results. Notably, our bound does not depend on the fragility dimension  $\eta$  which can scale exponential in  $d$ .
- We establish a regret bound of order  $O(d/\alpha \sqrt{T \log(\beta T/d)})$  for Thompson Sampling. To our knowledge, this is the first regret bound for any logistic bandit algorithm that scales only logarithmically with  $\beta > 0$  and is independent of the number of actions. Additionally, we show that if the action space encompasses the parameter space, the expected regret of Thompson Sampling is bounded in  $O(d \sqrt{T \log(\beta T/d)})$  with no dependence on  $\alpha$ .

The rest of the paper is organized as follows. Section 2 introduces the logistic bandit problem, defines the Bayesian expected regret, and the specific notations used. Section 3 introduces the Thompson Sampling algorithm and the information ratio analysis. Section 4 states and discusses our main results, providing the improved regret bounds. Section 5 presents the key ideas for analyzing the information ratio; and finally, Section 6 discusses our results and future extensions.

<sup>1</sup>This statistic is referred to as the *worst-case optimal log-odds* in the work of Dong et al. (2019).

## 2 Problem Setup

We consider a logistic bandit problem, where at each time step  $t \in \{1, \dots, T\}$ , an agent selects an action  $A_t \in \mathcal{A}$  and receives a binary reward  $R_t \in \{0, 1\}$  with probability following a logistic function:

$$\mathbb{P}(R_t = 1 | A_t = a, \Theta = \theta) = \frac{\exp(\beta \langle a, \theta \rangle)}{1 + \exp(\beta \langle a, \theta \rangle)}.$$

Here,  $\beta > 0$  is a known scale parameter, and  $\langle a, \theta \rangle$  denotes the inner product of the action vector  $a \in \mathcal{A}$  and the unknown parameter  $\theta \in \mathcal{O}$ . Throughout the paper, we denote the logistic function as  $\phi_\beta(x) := \frac{\exp(\beta x)}{1 + \exp(\beta x)}$ . As this function is strictly increasing, the probability of obtaining a reward,  $\phi_\beta(\langle a, \theta \rangle)$ , is maximized when the inner product between the action and parameter is maximized. In our setting, both the action  $a$  and the parameter vector  $\theta$  lie within the  $d$ -dimensional Euclidean unit ball,  $\mathbf{B}_d(0, 1)^2$ . For a given action space  $\mathcal{A}$  and parameter space  $\mathcal{O}$ , we define their *minimax alignment constant* as  $\alpha := \min_{\theta \in \mathcal{O}} \max_{a \in \mathcal{A}} \langle a, \theta \rangle$ . In the rest of the paper, we assume that the action and parameter spaces are such that  $\alpha \geq 0$ . This assumption is relatively mild, as it suffices for  $\mathcal{A}$  to contain two opposed actions  $a, a'$  (i.e.  $a = -a'$ ) to ensure  $\alpha \geq 0$  for any parameter set  $\mathcal{O}$ .

Following the Bayesian framework, we assume the parameter vector  $\Theta \in \mathcal{O}$  is sampled from a known prior distribution  $\mathbb{P}_\Theta$ . This prior, together with the reward distribution  $\mathbb{P}_{R|A,\Theta}$ , fully describes the logistic bandit problem. As the reward distribution depends only on the selected action and the parameter, it can be written as  $R_t = R(A_t, \Theta)$  for some random function  $R : \mathcal{A} \times \mathcal{O} \rightarrow \mathbb{R}$ . The agent's history at time  $t$  is denoted by  $H^t = \{A_1, R_1, \dots, A_{t-1}, R_{t-1}\}$ , representing all past actions and rewards observed up to time  $t$ . The goal of the agent is to sequentially select actions that maximize the total cumulated reward, or equivalently, that minimize the total expected regret defined as:

$$\mathbb{E}[\text{Regret}(T)] := \mathbb{E} \left[ \sum_{t=1}^T R(A^*, \Theta) - R(A_t, \Theta) \right],$$

where  $A^*$  is the *optimal action* corresponding to the parameter  $\Theta$ . We construct the optimal mapping  $\pi_\star(\theta) := \arg\max_{a \in \mathcal{A}} \mathbb{E}[R(a, \theta)]$  so that we can write  $A^* = \pi_\star(\Theta)$ . To ensure such a mapping exists, we make the technical assumption that the set of actions  $\mathcal{A}$  is compact. Following Dong et al. (2019), we assume without loss of generality that the mapping  $\pi_\star$  is one-to-one<sup>3</sup>.

Since  $\sigma$ -algebras of the history are often used in conditioning, we introduce the notations  $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot | H^t]$  and  $\mathbb{P}_t[\cdot] := \mathbb{P}[\cdot | H^t]$  to denote the conditional expectation and probability given the history  $H^t$ , respectively. Additionally, we define  $\mathbf{I}_t(A^*; R_t | A_t) := \mathbb{E}_t[\text{D}_{\text{KL}}(\mathbb{P}_{R_t | H^t, A^*, A_t} \| \mathbb{P}_{R_t | H^t, A_t})]$  as the disintegrated conditional mutual information between the optimal action  $A^*$  and the reward  $R_t$  conditioned on the action  $A_t$ , given the history  $H^t$ .

## 3 Thompson Sampling and the Information Ratio

An elegant algorithm for solving bandit problems is the *Thompson Sampling* algorithm. It works by randomly selecting actions according to their posterior probability of being optimal. More specifically, at each time step  $t \in \{1, \dots, T\}$ , the agent samples a parameter estimate  $\hat{\Theta}_t$  from the posterior distribution conditioned on the history  $H^t$  and selects the action that is optimal for the sampled parameter estimate,  $A_t = \pi_\star(\hat{\Theta}_t)$ . The pseudocode for the algorithm is given in Algorithm 1.

Studying the regret of Thompson Sampling, Russo and Van Roy (2016) introduced a key quantity to the analysis, the *information ratio* defined as the following random variable:

$$\Gamma_t := \frac{\mathbb{E}_t[R(A^*, \Theta) - R(A_t, \Theta)]^2}{\mathbf{I}_t(A^*; R(A_t, \Theta), A_t)}.$$

This ratio measures the trade-off between minimizing the current squared regret and gathering information about the optimal action; a small ratio indicating that a substantial gain of information about the optimal action compensated

<sup>2</sup>This setting is equivalent to the one considered by Faury et al. (2020) using  $\beta$  as the maximal norm for  $\theta \in \mathcal{O}$ .

<sup>3</sup>If a particular parameter is optimal for multiple actions, we can arbitrarily fix the mapping of that parameter to one of the optimal actions. Conversely, if a particular action is optimal for multiple parameters, we can introduce duplicate action labels to ensure a one-to-one correspondence between each parameter and its optimal action label. A rigorous explanation of this construction is provided in Appendix G.

**Algorithm 1** Thompson Sampling algorithm

---

```

1: Input: parameter prior  $\mathbb{P}_\Theta$ , mapping  $\pi_\star$ .
2: for  $t = 1$  to  $T$  do
3:   Sample a parameter estimate  $\hat{\Theta}_t \sim \mathbb{P}_{\Theta|H^t}$ .
4:   Take the corresponding optimal action  $A_t = \pi_\star(\hat{\Theta}_t)$ .
5:   Collect the reward  $R_t = R(A_t, \Theta)$ .
6:   Update the history  $H^{t+1} = H^t \cup \{A_t, R_t\}$ .
7: end for

```

---

for any significant regret. Russo and Van Roy (2016) use this concept to provide a general regret bound in  $O(\sqrt{\Gamma T H(A^\star)})$ , that depends on the time horizon  $T$ , the prior entropy of the optimal action  $H(A^\star)$ , and an algorithm- and problem-dependent upper bound on the average expected information ratio  $\Gamma$ .

A limitation of this approach is that the prior entropy of the optimal action,  $H(A^\star)$ , can grow arbitrarily large with the number of actions and gets infinite with continuous action space. We address this issue in Theorem 2, where we propose a regret bound depending instead on the entropy of a quantization of the parameter  $\Theta$  at scale  $\epsilon > 0$ .

## 4 Main Results

This section presents our main results on the Thompson Sampling regret for logistic bandits. In Theorem 2, we derive an information-theoretic regret bound for logistic bandits that holds for continuous and infinite parameter spaces. Following this, we state in Proposition 3 our principal contribution, a bound on the TS information ratio depending only on the problem’s dimension  $d$  and on the minimax alignment constant  $\alpha$ . By combining this result with our regret bound, we derive in Theorem 4, a bound on the expected regret of TS for logistic bandits, which scales as  $O(d/\alpha\sqrt{T\log(\beta T/d)})$ .

Our first theorem provides a regret bound that holds for large and continuous action spaces and relies on the entropy of the quantized parameter  $\Theta_\epsilon$ , defined in Definition 1, which is as the closest approximation for  $\Theta$  (as measured by the metric  $\rho$ ) on an  $\epsilon$ -net for  $(\mathcal{O}, \rho)$ . Compared to Dong and Van Roy (2018, Theorem 1), our result is compatible with bounds on the “standard” TS information ratio, rather than the “one-step compressed TS” information ratio. This distinction is crucial, as the latter is significantly more challenging to analyze in logistic bandits due to its intricate construction.

**Definition 1.** Let the set  $\mathcal{O}_\epsilon$  be an  $\epsilon$ -net for  $(\mathcal{O}, \rho)$  with associated projection mapping  $q : \mathcal{O} \rightarrow \mathcal{O}_\epsilon$  such that for all  $\theta \in \mathcal{O}$  we have  $\rho(\theta, q(\theta)) \leq \epsilon$ . We define the quantized parameter as  $\Theta_\epsilon := q(\Theta)$ .

**Theorem 2.** For all  $\beta > 0$ , under the logistic bandit setting with logistic function  $\phi_\beta(x)$ , let the quantized parameter  $\Theta_\epsilon$  be defined as in Definition 1 for some  $\epsilon > 0$ . If the average expected TS information ratio is bounded,  $\frac{1}{T} \sum_{t=1}^T \mathbb{E}[\Gamma_t] \leq \Gamma$ , for some  $\Gamma > 0$ , then the TS regret is bounded as

$$\mathbb{E}[\text{Regret}(T)] \leq \sqrt{\Gamma T (H(\Theta_\epsilon) + \epsilon \beta T)}.$$

The proof of Theorem 2 adapts the techniques of Neu et al. (2022, Theorem 2) and Gouverneur et al. (2023, Theorem 2) to the logistic bandits setting. It relies on an approximation of the conditional mutual information  $I(\Theta; R_t | A_t, H^t)$  as  $I(\Theta_\epsilon; R_t | A_t, H^t) + \beta \epsilon$  exploiting the fact that, for all  $a \in \mathcal{A}$ , the log-likelihood of  $R(a, \theta)$  is  $\beta$ -Lipschitz with respect to  $\theta \in \mathcal{O}$ .

*Proof.* We start by rewriting the TS expected regret using the information ratio:

$$\mathbb{E}[\text{Regret}(T)] = \sum_{t=1}^T \mathbb{E}[R(A^\star, \Theta) - R(A_t, \Theta)] = \sum_{t=1}^T \mathbb{E} \left[ \sqrt{\Gamma_t I_t(A^\star; R(A_t, \Theta), A_t)} \right].$$

We continue using Jensen’s inequality, followed by Cauchy-Schwarz inequality:

$$\mathbb{E}[\text{Regret}(T)] \leq \sum_{t=1}^T \sqrt{\mathbb{E}[\Gamma_t] I(A^\star; R(A_t, \Theta), A_t | H^t)} \leq \sqrt{\Gamma T \sum_{t=1}^T I(A^\star; R(A_t, \Theta), A_t | H^t)},$$

where in the last inequality, we used  $\sum_{t=1}^T \mathbb{E}_t[\Gamma_t] \leq \Gamma T$ . Applying the chain rule (Yury Polyanskiy, 2022, Theorem 3.7.b) we decompose the mutual information as

$$\begin{aligned} I(A^*; R(A_t, \Theta), A_t | H^t) &= I(A^*; A_t | H^t) + I_t(A^*; R(A_t, \Theta) | H^t, A_t) \\ &= I(\pi_*(\Theta); R(A_t, \Theta) | H^t, A_t) \\ &= I(\Theta; R(A_t, \Theta) | H^t, A_t), \end{aligned}$$

where we used the fact that the mutual information  $I_t(A^*; A_t | H^t) = 0$  as  $A^*$  and  $A_t$  are independent conditioned on the history  $H^t$  and the last equality follows as  $\pi_*$  is a one-to-one mapping (Yury Polyanskiy, 2022, Theorem 3.2.d).

Let  $f_{R_t | H^t, A_t, \Theta}$  denote the probability density of  $R_t$  conditioned on  $H^t, A_t, \Theta$  and  $f_{R_t | H^t, A_t}$  denote the probability density on  $H^t, A_t$ . Then, the mutual information terms can be written as

$$I(\Theta; R_t | H^t, A_t) = \mathbb{E} \left[ \log \frac{f_{R_t | H^t, A_t, \Theta}(R_t)}{f_{R_t | H^t, A_t}(R_t)} \right].$$

We let the set  $\mathcal{O}_\varepsilon$  be an  $\varepsilon$ -net for  $(\mathcal{O}, \rho)$  with associated mapping  $q : \mathcal{O} \rightarrow \mathcal{O}_\varepsilon$  and similarly to Neu et al. (2022, Theorem 2), we observe that the mutual information can be written as

$$\mathbb{E} \left[ \int_{\mathcal{O}} f_{\Theta | R_t, H^t, A_t}(\theta) \left( \log \frac{f_{R_t | A_t, \Theta=\theta}(R_t)}{f_{R_t | A_t, \Theta=q(\theta)}(R_t)} + \log \frac{f_{R_t | H^t, A_t, \Theta=q(\theta)}(R_t)}{f_{R_t | H^t, A_t}(R_t)} \right) d\theta \right], \quad (1)$$

using the fact that  $f_{R_t | H^t, A_t, \Theta} = f_{R_t | A_t, \Theta}$  by the conditional Markov chain  $R_t - A_t - H_t | \Theta$ .

Since the derivative of  $\log(\phi_\beta(x))$  is  $\beta/(1 + \exp(\beta x))$ , it is bounded by  $\beta$ , which makes it  $\beta$ -Lipschitz. Furthermore, for all  $a \in \mathcal{A}$  and  $\theta \in \mathcal{O}$ , the inner product  $\langle a, \theta \rangle \leq 1$ , implying that  $\log(f_{R_t | A_t, \Theta=\theta}(1))$  is also  $\beta$ -Lipschitz with respect to  $\theta$ . Similarly, the derivative of  $\log(1 - \phi_\beta(x))$  is bounded by  $\beta$ , making it  $\beta$ -Lipschitz as well. Consequently,  $\log(f_{R_t | A_t, \Theta=\theta}(0))$  is  $\beta$ -Lipschitz with respect to  $\theta \in \mathcal{O}$ . Thus, we conclude that for all  $\theta \in \mathcal{O}$ , we have  $|\log f_{R_t | A_t, \Theta=\theta}(R_t) - \log f_{R_t | A_t, \Theta=q(\theta)}(R_t)| \leq \beta \rho(\theta, q(\theta)) \leq \beta \varepsilon$ .

After defining the random variable  $\Theta_\varepsilon := q(\Theta)$ , the second term in eq. (1) is equal to  $I(\Theta_\varepsilon; R_t | H^t, A_t)$ . Summing the  $T$  mutual information  $I(\Theta_\varepsilon; R_t | H^t, A_t)$  and applying the chain rule, we obtain

$$\mathbb{E}[\text{Regret}(T)] \leq \sqrt{\Gamma T (I(\Theta_\varepsilon; H^T) + \varepsilon \beta T)}.$$

Finally, we upper bound  $I(\Theta_\varepsilon; H^T)$  by the entropy  $H(\Theta_\varepsilon)$  to obtain the claimed result.  $\square$

In the following, we present our main proposition, a bound on the TS information ratio that depends only on the problem dimension  $d$  and the minimax alignment constant  $\alpha$ .

**Proposition 3.** *For all  $\beta > 0$ , and for all  $\mathcal{A}, \mathcal{O} \subseteq \mathbf{B}_d(0, 1)$  with minimax alignment constant  $\alpha$ , under the logistic bandit setting with logistic function  $\phi_\beta(x)$  the TS information ratio is bounded as*

$$\Gamma_t \leq \frac{9}{2} d \alpha^{-2}.$$

At a high level, our proof consists of three parts: a lower bound on the conditional mutual information,  $I_t(A^*; R(A_t, \Theta), A_t)$ , an upper bound on the squared expected regret at time  $t$ ,  $\mathbb{E}_t[R(A^*, \Theta) - R(A_t, \Theta)]^2$ , and an upper bound on a ratio of expected variances by the study of the limit case  $\beta \rightarrow \infty$ . The key techniques for the proof of Proposition 3 are presented in Section 5.

Combining Proposition 3 with Theorem 2, we arrive at our main result: a bound on the expected TS regret in  $O(d/\alpha \sqrt{T \log(\beta T/d)})$ . To the best of our knowledge, this is the first regret bound for logistic bandits that scales only logarithmically with the logistic function's parameter  $\beta$  while remaining independent of the number of actions.

**Theorem 4.** *For all  $\beta > 0$ , and for all  $\mathcal{A}, \mathcal{O} \subseteq \mathbf{B}_d(0, 1)$  with minimax alignment constant  $\alpha$ , under the logistic bandit setting with logistic function  $\phi_\beta(x)$ , the TS regret is bounded as*

$$\mathbb{E}[\text{Regret}(T)] \leq 3d/\alpha \sqrt{T \log \left( \sqrt{3 + \frac{6\beta T}{d}} \right)}.$$

*Proof.* Combining Theorem 2 with Proposition 3, we upper bound the entropy  $H(\Theta_\varepsilon)$  by the cardinality of the  $\varepsilon$ -net to get a regret bound of  $3/\alpha \sqrt{dT/2 (\log(|\Theta_\varepsilon|) + \varepsilon\beta T)}$ . To define  $\Theta_\varepsilon$ , we set  $\mathcal{O}_\varepsilon$  as the  $\varepsilon$ -net of smallest cardinality. As the parameter space  $\mathcal{O}$  is within the Euclidean unit ball, we use Lemma 14 to control the covering number as  $\log(|\Theta_\varepsilon|) \leq d \log(1 + 2/\varepsilon)$  to upper bound the TS regret as

$$\mathbb{E}[\text{Regret}(T)] \leq 3/\alpha \sqrt{dT/2 \left( d \log \left( 1 + \frac{2}{\varepsilon} \right) + \varepsilon\beta T \right)}.$$

Finally, setting  $\varepsilon = d/(\beta T)$  and rearranging terms inside the logarithm yields the desired result.  $\square$

Importantly, the above theorem does not depend on the fragility dimension  $\eta$ , in contrast to the results of Dong et al. (2019). This distinction matters as, except in the case where  $\alpha = 1$ , the fragility dimension can grow exponentially with the dimension  $d$ . We can verify that our result, due to its logarithmic dependence on  $\beta$ , is compatible with Dong et al. (2019, Proposition 11), which shows that there exist logistic bandit problems for which no algorithm can achieve a Bayesian regret uniformly bounded by  $f(\alpha)p(d)T^{1-\epsilon}$ , for some function  $f$ , polynomial  $p$ , and  $\epsilon > 0$ .

The next two corollaries present cases where the dependence on the minimax alignment constant  $\alpha$  can be removed. The case in Corollary 6 is particularly relevant for applications where the action set can be treated as a design parameter, and where constructing large action spaces is not prohibitive. We illustrate the improvement of Corollary 5 over previous works through numerical experiments on a synthetic logistic bandit problem. The results are presented in Appendix D.

**Corollary 5.** *For all  $\beta > 0$ , under the logistic bandit setting with logistic function  $\phi_\beta(x)$ , let  $\mathcal{A} \subseteq \mathbf{B}_d(0, 1)$  and  $\mathcal{O} \subseteq \mathbf{S}_d(0, 1)$  be such that  $\mathcal{O} \subseteq \mathcal{A}$ . Then the TS regret is bounded as*

$$\mathbb{E}[\text{Regret}(T)] \leq 3d \sqrt{T \log \left( \sqrt{3 + \frac{6\beta T}{d}} \right)}.$$

*Proof.* If  $\mathcal{O} \subseteq \mathbf{S}_d(0, 1)$  and if  $\mathcal{O} \subseteq \mathcal{A}$ , then for each  $\theta \in \mathcal{O}$ , there exists an action  $a \in \mathcal{A}$  such that  $a = \theta$  and  $\langle a, \theta \rangle = 1$ , implying  $\alpha = 1$ . Using Theorem 4 concludes the proof.  $\square$

**Corollary 6.** *For all  $\beta > 0$ , under the logistic bandit setting with logistic function  $\phi_\beta(x)$ , there exists an action space  $\mathcal{A}$  with  $|\mathcal{A}| \leq 2d \cdot 3^{d-1}$  such that for any  $\mathcal{O} \subseteq \mathbf{S}_d(0, 1)$ , the TS regret is bounded as*

$$\mathbb{E}[\text{Regret}(T)] \leq 6d \sqrt{T \log \left( \sqrt{3 + \frac{6\beta T}{d}} \right)}.$$

*Proof.* Starting from Theorem 4, we have to construct  $\mathcal{A}$  such that its minimax alignment constant  $\alpha$  is greater or equal to  $\frac{1}{2}$  for any  $\mathcal{O} \subseteq \mathbf{S}_d(0, 1)$ . This is satisfied if  $\mathcal{A}$  is a  $\frac{1}{2}$ -net for  $\mathbf{S}_d(0, 1)$ . Setting  $\mathcal{A}$  as the  $\frac{1}{2}$ -net of minimal cardinality, from Lemma 15, we have  $|\mathcal{A}| \leq 2d \cdot 3^{d-1}$ .  $\square$

## 5 Analysis

This section presents the main technique of the proofs of our main proposition, Proposition 3. For the sake of clarity, we present here our results for the particular setting of Corollary 5, which ensures  $\alpha = 1$ . We prove how to extend those results to general spaces in Appendix C. Our proof can be divided into three parts: a lower bound on the mutual information (Section 5.1), an upper bound on the squared expected regret (Section 5.2), and an upper bound on a ratio of expected variances by analyzing the limit case as  $\beta \rightarrow \infty$  (Section 5.3). To alleviate the notations, we omit the subscript  $t$  for the rest of the section.

A key quantity to our analysis is the expected variance of the reward probability conditioned on the sampled action, expressed as  $\mathbb{E}[\mathbb{V}[\phi_\beta(\langle \hat{A}, \Theta \rangle) | \hat{A}]]$ . We will use it to lower bound on mutual information and a related quantity is used to upper bound the squared expected regret. Intuitively, when the variance of reward probability is high, the agent is exploring new actions, gathering information about  $\Theta$  but suffering regret. Conversely, if the variance of reward probability is low, it indicates that the agent has already identified near-optimal actions and is exploiting this knowledge.

Under the logistic bandit setting with logistic function  $\phi_\beta$ , the reward  $R(A_t, \Theta)$  is given by a Bernoulli random variable with associated probability  $\phi_\beta(\langle A_t, \Theta \rangle)$ . We denote it by  $\text{Bern}(\phi_\beta(\langle A_t, \Theta \rangle))$  to make the setting more explicit. With this notation, we write the information ratio as:

$$\Gamma = \frac{\mathbb{E}[\text{Bern}(\phi_\beta(\langle A^*, \Theta \rangle)) - \text{Bern}(\phi_\beta(\langle \hat{A}, \Theta \rangle))]^2}{\text{I}(A^*; \text{Bern}(\phi_\beta(\langle \hat{A}, \Theta \rangle)), \hat{A})}.$$

### 5.1 Lower bounding the mutual information

We start by stating a general lemma that relates the variance of a  $[0, 1]$  random variable  $U$  to the mutual information between  $U$  and a Bernoulli outcome with probability  $U$ . The proof, presented in Appendix A, uses the decomposition of mutual information as a difference of entropy and the Taylor expansion of the entropy function.

**Lemma 7.** *Let  $U$  be a random variable taking values in  $[0, 1]$  and  $\text{Bern}(U)$  be a Bernoulli random variable with probability  $U$ . Then it holds that,*

$$\text{I}(U; \text{Bern}(U)) \geq 2\mathbb{V}(U).$$

Using Lemma 7, we prove a lower bound on the mutual information  $\text{I}(A^*; \text{Bern}(\phi_\beta(\langle \hat{A}, \Theta \rangle)), \hat{A})$  is lower bounded by the expected variance of reward probability  $\mathbb{E}[\mathbb{V}[\phi_\beta(\langle \hat{A}, \Theta \rangle) | \hat{A}]]$ .

**Lemma 8.** *Let the logistic function be  $\phi_\beta(x)$ , then, it holds that*

$$\text{I}(A^*; \text{Bern}(\phi_\beta(\langle \hat{A}, \Theta \rangle)), \hat{A}) \geq 2\mathbb{E} \left[ \mathbb{V} \left[ \phi_\beta(\langle \hat{A}, \Theta \rangle) \mid \hat{\Theta} \right] \right].$$

*Proof.* We start by applying the chain rule. It comes that

$$\begin{aligned} \text{I}(\pi_\star(\Theta); \pi_\star(\hat{\Theta}), \text{Bern}(\phi_\beta(\langle \hat{A}, \Theta \rangle))) &\stackrel{(i)}{=} \text{I}(\Theta; \hat{\Theta}, \text{Bern}(\phi_\beta(\langle \hat{A}, \Theta \rangle))) \\ &\stackrel{(j)}{=} \text{I}(\Theta; \hat{\Theta}) + \text{I}(\Theta; \text{Bern}(\phi_\beta(\langle \hat{A}, \Theta \rangle)) \mid \hat{\Theta}) \\ &\stackrel{(k)}{=} \text{I}(\Theta; \text{Bern}(\phi_\beta(\langle \hat{A}, \Theta \rangle)) \mid \hat{\Theta}) \\ &\stackrel{(l)}{=} \mathbb{E}[\text{I}(\phi_\beta(\langle \hat{A}, \Theta \rangle); \text{Bern}(\phi_\beta(\langle \hat{A}, \Theta \rangle))) \mid \hat{\Theta} = \theta], \end{aligned}$$

where (i) follows as  $\pi_\star$  is a one-to-one mapping; (j) follows from the chain-rule; (k) follows as  $\Theta$  and  $\hat{\Theta}$  are independent conditioned on the history; and (l) follows as  $\phi_\beta(\langle a, \theta \rangle)$  is a one-to-one mapping conditioned on  $\hat{\Theta} = \theta$ . Finally, applying Lemma 7 yields the desired result.  $\square$

### 5.2 Upper bounding the squared expected regret

This part of the analysis is inspired by the proof techniques of Dong et al. (2019, Proposition 15) and similar to them, the two following lemmata will be of importance for our analysis.

**Lemma 9** (Dong et al. (2019, Lemma 16)). *Let  $U, V$  be random vectors in  $\mathbb{R}^d$ , and let  $\tilde{U}, \tilde{V}$  be independent random variables with distributions equal respectively to the marginals of  $U, V$ , then*

$$\mathbb{E}[(U^\top V)]^2 \leq d \cdot \mathbb{E}[(\tilde{U}^\top \tilde{V})^2].$$

**Lemma 10** (Dong et al. (2019, Lemma 18)). *Let  $f : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  be such that  $f(0) \geq 0$  and  $f(\zeta)/\zeta$  is non-decreasing over  $\zeta \geq 0$ . Then, for any non-negative random variable  $U$ , there is*

$$\frac{\mathbb{E}[f(U)]^2}{\mathbb{E}[U]^2} \leq \frac{\mathbb{V}[f(U)]}{\mathbb{V}[U]}.$$

Under the assumption that  $\mathcal{O} \subseteq \mathbf{S}_d(0, 1)$  with  $\mathcal{O} \subseteq \mathcal{A}$ , then for each  $\theta \in \mathcal{O}$ , there exists an action  $a \in \mathcal{A}$  such that  $a = \theta$  and  $\langle a, \theta \rangle = 1$ . This implies that  $\phi_\beta(\langle A^*, \Theta \rangle) - \phi_\beta(\langle \hat{A}, \Theta \rangle) = \phi_\beta(1) - \phi_\beta(\langle \hat{A}, \Theta \rangle)$ . To simplify notation, we define  $\psi_\beta(x) := \phi_\beta(1) - \phi_\beta(1 - x)$ , which relates the difference between the optimal action  $A^*$  and the sampled action  $\hat{A}$  to their corresponding reward differences. Specifically, we have

$$\psi_\beta(1 - \langle \hat{A}, \Theta \rangle) = \psi_\beta(\langle A^* - \hat{A}, \Theta \rangle) = \phi_\beta(\langle A^*, \Theta \rangle) - \phi_\beta(\langle \hat{A}, \Theta \rangle).$$

The function  $\psi_\beta(x)$  meets the first two conditions from Lemma 10 and when it is applied to the difference of inner products  $\langle A^*, \Theta \rangle - \langle \hat{A}, \Theta \rangle$ , it maps the interval  $[0, 2]$  to  $[0, 1]$ , and satisfies  $\psi_\beta(0) = \phi_\beta(1) - \phi_\beta(1 - 0) = 0$ . However, it does not meet the third condition, as  $\psi_\beta(x)/x$  initially increases, reaches a maximum between 1 and 2, and then decreases (see Remark 17). To address this issue, we introduce a modified function, referred to as the *logistic surrogate*, which serves as the tightest upper bound on  $\psi_\beta(x)$  that satisfies the last requirement from Lemma 10.

**Definition 11** (Logistic surrogate). *We construct the logistic surrogate function  $\varphi_\beta(x)$  as the tightest upper bound on  $\psi_\beta(x)$  such that  $\varphi_\beta(x)/x$  is non-decreasing over  $x \geq 0$ .*

Namely, let  $\delta_\beta = \arg \max_{x \in [0, 2]} \frac{\psi_\beta(x)}{x}$ , we define  $\varphi_\beta$  as

$$\varphi_\beta(x) = \begin{cases} \psi_\beta(x) & x \in [0, \delta_\beta] \\ \psi_\beta(\delta_\beta) + (x - \delta_\beta) \cdot \psi_\beta(\delta_\beta)/\delta_\beta & x \in ]\delta_\beta, 2] \end{cases}.$$

We are now equipped to state and prove an upper bound on the squared expected regret.

**Lemma 12.** *Let the logistic surrogate be defined as in Definition 11. Then, it holds that*

$$\mathbb{E}[\text{Bern}(\phi_\beta(\langle A^*, \Theta \rangle)) - \text{Bern}(\phi_\beta(\langle \hat{A}, \Theta \rangle))]^2 \leq d \cdot \mathbb{E} \left[ \mathbb{V} \left[ \varphi_\beta \left( 1 - \langle \hat{A}, \Theta \rangle \right) \mid \hat{\Theta} \right] \right].$$

*Proof.* Integrating over the randomness of the Bernoulli outcome, the squared expected regret can be expressed as  $\mathbb{E}[(\phi_\beta(\langle A^*, \Theta \rangle) - \phi_\beta(\langle \hat{A}, \Theta \rangle))]^2 = \mathbb{E}[\psi_\beta(1 - \langle \hat{A}, \Theta \rangle)]^2$ . Since by definition  $\varphi_\beta(x) \geq \psi_\beta(x)$ , we have  $\mathbb{E}[\psi_\beta(1 - \langle \hat{A}, \Theta \rangle)]^2 \leq \mathbb{E}[\mathbb{E}[\varphi_\beta(1 - \langle \hat{A}, \Theta \rangle) \mid \hat{\Theta}]]^2$ .

We now apply Lemma 10 and have that

$$\begin{aligned} \mathbb{E}[\mathbb{E}[\varphi_\beta(1 - \langle \hat{A}, \Theta \rangle) \mid \hat{\Theta}]]^2 &\leq \mathbb{E} \left[ \underbrace{\sqrt{\frac{\mathbb{V} \left[ \varphi_\beta \left( 1 - \langle \hat{A}, \Theta \rangle \right) \mid \hat{\Theta} \right]}{\mathbb{V} \left[ 1 - \langle \hat{A}, \Theta \rangle \mid \hat{\Theta} \right]}}}_{:=U(\hat{\Theta})} \mathbb{E} \left[ 1 - \langle \hat{A}, \Theta \rangle \mid \hat{\Theta} \right] \right]^2 \\ &= \mathbb{E} \left[ U(\hat{\Theta}) \langle \hat{A}, \hat{\Theta} \rangle - \langle \hat{A}, \Theta \rangle \right]^2 = \mathbb{E} \left[ \langle U(\hat{\Theta}) \hat{A}, \Theta - \hat{\Theta} \rangle \right]^2 \end{aligned}$$

We use Lemma 9 with  $U=U(\hat{\Theta})\hat{A}$  and  $V=\Theta-\hat{\Theta}$  and rearrange terms to obtain the claimed result:

$$\begin{aligned} \mathbb{E} \left[ \langle U(\hat{\Theta}) \hat{A}, \Theta - \hat{\Theta} \rangle \right]^2 &\leq d \cdot \mathbb{E} \left[ \left( \langle U(\hat{\Theta}) \hat{A}, \Theta - \hat{\Theta} \rangle \right)^2 \right] = d \cdot \mathbb{E} \left[ U(\hat{\Theta})^2 \mathbb{E} \left[ \langle \hat{A}, \Theta - \hat{\Theta} \rangle^2 \mid \hat{\Theta} \right] \right] \\ &= d \cdot \mathbb{E} \left[ \frac{\mathbb{V} \left[ \varphi_\beta \left( 1 - \langle \hat{A}, \Theta \rangle \right) \mid \hat{\Theta} \right]}{\mathbb{V} \left[ 1 - \langle \hat{A}, \Theta \rangle \mid \hat{\Theta} \right]} \mathbb{V} \left[ \langle \hat{A}, \Theta \rangle \mid \hat{\Theta} \right] \right] \\ &= d \cdot \mathbb{E} \left[ \mathbb{V} \left[ \varphi_\beta \left( 1 - \langle \hat{A}, \Theta \rangle \right) \mid \hat{\Theta} \right] \right]. \end{aligned}$$

□

Combining Lemma 8 and Lemma 12, we get that the information ratio  $\Gamma$  is bounded by

$$\Gamma \leq d/2 \cdot \frac{\mathbb{E} \left[ \mathbb{V} \left[ \varphi_\beta \left( 1 - \langle \hat{A}, \Theta \rangle \right) \mid \hat{\Theta} \right] \right]}{\mathbb{E} \left[ \mathbb{V} \left[ \psi_\beta \left( 1 - \langle \hat{A}, \Theta \rangle \right) \mid \hat{\Theta} \right] \right]},$$

where we use the fact that  $\mathbb{V} \left[ \phi_\beta(\langle \hat{A}, \Theta \rangle) \mid \hat{\Theta} \right] = \mathbb{V} \left[ \psi_\beta(1 - \langle \hat{A}, \Theta \rangle) \mid \hat{\Theta} \right]$  by the definition of  $\psi_\beta$ . The next part of the proof takes care of controlling the ratio of expected variances over  $\varphi_\beta$  and  $\psi_\beta$ .



### 5.3 Bounding the ratio of expected variances over the functions $\varphi_\beta$ and $\psi_\beta$

By definition, the function  $\psi_\beta$  and its surrogate  $\varphi_\beta$  are equal for  $x \in [0, \delta_\beta]$  and then diverge linearly at a rate of  $\psi_\beta(\delta_\beta)/\delta_\beta$ . We observe, in Remark 17, that  $\delta_\beta$  is a decreasing function of  $\beta$  and that the slope  $\psi_\beta(\delta_\beta)/\delta_\beta$  strictly increases with  $\beta$ . This observation suggests that studying the case  $\beta \rightarrow \infty$  could provide a general upper bound. Indeed, taking the limit case  $\beta \rightarrow \infty$ , the domain where the two functions differ is maximized, and the rate at which they differ is the largest.

We show in Lemma 16, presented in Appendix B, that under some simple preliminary transformations, increasing the value of  $\beta$  leads to a larger ratio of expected variances, and therefore, the case  $\beta \rightarrow \infty$  can serve to derive general upper bounds. Quite satisfyingly, this limit case provides a lot of simplifications. We prove in Lemma 16, that the ratio of expected variance between  $\psi_\beta$  and  $\varphi_\beta$  is upper bounded by the ratio of expected variance between  $\bar{\psi}$  and  $\bar{\varphi}$  defined as

$$\bar{\psi}(x) = \begin{cases} 0 & x \in [0, 1] \\ 1 & x \in ]1, 2] \end{cases}, \quad \text{and} \quad \bar{\varphi}(x) = \begin{cases} 0 & x \in [0, 1] \\ 1 + 2(x - 1) & x \in ]1, 2] \end{cases}. \quad (2)$$

**Lemma 13.** *Let  $\bar{\psi}$  and  $\bar{\varphi}$  be defined as in eq. (2). Then, it holds that*

$$\frac{\mathbb{E} \left[ \mathbb{V} \left[ \bar{\varphi} (1 - \langle \hat{A}, \Theta \rangle) \mid \hat{\Theta} \right] \right]}{\mathbb{E} \left[ \mathbb{V} \left[ \bar{\psi} (1 - \langle \hat{A}, \Theta \rangle) \mid \hat{\Theta} \right] \right]} \leq 9.$$

**Sketch of proof** Analyzing  $\bar{\psi}$ , we note that  $\mathbb{E}[\mathbb{V}[\bar{\psi}(1 - \langle \hat{A}, \Theta \rangle) \mid \hat{\Theta}]]$  is equal to the expected variance of a Bernoulli random variable with probability given by  $Q(\hat{A}) = \mathbb{E}[I(\langle \hat{A}, \Theta \rangle)]$  where  $I(\langle \hat{A}, \Theta \rangle) := \mathbb{1}_{\{\langle \hat{A}, \Theta \rangle < 0\}}$ . The expected variance can then be written as  $\mathbb{E}[I(\langle \hat{A}, \Theta \rangle)^2] - \mathbb{E}[\mathbb{E}[I(\langle \hat{A}, \Theta \rangle)^2]]$  where in the second term, the outer expectation is on  $\hat{A}$ , and the inner one is on  $\Theta$ . After rearranging terms, we can write  $\mathbb{E}[\mathbb{V}[\bar{\varphi}(1 - \langle \hat{A}, \Theta \rangle) \mid \hat{\Theta}]]$  as  $\mathbb{E}[I(\langle \hat{A}, \Theta \rangle)(1 - 2\langle \hat{A}, \Theta \rangle)^2] - \mathbb{E}[\mathbb{E}[I(\langle \hat{A}, \Theta \rangle)(1 - 2\langle \hat{A}, \Theta \rangle)^2]]$  where again for the second term, the outer expectation is on  $\hat{A}$ , and the inner one is on  $\Theta$ . Taking the supremum over the possible values of  $(1 - 2\langle a, \theta \rangle) \in [-1, 3]$  concludes the proof.

## 6 Conclusion and Future Work

In this work, we analyzed the performance of the Thompson Sampling algorithm for logistic bandit problems. Following the information-theoretic framework from Russo and Van Roy (2016), we study the information ratio, a key statistic that captures the trade-off between exploration and exploitation in logistic bandits. Our main result establishes that the information ratio of Thompson Sampling for logistic bandits can be bounded using only the dimension of the problem,  $d$ , and  $\alpha$ , a minimax alignment constant between the action and parameter spaces. Importantly, this bound is independent of the logistic function’s slope parameter,  $\beta$ .

Building on this result, we derive a regret bound of  $O(d/\alpha \sqrt{T \log(\beta T/d)})$ , which scales only logarithmically with  $\beta$ , representing a significant improvement over prior works. To the best of our knowledge, this is the first regret bound for logistic bandits that achieves logarithmic dependence on  $\beta$  while remaining independent of the action set’s cardinality. Finally, we presented specific settings where the dependence on  $\alpha$  can be controlled. For instance, when the action space fully encompasses the parameter space, the regret of Thompson Sampling scales as  $\tilde{O}(d\sqrt{T})$ .

A promising direction for future work is to extend our analysis to the broader class of generalized linear bandits. The properties of the logistic function that we leverage in our analysis could be shared by other class of link function and be used to derive regret bounds using a similar analysis of the information ratio as we performed in Section 5.

Another interesting research direction is to use the result in this paper to derive regret bounds for logistic bandits in the frequentist setting. A promising way is to apply our information theoretic analysis to the optimistic information directed sampling algorithm introduced by Neu et al. (2024). We believe that this approach could lead to deriving new and improved frequentist bounds for logistic bandits.

## References

D. Russo and B. Van Roy, “An information-theoretic analysis of Thompson sampling,” *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2442–2471, 2016.

- O. Chapelle and L. Li, “An Empirical Evaluation of Thompson Sampling,” in *Advances in Neural Information Processing Systems*, vol. 24. Curran Associates, Inc., 2011. [Online]. Available: <https://proceedings.neurips.cc/paper/2011/hash/e53a0a2978c28872a4505bdb51db06dc-Abstract.html>
- D. Russo and B. Van Roy, “Learning to Optimize via Information-Directed Sampling,” *Operations Research*, vol. 66, no. 1, pp. 230–252, Feb. 2018. [Online]. Available: <http://pubsonline.informs.org/doi/10.1287/opre.2017.1663>
- S. Filippi, O. Cappe, A. Garivier, and C. Szepesvári, “Parametric Bandits: The Generalized Linear Case,” in *Advances in Neural Information Processing Systems*, vol. 23. Curran Associates, Inc., 2010. [Online]. Available: <https://proceedings.neurips.cc/paper/2010/hash/c2626d850c80ea07e7511bbae4c76f4b-Abstract.html>
- L. Li, Y. Lu, and D. Zhou, “Provably Optimal Algorithms for Generalized Linear Contextual Bandits,” in *Proceedings of the 34th International Conference on Machine Learning*. PMLR, Jul. 2017, pp. 2071–2080, iSSN: 2640-3498. [Online]. Available: <https://proceedings.mlr.press/v70/li17c.html>
- L. Faury, M. Abeille, C. Calauzènes, and O. Fercoq, “Improved optimistic algorithms for logistic bandits,” PMLR, pp. 3052–3060, 2020.
- D. Russo and B. Van Roy, “Learning to Optimize via Posterior Sampling,” *Mathematics of Operations Research*, vol. 39, no. 4, pp. 1221–1243, Nov. 2014. [Online]. Available: <https://pubsonline.informs.org/doi/10.1287/moor.2014.0650>
- S. Dong, T. Ma, and B. V. Roy, “On the Performance of Thompson Sampling on Logistic Bandits,” *Conference on Learning Theory*, 2019.
- M. Abeille and A. Lazaric, “Linear Thompson Sampling Revisited,” in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. PMLR, Apr. 2017, pp. 176–184, iSSN: 2640-3498. [Online]. Available: <https://proceedings.mlr.press/v54/abeille17a.html>
- H. B. McMahan and M. Streeter, “Open Problem: Better Bounds for Online Logistic Regression,” *Conference on Learning Theory*, 2012.
- 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, Dec. 1933. [Online]. Available: <https://doi.org/10.1093/biomet/25.3-4.285>
- D. J. Russo, B. V. Roy, A. Kazerouni, I. Osband, and Z. Wen, “A Tutorial on Thompson Sampling,” *Foundations and Trends® in Machine Learning*, vol. 11, no. 1, pp. 1–96, Jul. 2018, publisher: Now Publishers, Inc. [Online]. Available: <https://www.nowpublishers.com/article/Details/MAL-070>
- S. Dong and B. Van Roy, “An information-theoretic analysis for thompson sampling with many actions,” *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- B. Kveton, M. Zaheer, C. Szepesvari, L. Li, M. Ghavamzadeh, and C. Boutilier, “Randomized exploration in generalized linear bandits,” in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 2066–2076.
- G. Neu, I. Olkhovskaia, M. Papini, and L. Schwartz, “Lifting the Information Ratio: An Information-Theoretic Analysis of Thompson Sampling for Contextual Bandits,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 9486–9498, Dec. 2022. [Online]. Available: [https://proceedings.neurips.cc/paper\\_files/paper/2022/hash/3d84d9b523e6e82916d496e58761002e-Abstract-Conference.html](https://proceedings.neurips.cc/paper_files/paper/2022/hash/3d84d9b523e6e82916d496e58761002e-Abstract-Conference.html)
- A. Gouverneur, B. Rodríguez-Gálvez, T. J. Oechtering, and M. Skoglund, “Thompson sampling regret bounds for contextual bandits with sub-gaussian rewards,” *IEEE*, pp. 1306–1311, 2023.
- Y. W. Yury Polyanskiy, *Information Theory - From Coding to Learning*, 1st ed. Cambridge University Press, Oct. 2022.
- G. Neu, M. Papini, and L. Schwartz, “Optimistic Information Directed Sampling,” Jun. 2024, arXiv:2402.15411 [cs]. [Online]. Available: <http://arxiv.org/abs/2402.15411>
- J. Duchi, “Lecture notes for statistics 311/electrical engineering 377,” URL: [https://stanford.edu/class/stats311/Lectures/full notes. pdf](https://stanford.edu/class/stats311/Lectures/full%20notes.pdf). Last visited on, vol. 2, p. 23, 2016.
- R. van Handel, *Probability in High Dimension*. Princeton University, Dec. 2016, vol. APC 550 Lecture Notes.
- D. Russo and B. Van Roy, “Learning to Optimize via Information-Directed Sampling,” in *Advances in Neural Information Processing Systems*, vol. 27. Curran Associates, Inc., 2014. [Online]. Available: [https://proceedings.neurips.cc/paper\\_files/paper/2014/hash/301ad0e3bd5cb1627a2044908a42fdc2-Abstract.html](https://proceedings.neurips.cc/paper_files/paper/2014/hash/301ad0e3bd5cb1627a2044908a42fdc2-Abstract.html)
- R. M. Gray, *Entropy and Information Theory - First Edition, Corrected*. Information Systems Laboratory Electrical Engineering Department Stanford University: Springer-Verlag, 2013.

## Appendix

The appendix is organized as follows:

- Appendix A introduces lemmata that will be useful later for our proofs;
- Appendix B formalizes the proof for controlling the ratio of expected variances between the functions  $\varphi_\beta$  and  $\psi_\beta$ ;
- Appendix C extends our information ratio analysis to general action and parameter spaces;
- Appendix D illustrates the improvement of our bounds compared to previous regret guarantees through numerical experiments;
- Appendix E illustrates how applying directly linear bandits bounds to the logistic bandit setting leads to regret bounds scaling exponentially with  $\beta$ ;
- Appendix F elaborates on the gaps in the previous literature mentioned in Section 1;
- Appendix G rigorously explains the construction of the mapping  $\pi_\star$ .

### A Useful lemmata

**Lemma 7.** *Let  $U$  be a random variable taking values in  $[0, 1]$  and  $\text{Bern}(U)$  be a Bernoulli random variable with probability  $U$ . Then it holds that,*

$$I(U; \text{Bern}(U)) \geq 2\mathbb{V}(U).$$

*Proof.* Using Yury Polyanskiy (2022, Theorem 3.4.d), we decompose the mutual information between  $U$  and  $\text{Bern}(U)$  as

$$I(U; \text{Bern}(U)) = h(\text{Bern}(U)) - h(\text{Bern}(U)|U).$$

Following Duchi (2016, Example 2.2) notation, we define  $h_2(p) := -p \log(p) - (1-p) \log(1-p)$  for  $p \in [0, 1]$  and rewrite the mutual information as

$$I(U; \text{Bern}(U)) = h_2(\mathbb{E}[U]) - \mathbb{E}[h_2(U)]. \quad (3)$$

From a Taylor expansion of  $h_2(x)$  we have that

$$h_2(x) = h_2(p) + (x-p)h_2'(p) + \frac{1}{2}(x-p)^2 h_2''(\xi),$$

for some  $\xi \in (0, 1)$  as  $h_2''$  is continuous on the interval  $[0, 1]$ . We compute the second derivative of  $h_2$  and get  $h_2''(\xi) = -\frac{1}{\xi(1-\xi)}$  for  $\xi \in (0, 1)$ . This function is concave and maximal at  $\xi = 1/2$ , where it takes the value  $h_2''(1/2) = -4$ . We then have that for all  $x \in [0, 1]$  and all  $p \in [0, 1]$ ,

$$h_2(x) \leq h_2(p) + (x-p)h_2'(p) - 2(x-p)^2.$$

Using this fact for  $x = U$  and  $p = \mathbb{E}[U]$ , we have that

$$h_2(U) \leq h_2(\mathbb{E}[U]) + (U - \mathbb{E}[U])h_2'(\mathbb{E}[U]) - 2(U - \mathbb{E}[U])^2.$$

Applying the last inequality to the second term in eq. (3), it comes that

$$I(U; \text{Bern}(U)) \geq \mathbb{E} [h_2(\mathbb{E}[U]) - h_2(\mathbb{E}[U]) - (U - \mathbb{E}[U])h_2'(\mathbb{E}[U]) + 2(U - \mathbb{E}[U])^2].$$

Finally, simplifying terms and taking the expectation gives the desired result.  $\square$

The two following lemmata are be particularly useful to control the covering number in Euclidean balls and spheres.

**Lemma 14** (van Handel (2016, Lemma 5.13)). *Let  $\mathbf{B}_d(0, 1)$  denote the  $d$ -dimensional closed Euclidean unit ball. We have  $|\mathcal{N}(\mathbf{B}_d(0, 1), \|\cdot\|_2, \varepsilon)| = 1$  for  $\varepsilon \geq 1$  and for  $0 < \varepsilon < 1$ , we have*

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

**Lemma 15** (Yury Polyanskiy (2022, Corollary 27.4)). *Let  $\mathbf{S}_d(0, 1)$  denote the  $d$ -dimensional Euclidean unit sphere. We have  $|\mathcal{N}(\mathbf{S}_d(0, 1), \|\cdot\|_2, \varepsilon)| = 1$  for  $\varepsilon \geq 1$  and for  $0 < \varepsilon < 1$ , we have*

$$\left(\frac{1}{2\varepsilon}\right)^{d-1} \leq |\mathcal{N}(\mathbf{S}_d(0, 1), \|\cdot\|_2, \varepsilon)| \leq 2d \left(1 + \frac{1}{\varepsilon}\right)^{d-1}.$$

## B Bounding the ratio of expected variances over the functions $\varphi_\beta$ and $\psi_\beta$

**Lemma 13.** *Let  $\bar{\psi}$  and  $\bar{\varphi}$  be defined as in eq. (2). Then, it holds that*

$$\frac{\mathbb{E} \left[ \mathbb{V} \left[ \bar{\varphi} \left( 1 - \langle \hat{A}, \Theta \rangle \right) \mid \hat{\Theta} \right] \right]}{\mathbb{E} \left[ \mathbb{V} \left[ \bar{\psi} \left( 1 - \langle \hat{A}, \Theta \rangle \right) \mid \hat{\Theta} \right] \right]} \leq 9.$$

*Proof.* We start by analyzing  $\mathbb{E} \left[ \mathbb{V} \left[ \bar{\psi} \left( 1 - \langle \hat{A}, \Theta \rangle \right) \mid \hat{\Theta} \right] \right]$ . We note that  $\bar{\psi} \left( 1 - \langle \hat{A}, \Theta \rangle \right)$  is equal to 1 if  $\langle \hat{A}, \Theta \rangle < 0$  and is equal to 0 otherwise. To distinguish those two cases, we introduce the notation  $I(\langle \hat{A}, \Theta \rangle) := \mathbb{1}_{\{\langle \hat{A}, \Theta \rangle < 0\}}$ . We observe that  $\mathbb{E} \left[ \mathbb{V} \left[ \bar{\psi} \left( 1 - \langle \hat{A}, \Theta \rangle \right) \mid \hat{\Theta} \right] \right]$  is equal to the expected variance of a Bernoulli random variable with probability given by  $Q(\hat{A}) := \mathbb{E}[I(\langle \hat{A}, \Theta \rangle)]$  and can therefore be written as  $\mathbb{E} \left[ \mathbb{V} \left[ \bar{\psi} \left( 1 - \langle \hat{A}, \Theta \rangle \right) \mid \hat{\Theta} \right] \right] = \mathbb{E}[Q(\hat{A})(1 - Q(\hat{A}))]$ .

The last part of the proof concerns  $\mathbb{E} \left[ \mathbb{V} \left[ \bar{\varphi} \left( 1 - \langle \hat{A}, \Theta \rangle \right) \mid \hat{\Theta} \right] \right]$ . Similarly, we can distinguish between two cases: either  $\langle \hat{A}, \Theta \rangle > 0$  and  $\bar{\varphi}(1 - \langle \hat{A}, \Theta \rangle) = 0$ , or  $\langle \hat{A}, \Theta \rangle < 0$  and  $\bar{\varphi}(1 - \langle \hat{A}, \Theta \rangle) = 1 - 2\langle \hat{A}, \Theta \rangle$ . Introducing the notation  $G(\hat{A}) := \mathbb{E}[I(\langle \hat{A}, \Theta \rangle)\langle \hat{A}, \Theta \rangle]$ , we can write

$$\begin{aligned} \mathbb{E} \left[ \mathbb{V} \left[ \bar{\varphi} \left( 1 - \langle \hat{A}, \Theta \rangle \right) \mid \hat{\Theta} \right] \right] &= \mathbb{E} \left[ \mathbb{E} \left[ \left( \bar{\varphi} \left( 1 - \langle \hat{A}, \Theta \rangle \right) - \mathbb{E} \left[ \bar{\varphi} \left( 1 - \langle \hat{A}, \Theta \rangle \right) \mid \hat{\Theta} \right] \right)^2 \mid \hat{\Theta} \right] \right] \\ &= \mathbb{E} \left[ I(\langle \hat{A}, \Theta \rangle) \left( 1 - 2\langle \hat{A}, \Theta \rangle - (Q(\hat{A}) + 2G(\hat{A})) \right)^2 \right] \\ &\quad + \mathbb{E} \left[ \left( 1 - I(\langle \hat{A}, \Theta \rangle) \right) \left( 0 - (Q(\hat{A}) + 2G(\hat{A})) \right)^2 \right]. \end{aligned}$$

Distributing the square and simplifying terms, we obtain

$$\begin{aligned} &\mathbb{E} \left[ I(\langle \hat{A}, \Theta \rangle) \left( 1 - 2\langle \hat{A}, \Theta \rangle \right)^2 \right] - 2\mathbb{E} \left[ I(\langle \hat{A}, \Theta \rangle) \left( 1 - 2\langle \hat{A}, \Theta \rangle \right) (Q(\hat{A}) + 2G(\hat{A})) \right] \\ &+ \mathbb{E} \left[ I(\langle \hat{A}, \Theta \rangle) (Q(\hat{A}) + 2G(\hat{A}))^2 \right] + \mathbb{E} \left[ \left( 1 - I(\langle \hat{A}, \Theta \rangle) \right) (Q(\hat{A}) + 2G(\hat{A}))^2 \right] \\ &= \mathbb{E} \left[ I(\langle \hat{A}, \Theta \rangle) \left( 1 - 2\langle \hat{A}, \Theta \rangle \right)^2 \right] - \mathbb{E} \left[ (Q(\hat{A}) + 2G(\hat{A}))^2 \right]. \end{aligned}$$

To get to the last part of the proof, we rewrite explicitly  $Q(\hat{A}) + 2G(\hat{A})$  as  $\mathbb{E} \left[ I(\langle \hat{A}, \Theta \rangle) \left( 1 - 2\langle \hat{A}, \Theta \rangle \right) \right]$  and optimize over the values of  $(1 - 2\langle \hat{A}, \Theta \rangle)$ . It then comes

$$\begin{aligned} &\mathbb{E} \left[ I(\langle \hat{A}, \Theta \rangle) \left( 1 - 2\langle \hat{A}, \Theta \rangle \right)^2 \right] - \mathbb{E} \left[ \mathbb{E} \left[ I(\langle \hat{A}, \Theta \rangle) \left( 1 - 2\langle \hat{A}, \Theta \rangle \right) \right]^2 \right] \\ &\leq \sup_{\zeta \in [-1, 3]} \mathbb{E} \left[ I(\langle \hat{A}, \Theta \rangle) \zeta^2 \right] - \mathbb{E} \left[ \mathbb{E} \left[ I(\langle \hat{A}, \Theta \rangle) \zeta \right]^2 \right] = 9 \cdot \mathbb{E}[Q(\hat{A})(1 - Q(\hat{A}))], \end{aligned}$$

which concludes the proof.  $\square$

**Lemma 16.** *Let  $\psi_\beta(x) = \phi_\beta(1) - \phi(1 - x)$  and the logistic surrogate  $\varphi_\beta$  as in Definition 11 and let  $\bar{\psi}$  and  $\bar{\varphi}$  be defined as in eq. (2). Then, for all  $\beta > 0$ , it holds that*

$$\frac{\mathbb{E} \left[ \mathbb{V} \left[ \varphi_\beta \left( 1 - \langle \hat{A}, \Theta \rangle \right) \mid \Theta \right] \right]}{\mathbb{E} \left[ \mathbb{V} \left[ \psi_\beta \left( 1 - \langle \hat{A}, \Theta \rangle \right) \mid \hat{\Theta} \right] \right]} \leq \frac{\mathbb{E} \left[ \mathbb{V} \left[ \bar{\varphi} \left( 1 - \langle \hat{A}, \Theta \rangle \right) \mid \hat{\Theta} \right] \right]}{\mathbb{E} \left[ \mathbb{V} \left[ \bar{\psi} \left( 1 - \langle \hat{A}, \Theta \rangle \right) \mid \hat{\Theta} \right] \right]}.$$

*Proof.* Beginning with the ratio of expected variances between  $\varphi_\beta$  and  $\psi_\beta$ , we will apply a series of transformations to the functions  $\varphi_\beta$  and  $\psi_\beta$ , ultimately yielding the functions  $\bar{\varphi}$  and  $\bar{\psi}$ . These transformations are chosen to ensure they can only increase the ratio of expected variances.

By definition, the function  $\psi_\beta$  and its surrogate  $\varphi_\beta$  are identical for  $x \in [0, \delta_\beta]$  and then diverge linearly at a rate of  $\psi_\beta(\delta_\beta)/\delta_\beta$  on the interval  $x \in [\delta_\beta, 2]$ . We illustrate this on Figure 3. Focusing on the domain where the two functions coincide, we observe that the transformation  $f(x) = \max(x, \psi_\beta(1))$  reduces the expected variance for both  $\psi_\beta$  and  $\varphi_\beta$ . However, since  $\psi_\beta(x)$  is less than or equal to  $\varphi_\beta(x)$  for all  $x \in [0, 2]$ , and both functions exceed  $\psi_\beta(1)$  on the interval  $[1, 2]$ , the transformation  $f$  proportionally reduces the expected variance of  $\psi_\beta$  more than that of  $\varphi_\beta$ . As a result, the transformation increases the ratio of expected variances between the two functions. As  $\psi_\beta$  and  $\varphi_\beta$  are strictly increasing functions, the resulting functions, illustrated on Figure 4, can be written as

$$f(\psi_\beta(x)) = \begin{cases} \psi_\beta(1) & x \in [0, 1] \\ \psi_\beta(x) & x \in ]1, 2] \end{cases}, \quad \text{and} \quad f(\varphi_\beta(x)) = \begin{cases} \psi_\beta(1) & x \in [0, 1] \\ \varphi_\beta(x) & x \in ]1, 2] \end{cases}.$$

The second transformation we apply concerns only the function  $f(\psi_\beta(x))$ . We crop all the values larger than  $\psi(\delta_\beta)$  by applying the transformation  $g(x) = \min(x, \psi(\delta_\beta))$ . As  $f(\psi_\beta(x))$  is an increasing function, the function  $g(f(\psi_\beta(x)))$ , illustrated on Figure 5, can be written as

$$g(f(\psi_\beta(x))) = \begin{cases} \psi_\beta(1) & x \in [0, 1] \\ \psi_\beta(x) & x \in ]1, \delta_\beta] \\ \psi_\beta(\delta_\beta) & x \in ]\delta_\beta, 2] \end{cases}.$$

The transformation  $g$  reduces the variance of the function  $f(\psi_\beta(x))$  as it both decreases the values of  $f(\psi_\beta(x))$  and the derivative of  $f(\psi_\beta(x))$  for all  $x \in ]\delta_\beta, 2]$ .

The third transformation we apply is increasing the value of  $\beta$ . As  $\beta$  increases, the derivative of  $f(\varphi_\beta(x))$  increases everywhere,

$$\frac{d}{dx} f(\varphi_\beta(x)) = \begin{cases} 0 & x \in [0, 1] \\ \frac{\beta \exp(-\beta(1-x))}{(1+\exp(-\beta(1-x)))^2} & x \in ]1, \delta_\beta] \\ \psi_\beta(\delta_\beta)/\delta_\beta & x \in ]\delta_\beta, 2] \end{cases}.$$

and the expected variance of  $f(\varphi_\beta)$  increases. Regarding  $g(f(\psi_\beta(x)))$ , we can show that that for all  $x \in [0, 2]$ , the ratio  $f(\varphi_\beta(x))/g(f(\psi_\beta(x)))$  increases with  $\beta$ . Indeed, this ratio is equal to 1 for all  $x \in [0, \delta_\beta]$  and increases for all  $x \in ]\delta_\beta, 2]$  as

$$\frac{f(\varphi_\beta(x))}{g(f(\psi_\beta(x)))} = \frac{\varphi_\beta(\delta_\beta) + \varphi_\beta(\delta_\beta)/\delta_\beta \cdot (x - \delta_\beta)}{\varphi_\beta(\delta_\beta)} = 1 + \frac{(x - \delta_\beta)}{\delta_\beta},$$

and as  $\delta_\beta$  is a decreasing function of  $\beta$  (see Remark 17), the ratio  $(x - \delta_\beta)/\delta_\beta$  is a increasing function of  $\beta$  for all  $x \in ]\delta_\beta, 2]$ . This fact ensures that the expected variance of  $g(f(\psi_\beta(x)))$  cannot increase proportionally more than the expected variance of  $f(\varphi_\beta(x))$ . We can therefore study the ratio of expected variances between  $f(\varphi_\infty)$  and  $g(f(\psi_\infty))$ .

The last operation we apply is a convenient shifting and scaling. We define  $h(x) = (x - g(f(\psi_\beta(1))))/(g(f(\psi_\infty(2))) - g(f(\psi_\beta(1))))$  and apply it on both  $g(f(\psi_\infty))$  and  $f(\varphi_\infty)$  these operations do not affect the ratio of expected variances. The resulting functions are illustrated on Figure 6.

To express the resulting functions, we have to analyze the function  $\psi_\beta(x)$  for  $\beta$  tending to infinity for values  $x \in ]1, 2]$ .

We recall that  $\psi_\beta(x) = \phi_\beta(1) - \phi_\beta(1 - x)$  and can equivalently be written as

$$\psi_\beta(x) = \frac{1}{1 + \exp(-\beta)} - \frac{1}{1 + \exp(-\beta(1 - x))}.$$

We have to distinguish between three cases for  $(x - 1)$ : negative, zero, or positive. For values of  $x \in ]1, 2]$ , we have that  $(1 - x) < 0$  and that  $\lim_{\beta \rightarrow \infty} \psi_\beta(x) = 1$ , if  $x = 1$ , we have that  $\lim_{\beta \rightarrow \infty} \psi_\beta(x) = 1/2$  and for values of  $x \in [0, 1[$ , we have that  $(1 - x) > 0$  and that  $\lim_{\beta \rightarrow \infty} \psi_\beta(x) = 0$ . We can then write

$$\psi_\infty(x) = \begin{cases} 0 & x \in [0, 1[ \\ 1/2 & x = 1 \\ 1 & x \in ]1, 2] \end{cases}.$$

We can now construct the corresponding  $\varphi_\infty(x)$ . We note that  $\frac{\psi_\infty(x)}{x}$  is maximized when taking the limit to  $x = 1^+$  from the right:  $\lim_{x \rightarrow 1^+} \frac{\psi_\infty(x)}{x} = 1$ . It comes that  $\varphi_\infty(x)$  can be written as

$$\varphi_\infty(x) = \begin{cases} 0 & x \in [0, 1[ \\ 1/2 & x = 1 \\ 1 + (x - 1) & x \in ]1, 2] \end{cases}.$$

We denote the resulting functions  $h(g(f(\psi_\infty(x))))$  and  $h(f(\varphi_\infty(x)))$  respectively as  $\bar{\psi}$  and  $\bar{\varphi}$ . We note that they can be written quite simply as

$$\bar{\psi}(x) = \begin{cases} 0 & x \in [0, 1] \\ 1 & x \in ]1, 2] \end{cases}, \quad \text{and} \quad \begin{cases} 0 & x \in [0, 1] \\ 1 + 2(x - 1) & x \in ]1, 2] \end{cases}.$$

□

**Remark 17.** We illustrate the function  $\psi_\beta(x)/x$  on Figure 1 and the behavior of  $\delta_\beta$  and  $\psi_\beta(\delta_\beta)/\delta_\beta$  for increasing values of  $\beta$  on Figure 2. The derivative of the function  $\psi_\beta(x)/x$  is given by

$$\frac{d}{dx} \left( \frac{\psi_\beta(x)}{x} \right) = \frac{1}{x} \left( \frac{d}{dx} \psi_\beta(x) - \frac{\psi_\beta(x)}{x} \right).$$

We note that it is equal to zero for values of  $x \in ]0, 2]$  such that  $\frac{d}{dx} \psi_\beta(x) = \frac{\psi_\beta(x)}{x}$ . By definition of  $\delta_\beta$ , we have  $\frac{d}{dx} \psi_\beta(\delta_\beta) = \frac{\psi_\beta(\delta_\beta)}{\delta_\beta}$ .

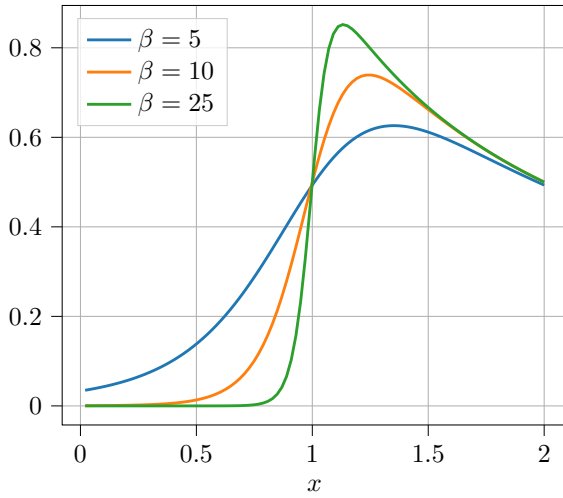


Figure 1: Illustration of the function  $\psi_\beta(x)/x$  for different values of  $x$ . The maximum of the function is attained for  $x = \delta_\beta$ .

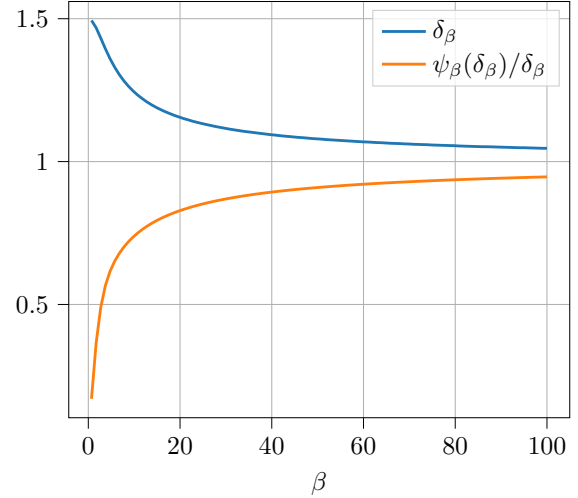


Figure 2: Illustration of  $\delta_\beta$  and  $\psi_\beta(\delta_\beta)/\delta_\beta$  as functions of  $\beta$ . One can observe that  $\delta_\beta$  decreases with  $\beta$  while  $\psi_\beta(\delta_\beta)/\delta_\beta$  increases.

## C Extension to general spaces

To extend the proof technique of Section 5.2 and Section 5.3, we first need to introduce the *alignment function*  $\alpha(\theta) := \langle \pi_\star(\theta), \theta \rangle$ . We can define the *extended logistic function*  $\psi_\beta(x, \theta) := \phi_\beta(\alpha(\theta)) - \phi_\beta(\alpha(\theta) - x)$  and note that

$$\psi_\beta(\alpha(\hat{\Theta}) - \langle \hat{A}, \Theta \rangle, \hat{\Theta}) = \phi_\beta(\alpha(\hat{\Theta})) - \phi_\beta(\langle \hat{A}, \Theta \rangle) = \phi_\beta(\langle \hat{A}, \hat{\Theta} \rangle) - \phi_\beta(\langle \hat{A}, \Theta \rangle).$$

Integrating the randomness of the Bernoulli process, we can write the expected regret using  $\psi_\beta(x, \theta)$ :

$$\mathbb{E}[\psi_\beta(\alpha(\hat{\Theta}) - \langle \hat{A}, \Theta \rangle, \hat{\Theta})] = \mathbb{E}[\phi_\beta(\langle \hat{A}, \hat{\Theta} \rangle) - \phi_\beta(\langle \hat{A}, \Theta \rangle)] = \mathbb{E}[\phi_\beta(\langle A^\star, \Theta \rangle) - \phi_\beta(\langle \hat{A}, \Theta \rangle)],$$

where we used the fact that the pair  $(A^\star, \Theta)$  and  $(\hat{A}, \hat{\Theta})$  are identically distributed.

Similarly to the proof in Section 5.2, we construct a function  $\varphi_\beta(x, \theta)$  as the tightest upper bound on  $\psi_\beta(x, \theta)$  that satisfies the requirements of Lemma 10.

**Definition 18** (Extended logistic surrogate). *We construct the extended logistic surrogate function  $\varphi_\beta(x, \theta)$  as the tightest upper bound on  $\psi_\beta(x, \theta)$  such that  $\varphi_\beta(x, \theta)/x$  is non-decreasing over  $x \geq 0$  for all  $\theta \in \mathcal{O}$ .*

Namely, let  $\delta_\beta(\theta) = \arg \max_{x \in [0, 1 + \alpha(\theta)]} \frac{\psi_\beta(x, \theta)}{x}$ , we define the function  $\varphi_\beta(x, \theta)$  as

$$\varphi_\beta(x, \theta) = \begin{cases} \psi_\beta(x, \theta) & x \in [0, \delta_\beta(\theta)] \\ \psi_\beta(\delta_\beta(\theta), \theta) + (x - \delta_\beta(\theta)) \cdot \psi_\beta(\delta_\beta(\theta), \theta) / \delta_\beta(\theta) & x \in ]\delta_\beta(\theta), 1 + \alpha(\theta)] \end{cases}.$$

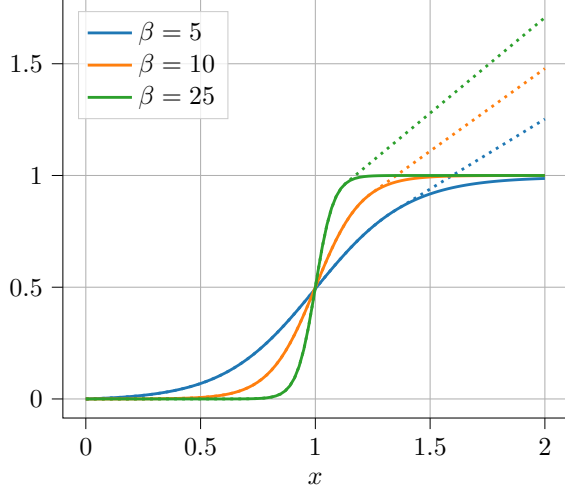


Figure 3: Illustration of the function  $\psi_\beta$  (in solid line) and the function  $\varphi_\beta$  (in dotted line) for different values of  $\beta$ .

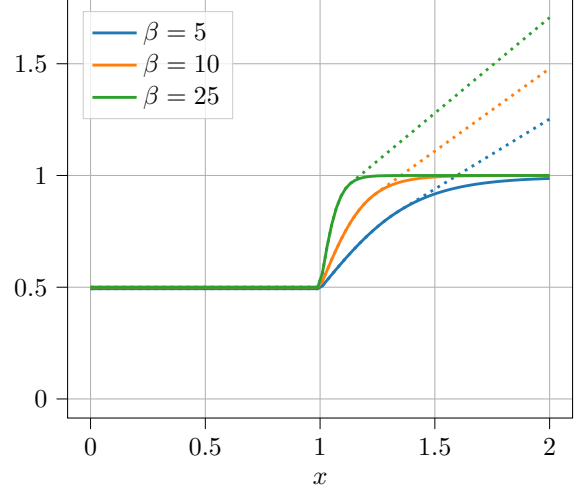


Figure 4: Illustration of the function  $f(\psi_\beta)$  (in solid line) and the function  $f(\varphi_\beta)$  (in dotted line) for different values of  $\beta$ .

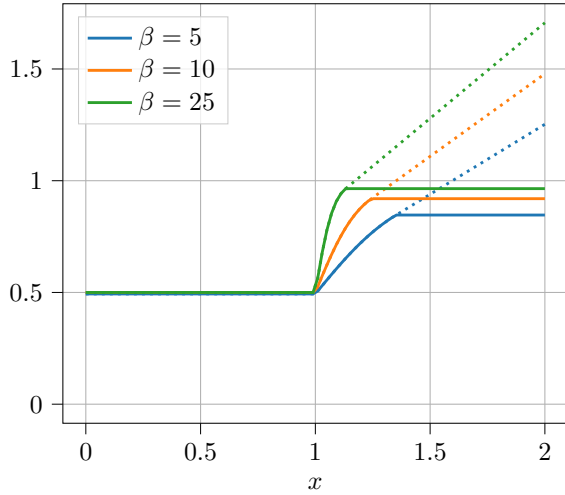


Figure 5: Illustration of the function  $g(f(\psi_\beta))$  (in solid line) and the function  $f(\varphi_\beta)$  (in dotted line) for different values of  $\beta$ .

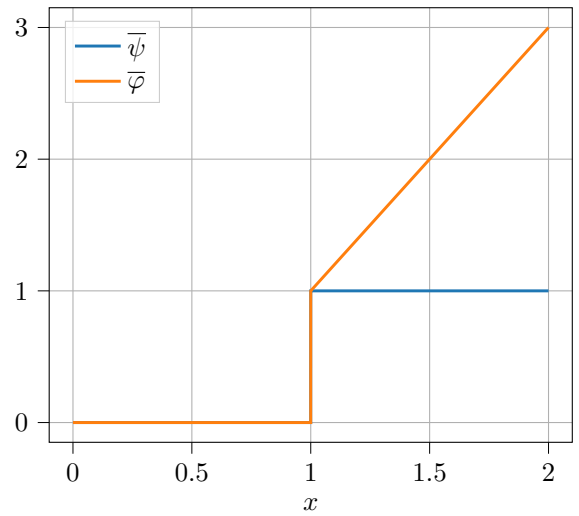


Figure 6: Illustration of the function  $\bar{\psi}$  (in blue) and the function  $\bar{\varphi}$  (in orange).

We are now equipped to extend Lemma 12 to general action and parameter spaces.

**Lemma 19.** *Let the extended logistic surrogate be defined as in Definition 18. Then, it holds that*

$$\mathbb{E}[\text{Bern}(\phi_\beta(\langle \hat{A}, \Theta \rangle)) - \text{Bern}(\phi_\beta(\langle \hat{A}, \Theta \rangle))]^2 \leq d \cdot \mathbb{E} \left[ \mathbb{V} \left[ \varphi_\beta(\alpha(\hat{\Theta}) - \langle \hat{A}, \Theta \rangle, \hat{\Theta}) \mid \hat{\Theta} \right] \right].$$

*Proof.* The proof follows closely the technique used to prove Lemma 12. We note that conditioned on  $\hat{\Theta} = \theta$ , the extended logistic surrogate is a mapping from  $[0, 1 + \delta_\beta(\theta)]$  to  $[0, 1]$ , that  $\varphi(0, \theta) = \phi_\beta(\alpha(\theta)) - \phi_\beta(\alpha(\theta)) = 0$  and fulfills the assumptions of Lemma 10.  $\square$

Noting that  $\mathbb{V}[\phi_\beta(\langle \hat{A}, \Theta \rangle) \mid \hat{\Theta}] = \mathbb{V}[\psi_\beta(\alpha(\hat{\Theta}) - \langle \hat{A}, \Theta \rangle, \hat{\Theta}) \mid \hat{\Theta}]$  and using Lemma 8, we have that the information ration  $\Gamma$  can be bounded by

$$\Gamma \leq d/2 \cdot \frac{\mathbb{E}[\mathbb{V}[\varphi_\beta(\alpha(\hat{\Theta}) - \langle \hat{A}, \Theta \rangle, \hat{\Theta}) \mid \hat{\Theta}]]}{\mathbb{E}[\mathbb{V}[\psi_\beta(\alpha(\hat{\Theta}) - \langle \hat{A}, \Theta \rangle, \hat{\Theta}) \mid \hat{\Theta}]]}.$$

Similarly to the analysis for the  $\mathcal{O} \subseteq \mathcal{A}$ , we can derive an upper bound by studying the case  $\beta \rightarrow \infty$  after applying the same preliminary transformations,  $f$ ,  $g$ , and  $h$  as used in Lemma 16 on the functions  $\varphi_\beta(x, \theta)$  and  $\psi_\beta(x, \theta)$ . To express the resulting functions  $h(g(f(\psi_\infty(x))))$  and  $h(f(\varphi_\infty(x)))$ , we need to study the extended logistic function  $\psi_\beta(x, \theta)$  and the corresponding extended logistic surrogate  $\varphi_\beta(x, \theta)$  for  $\beta$  tending to infinity.

Starting with  $\psi_\beta(x, \theta)$ , we recall that  $\psi_\beta(x, \theta) = \phi_\beta(\alpha(\theta)) - \phi_\beta(\alpha(\theta) - x)$  can be written as

$$\psi_\beta(x, \theta) = \frac{1}{1 + \exp(-\beta\alpha(\theta))} - \frac{1}{1 + \exp(-\beta(\alpha(\theta) - x))}.$$

Again, we can distinguish between three cases for  $(\alpha(\theta) - x)$ : negative, zero, or positive. For values of  $x \in ]\alpha(\theta), 1 + \alpha(\theta)]$ , we have that  $(\alpha(\theta) - x) < 0$  and that  $\lim_{\beta \rightarrow \infty} \psi_\beta(x, \theta) = 1$ , if  $x = \alpha(\theta)$ , we have that  $\lim_{\beta \rightarrow \infty} \psi_\beta(x, \theta) = 1/2$  and for values of  $x \in [0, \alpha(\theta)[$ , we have that  $(\alpha(\theta) - x) > 0$  and that  $\lim_{\beta \rightarrow \infty} \psi_\beta(x, \theta) = 0$ . We can then write

$$\psi_\infty(x, \theta) = \begin{cases} 0 & x \in [0, \alpha(\theta)[ \\ 1/2 & x = \alpha(\theta) \\ 1 & x \in ]\alpha(\theta), 1 + \alpha(\theta)] \end{cases}.$$

We continue and construct the corresponding  $\varphi_\infty(x, \theta)$ . By definition, we have that  $\alpha(\theta) \leq 1$  and we note that  $\frac{\psi_\infty(x, \theta)}{x}$  is maximized when taking the limit to  $x = \alpha(\theta)^+$  from the right:  $\lim_{x \rightarrow \alpha(\theta)^+} \frac{\psi_\infty(x)}{x} = \frac{1}{\alpha(\theta)}$ . It comes that  $\varphi_\infty(x)$  can be written as

$$\varphi_\infty(x) = \begin{cases} 0 & x \in [0, \alpha(\theta)[ \\ 1/2 & x = \alpha(\theta) \\ 1 + \frac{x - \alpha(\theta)}{\alpha(\theta)} & x \in ]\alpha(\theta), 1 + \alpha(\theta)] \end{cases}.$$

We can now construct the functions  $\bar{\psi}(x, \theta) := h(g(f(\psi_\infty(x, \theta))))$  and  $\bar{\varphi}(x, \theta) := h(f(\varphi_\infty(x, \theta)))$  for  $f$ ,  $g$ , and  $h$  defined as in Lemma 16. We note that the resulting functions can be written as

$$\bar{\psi}(x, \theta) = \begin{cases} 0 & x \in [0, \alpha(\theta)] \\ 1 & x \in ]\alpha(\theta), 1 + \alpha(\theta)] \end{cases}, \quad \text{and} \quad \bar{\varphi}(x, \theta) = \begin{cases} 0 & x \in [0, \alpha(\theta)] \\ 1 + \frac{2(x - \alpha(\theta))}{\alpha(\theta)} & x \in ]\alpha(\theta), 1 + \alpha(\theta)] \end{cases}, \quad (4)$$

a similar form as the functions  $\bar{\psi}$  and  $\bar{\varphi}$  be defined as in (2).

**Lemma 20.** *Let  $\bar{\psi}(x, \theta)$  and  $\bar{\varphi}(x, \theta)$  be defined as in (4). Then, it holds that*

$$\frac{\mathbb{E}[\mathbb{V}[\bar{\varphi}(\alpha(\hat{\Theta}) - \langle \hat{A}, \Theta \rangle, \hat{\Theta}) \mid \hat{\Theta}]]}{\mathbb{E}[\mathbb{V}[\bar{\psi}(\alpha(\hat{\Theta}) - \langle \hat{A}, \Theta \rangle, \hat{\Theta}) \mid \hat{\Theta}]]} \leq \frac{9}{\alpha}.$$

*Proof.* The proof follows the one for Lemma 13. Starting with  $\mathbb{E}[\mathbb{V}[\bar{\psi}(\alpha(\hat{\Theta}) - \langle \hat{A}, \Theta \rangle, \hat{\Theta}) \mid \hat{\Theta}]]$ , we note that  $\bar{\psi}(\alpha(\hat{\Theta}) - \langle \hat{A}, \Theta \rangle, \hat{\Theta})$  is equal to 1 if  $\langle \hat{A}, \Theta \rangle < 0$  and is equal to 0 otherwise. It can then be written as  $\mathbb{E}[Q(\hat{A})(1 - Q(\hat{A}))]$  using the notations from the proof of Lemma 13. Similarly for  $\mathbb{E}[\mathbb{V}[\bar{\varphi}(\alpha(\hat{\Theta}) - \langle \hat{A}, \Theta \rangle, \hat{\Theta}) \mid \hat{\Theta}]]$ , we distinguish two cases: either  $\langle \hat{A}, \Theta \rangle \geq 0$  and the function is equal to 0 or  $\langle \hat{A}, \Theta \rangle < 0$  and the function is equal to  $1 - 2\frac{\langle \hat{A}, \Theta \rangle}{\langle \hat{A}, \hat{\Theta} \rangle}$ . Using a similar decomposition as in the proof of Lemma 13, we write  $\mathbb{E}[\mathbb{V}[\bar{\varphi}(\alpha(\hat{\Theta}) - \langle \hat{A}, \Theta \rangle, \hat{\Theta}) \mid \hat{\Theta}]]$  as

$$\mathbb{E}\left[I(\langle \hat{A}, \Theta \rangle) \left(1 - 2\frac{\langle \hat{A}, \Theta \rangle}{\langle \hat{A}, \hat{\Theta} \rangle}\right)^2\right] - \mathbb{E}\left[\mathbb{E}\left[I(\langle \hat{A}, \Theta \rangle) \left(1 - 2\frac{\langle \hat{A}, \Theta \rangle}{\langle \hat{A}, \hat{\Theta} \rangle}\right)\right]^2\right],$$



where in the second term, the outer expectation is on  $\hat{A}, \hat{\Theta}$ , and the inner expectation is on  $\Theta$ . Then taking the supremum over the possible values of  $1 - 2\frac{\langle \hat{A}, \Theta \rangle}{\langle \hat{A}, \hat{\Theta} \rangle}$  which ranges from  $[1 - 2/\alpha, 1 + 2/\alpha]$  we get:

$$\begin{aligned} \mathbb{E} \left[ \mathbb{V} \left[ \bar{\varphi} \left( \alpha(\hat{\Theta}) - \langle \hat{A}, \Theta \rangle, \hat{\Theta} \right) \mid \hat{\Theta} \right] \right] &\leq \sup_{\zeta \in [1-2/\alpha, 1+2/\alpha]} \mathbb{E} \left[ I(\langle \hat{A}, \Theta \rangle) \zeta^2 \right] - \mathbb{E} \left[ \mathbb{E} \left[ I(\langle \hat{A}, \Theta \rangle) \zeta \right]^2 \right] \\ &= (1 + 2/\alpha)^2 \cdot \mathbb{E}[Q(\hat{A})(1 - Q(\hat{A}))]. \end{aligned}$$

Finally, as  $\alpha \in [0, 1]$ , we can upper bound  $(1 + 2/\alpha)^2$  by  $9\alpha^{-2}$  and we conclude the proof.  $\square$

## D Numerical simulations

To illustrate the improvement of our regret analysis compared to previous work, we performed numerical experiments on a synthetic problem. We considered a logistic bandit problem in dimension  $d = 10$ , with time horizon  $T = 200$ , and with parameter  $\beta$  ranging between  $[0.25, 10]$ . For both action space and parameter space, we used the closed  $d$ -dimensional unit sphere,  $\mathcal{A} = \mathcal{O} = \mathbf{S}_d(0, 1)$  and assumed a uniform prior distribution for the parameter  $\Theta$ . We computed the expected regret of the Thompson Sampling algorithm using an MCMC method and compared it to three Bayesian regret bounds that hold for continuous spaces: our Corollary 5, Russo and Van Roy (2014b, Proposition 4), and Dong et al. (2019, Proposition 17) which can be adapted to be compatible with Dong and Van Roy (2018, Theorem 1) (see Appendix E).

The results are presented in Figure 7. The left sub-figure shows the evolution of the expected regret and the regret bounds for two different values of  $\beta = \{2, 4\}$ . For both values of  $\beta$ , our bound is tighter throughout the entire time horizon and is less sensitive to increasing  $\beta$  compared to Russo and Van Roy (2014b, Proposition 4) and Proposition 21 adapting Dong and Van Roy (2018, Theorem 2). This behavior is illustrated in the right sub-figure, where the different regret bounds at  $t = 200$  are compared for values of  $\beta$  ranging between  $[0.25, 10]$ . Our bound remains competitive across all values of  $\beta$  and is tighter than Dong et al. (2019, Proposition 17) for values of  $\beta \geq 2$ . Importantly, we observe that while our bound increases only logarithmically, both Russo and Van Roy (2014b, Proposition 4) and Dong et al. (2019, Proposition 17) increase exponentially with  $\beta$ .

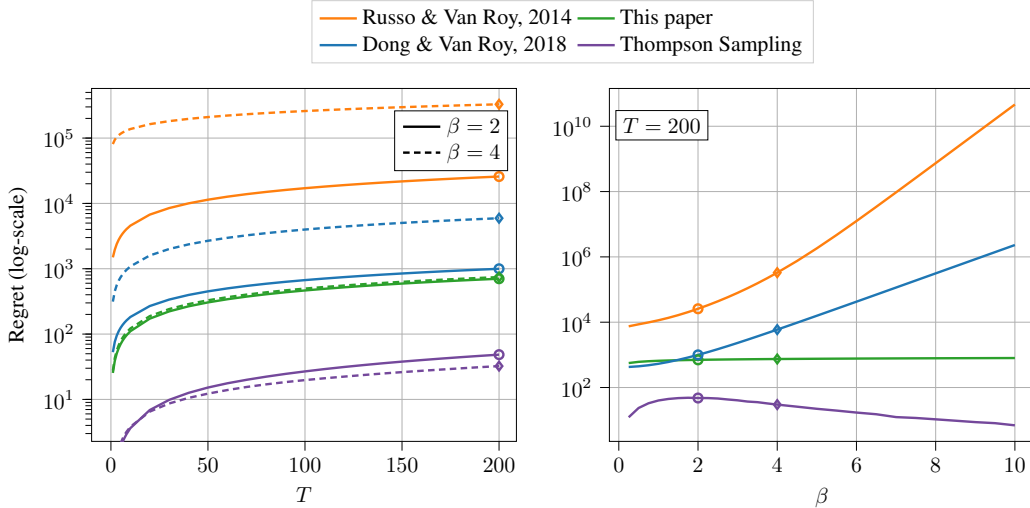


Figure 7: Comparison of Bayesian regret bounds for the logistic bandit setting ( $d = 10$ ,  $T = 200$ ,  $\beta \in [0.25, 10]$ ,  $\mathcal{A} = \mathcal{O} = \mathbf{S}_d(0, 1)$ , uniform prior on  $\Theta$ ). The left sub-figure compares the evolution of the bounds and the expected regret for  $\beta \in \{2, 4\}$ . The right sub-figure illustrates the behavior of the bounds and the expected regret at time  $T = 200$  for values of  $\beta$  ranging in  $[0.25, 10]$ .

## E Adapting information ratio bounds from linear to logistic bandits

This section presents one way to rigorously adapt the TS information bound from Dong et al. (2019, Proposition 17) to the *one-step compressed Thompson Sampling* such that it can be combined with the regret bound from Dong and Van Roy (2018, Theorem 1).

**Proposition 21** (Dong et al. (2019, Proposition 17) extended to *one-step compressed TS*). *For all  $\beta > 0$ , under the logistic bandit setting with logistic function  $\phi_\beta(x)$ , letting  $\tilde{\Theta}_t^*$  and  $\tilde{\Theta}_t$  satisfy the conditions in Dong and Van Roy (2018, Proposition 2), the one-step compressed TS information ratio is bounded as*

$$\tilde{\Gamma}(\tilde{\Theta}_t^*, \tilde{\Theta}_t) := \frac{\mathbb{E}_t[R(\pi_\star(\tilde{\Theta}_t^*), \Theta) - R(\pi_\star(\tilde{\Theta}_t), \Theta)]^2}{I_t(\tilde{\Theta}_t^*; R(\pi_\star(\tilde{\Theta}_t), \Theta), \tilde{\Theta}_t)} \leq d \frac{(1 + \exp(\beta))^4}{32 \exp(\beta)^2}.$$

*Proof.* We combine the proof techniques from Dong et al. (2019, Proposition 17) and (Dong and Van Roy, 2018, Proposition 3). To simplify the exposure we will omit the subscript  $t$  and reuse the notations introduced in Section 5. With these notations, the one-step compressed information ratio can be written as

$$\frac{\mathbb{E}[\text{Bern}(\phi_\beta(\langle \pi_\star(\tilde{\Theta}^*), \Theta \rangle)) - \text{Bern}(\phi_\beta(\langle \pi_\star(\tilde{\Theta}), \Theta \rangle))]^2}{I(\tilde{\Theta}^*; \text{Bern}(\phi_\beta(\langle \pi_\star(\tilde{\Theta}), \Theta \rangle)), \tilde{\Theta})}.$$

Similarly to Dong et al. (2019, Proposition 17), we let  $L_1 = \inf_{x \in [-1, 1]} |\phi_\beta(x)'| = \frac{\beta \exp(\beta)}{(1 + \exp(\beta))^2}$  and  $L_2 = \sup_{x \in [-1, 1]} |\phi_\beta(x)'| = \frac{\beta}{4}$ . We will start by analyzing the numerator which we write as

$$\begin{aligned} \mathbb{E}[\text{Bern}(\phi_\beta(\langle \pi_\star(\tilde{\Theta}^*), \Theta \rangle)) - \text{Bern}(\phi_\beta(\langle \pi_\star(\tilde{\Theta}), \Theta \rangle))]^2 &= \mathbb{E}[\phi_\beta(\langle \pi_\star(\tilde{\Theta}^*), \Theta \rangle) - \phi_\beta(\langle \pi_\star(\tilde{\Theta}), \Theta \rangle)]^2 \\ &\leq L_2^2 \cdot \mathbb{E}[\langle \pi_\star(\tilde{\Theta}^*), \Theta \rangle - \langle \pi_\star(\tilde{\Theta}), \Theta \rangle]^2, \end{aligned}$$

where the inequality follows from  $\frac{\phi_\beta(x_1) - \phi_\beta(x_2)}{x_1 - x_2} \leq \sup_{x \in [-1, 1]} |\phi_\beta(x)'|$  for all  $x_1, x_2 \in [-1, 1]$ .

We then focus on the denominator, which we lower bound similarly as in (Dong and Van Roy, 2018, Lemma 2). Introducing the notation  $R(\tilde{\theta}) = \text{Bern}(\phi_\beta(\langle \pi_\star(\tilde{\theta}), \Theta \rangle))$ , we can write

$$\begin{aligned} I(\tilde{\Theta}^*; \text{Bern}(\phi_\beta(\langle \pi_\star(\tilde{\Theta}), \Theta \rangle)), \tilde{\Theta}) &\stackrel{(i)}{=} I(\tilde{\Theta}^*; \tilde{\Theta}) + I(\tilde{\Theta}^*; \text{Bern}(\phi_\beta(\langle \pi_\star(\tilde{\Theta}), \Theta \rangle)) | \tilde{\Theta}) \\ &\stackrel{(j)}{=} \sum_{\tilde{\theta} \in \mathcal{O}_\epsilon} \sum_{\tilde{\theta}^* \in \mathcal{O}_\epsilon} \mathbb{P}[\tilde{\Theta} = \tilde{\theta}] \mathbb{P}[\tilde{\Theta}^* = \tilde{\theta}^*] D_{\text{KL}}(\mathbb{P}_{R(\tilde{\theta}) | \tilde{\Theta} = \tilde{\theta}} \| \mathbb{P}_{R(\tilde{\theta})}) \\ &\stackrel{(k)}{\geq} \sum_{\tilde{\theta} \in \mathcal{O}_\epsilon} \sum_{\tilde{\theta}^* \in \mathcal{O}_\epsilon} \mathbb{P}[\tilde{\Theta} = \tilde{\theta}] \mathbb{P}[\tilde{\Theta}^* = \tilde{\theta}^*] 2(\mathbb{E}[R(\tilde{\theta}) | \tilde{\Theta} = \tilde{\theta}] - \mathbb{E}[R(\tilde{\theta})])^2, \end{aligned}$$

where  $\mathcal{O}_\epsilon$  is the set of values for the random variables  $\tilde{\theta}^*$  and  $\tilde{\theta}$  as defined in Dong and Van Roy (2018, Proposition 2). The inequality (i) follows from the chain-rule; (j) follows from  $\tilde{\Theta}_t^*$  and  $\tilde{\Theta}_t$  being independent as they satisfy the conditions in Dong and Van Roy (2018, Proposition 2); (k) is obtained using the Donsker-Varadhan inequality (Gray, 2013, Theorem 5.2.1) as in Russo and Van Roy (2016, Lemma 3). We then continue to lower bound

$$\begin{aligned} (k) &= 2 \sum_{\tilde{\theta} \in \mathcal{O}_\epsilon} \sum_{\tilde{\theta}^* \in \mathcal{O}_\epsilon} \mathbb{P}[\tilde{\Theta} = \tilde{\theta}] \mathbb{P}[\tilde{\Theta}^* = \tilde{\theta}^*] (\mathbb{E}[\phi_\beta(\langle \pi_\star(\tilde{\theta}), \Theta \rangle) | \tilde{\Theta} = \tilde{\theta}] - \mathbb{E}[\phi_\beta(\langle \pi_\star(\tilde{\theta}), \Theta \rangle)])^2 \\ &\geq 2L_1^2 \sum_{\tilde{\theta} \in \mathcal{O}_\epsilon} \sum_{\tilde{\theta}^* \in \mathcal{O}_\epsilon} \mathbb{P}[\tilde{\Theta} = \tilde{\theta}] \mathbb{P}[\tilde{\Theta}^* = \tilde{\theta}^*] (\mathbb{E}[\langle \pi_\star(\tilde{\theta}), \Theta \rangle | \tilde{\Theta} = \tilde{\theta}] - \mathbb{E}[\langle \pi_\star(\tilde{\theta}), \Theta \rangle])^2, \end{aligned}$$

using the fact that  $\frac{\phi_\beta(x_1) - \phi_\beta(x_2)}{x_1 - x_2} \geq \inf_{x \in [-1, 1]} |\phi_\beta(x)'|$  for all  $x_1, x_2 \in [-1, 1]$ .

Combining the inequality on the numerator and the denominator, we get that

$$\tilde{\Gamma}(\tilde{\Theta}_t^*, \tilde{\Theta}_t) \leq \frac{L_2^2}{2L_1^2} \frac{\left( \sum_{\theta' \in \mathcal{O}_\epsilon} \mathbb{P}[\tilde{\Theta} = \theta'] (\mathbb{E}[\langle \pi_\star(\theta'), \Theta \rangle | \tilde{\Theta} = \theta'] - \mathbb{E}[\langle \pi_\star(\theta'), \Theta \rangle]) \right)^2}{\sum_{\tilde{\theta} \in \mathcal{O}_\epsilon} \sum_{\tilde{\theta}^* \in \mathcal{O}_\epsilon} \mathbb{P}[\tilde{\Theta} = \tilde{\theta}] \mathbb{P}[\tilde{\Theta}^* = \tilde{\theta}^*] (\mathbb{E}[\langle \pi_\star(\tilde{\theta}), \Theta \rangle | \tilde{\Theta} = \tilde{\theta}] - \mathbb{E}[\langle \pi_\star(\tilde{\theta}), \Theta \rangle])^2}. \quad (5)$$

Without loss of generality, we write  $\mathcal{O}_\epsilon = \{\tilde{\theta}_1, \dots, \tilde{\theta}_{|\mathcal{O}_\epsilon|}\}$ . Now, we define the random matrix  $M \in \mathbb{R}^{|\mathcal{O}_\epsilon| \times |\mathcal{O}_\epsilon|}$  where for each  $i, j \in \{1, \dots, |\mathcal{O}_\epsilon|\}$  the corresponding entry is given by

$$M_{i,j} = \sqrt{\mathbb{P}[\tilde{\Theta}^* = \tilde{\theta}_i] \mathbb{P}[\tilde{\Theta}^* = \tilde{\theta}_j]} (\mathbb{E}[\langle \pi_*(\tilde{\theta}_i), \Theta \rangle | \tilde{\Theta}^* = \tilde{\theta}_j] - \mathbb{E}[\langle \pi_*(\tilde{\theta}_i), \Theta \rangle]).$$

We note that we can rewrite eq. (5) using the trace and the Frobenius norm of the matrix  $M$ , as

$$\tilde{\Gamma}(\tilde{\Theta}_t^*, \tilde{\Theta}_t) \leq \frac{L_2^2}{2L_1^2} \frac{\text{Trace}(M)^2}{\|M\|_F} \leq \frac{L_2^2}{2L_1^2} \text{Rank}(M),$$

where the last inequality is obtained from (Russo and Van Roy, 2016, Fact 10).

The proof concludes showing the rank of the matrix  $M$  is upper bounded by  $d$ . For the sake of brevity, we define  $\bar{\Theta} := \mathbb{E}[\Theta]$  and  $Q_i := \mathbb{E}[\Theta | \tilde{\Theta}^* = \tilde{\theta}_i]$  for all  $i \in \{1, \dots, |\mathcal{O}_\epsilon|\}$ . We then have  $\mathbb{E}[\langle \pi_*(\tilde{\theta}_i), \Theta \rangle | \tilde{\Theta}^* = \tilde{\theta}_j] = \langle \pi_*(\tilde{\theta}_i), Q_j \rangle$  and  $\mathbb{E}[\langle \pi_*(\tilde{\theta}_i), \Theta \rangle] = \langle \pi_*(\tilde{\theta}_i), \bar{\Theta} \rangle$ . Since the inner product is linear, we can rewrite each entry  $M_{i,j}$  of the matrix  $M$  as

$$\sqrt{\mathbb{P}[\tilde{\Theta}^* = \tilde{\theta}_i] \mathbb{P}[\tilde{\Theta}^* = \tilde{\theta}_j]} \langle \pi_*(\tilde{\theta}_i), Q_j - \bar{\Theta} \rangle.$$

Equivalently, the matrix  $M$  can be written as

$$\begin{bmatrix} \sqrt{\mathbb{P}[\tilde{\Theta}^* = \tilde{\theta}_1]} \pi_*(\tilde{\theta}_1) \\ \vdots \\ \sqrt{\mathbb{P}[\tilde{\Theta}^* = \tilde{\theta}_{|\mathcal{O}_\epsilon|}]} \pi_*(\tilde{\theta}_{|\mathcal{O}_\epsilon|}) \end{bmatrix} \begin{bmatrix} \sqrt{\mathbb{P}[\tilde{\Theta}^* = \tilde{\theta}_1]} (Q_1 - \bar{\Theta}) & \dots & \sqrt{\mathbb{P}[\tilde{\Theta}^* = \tilde{\theta}_{|\mathcal{O}_\epsilon|}]} (Q_{|\mathcal{O}_\epsilon|} - \bar{\Theta}) \end{bmatrix}.$$

This rewriting highlights that  $M$  can be written as the product of a  $|\mathcal{O}_\epsilon|$  by  $d$  matrix and a  $d$  by  $|\mathcal{O}_\epsilon|$  matrix and therefore has a rank lower or equal than  $\min(d, |\mathcal{O}_\epsilon|)$ .  $\square$

## F Regarding the gaps in previous literature

In Section 1, we mention that the main results of Dong et al. (2019) are incomplete because of two shortcomings. The first one concerns a gap in their analysis of the Thompson Sampling information ratio for values of  $\beta > 2$ . The second one regards a mistake in their regret analysis, which combines incompatible results. We elaborate on both shortcomings below.

**Regarding the first shortcoming** We identified an issue in the proof of Dong et al. (2019, Theorem 5) at the end of the proof on page 20, where the inequality  $\chi > \xi > 0.1\lambda$  is stated without justification. This inequality plays a crucial role in deriving their bound on the Thompson Sampling information ratio, but the only evidence provided is Dong et al. (2019, Figure 3), which illustrates the functions  $\chi(\lambda, \beta)$  and  $\xi(\lambda, \beta)$  for the specific case of  $\beta = 2$ . While this figure suggests that the inequality holds for  $\beta = 2$ , it cannot be used to conclude that the inequality holds in general for  $\beta \geq 2$ . Additionally, we note that the computation of  $\chi(\lambda, \beta)$  and  $\xi(\lambda, \beta)$  for given values of  $\lambda$  and  $\beta$  is highly intricate, and despite our best efforts, we were unable to reproduce Dong et al. (2019, Figure 3).

**Regarding the second shortcoming** As mentioned earlier, the regret analysis in Dong et al. (2019, Theorems 1 and 5) combines incompatible results. Specifically, the paper uses a uniform bound on the information ratio of the “standard” Thompson Sampling (provided in Dong et al. (2019, Appendix B, Eq. (18))) together with Dong and Van Roy (2018, Theorem 1), which requires a uniform bound on the information ratio of the *one-step compressed Thompson Sampling*. This inconsistency invalidates the regret bounds derived in Dong et al. (2019, Theorems 1 and 5). We emphasize that the problem is “hidden” in Dong et al. (2019, Proposition 9), which incorrectly restates Dong and Van Roy (2018, Theorem 4). It is important to note that there is no straightforward way to extend the result from Dong and Van Roy (2018, Theorem 4) such that it the regret bound works with bounds on the “standard” Thompson Sampling (and not with *one-step compressed Thompson Sampling*), as this compressed version of Thompson Sampling is central to Dong and Van Roy (2018, Theorem 4).

While it is possible to use Dong and Van Roy (2018, Proposition 1) directly with a uniform bound on the “standard” Thompson Sampling information ratio, this approach is limited because Dong and Van Roy (2018, Proposition 1) provides a loose bound. Specifically, this bound depends on the cardinality of the parameter space  $\Theta$  through the entropy  $H(\Theta^*)$  (or on the cardinality of the action space through  $H(A^*)$  in the original version Russo and Van Roy (2016, Proposition 1)). This issue is highlighted in Dong and Van Roy (2018) at the end of Section 3, and serves as a motivation for the introduction of the *one-step-compressed Thompson Sampling* regret analysis in the paper. Combining Dong and Van Roy (2018, Proposition 1) with our bound on the Thompson Sampling information ratio, Proposition 3, results in a regret bound of the order  $O(\sqrt{dT \log(|\mathcal{O}|)})$ , which becomes vacuous for infinite or continuous parameter spaces.

## G Constructing $\pi_*$ as a one-to-one mapping

If the mapping  $\pi_*(\theta)$  is not one-to-one, it could either be that a particular parameter is optimal for several actions or that a particular action is optimal for several parameters.

In the first case, for example, if actions  $a_1, a_2 \in \mathcal{A}$  are optimal for the same parameter  $\theta_1 \in \mathcal{O}$ , then it implies that  $\mathbb{E}[R(a_1, \theta_1)] = \mathbb{E}[R(a_2, \theta_1)] \geq \mathbb{E}[R(a, \theta_1)]$  for all  $a \in \mathcal{A}$  with  $a \neq a_1$  and  $a \neq a_2$ . In this scenario, we can arbitrarily set  $\pi_*(\theta_1) = a_1$  without affecting the regret of Thompson Sampling, as  $\mathbb{E}[R(a_1, \theta_1)] = \mathbb{E}[R(a_2, \theta_1)]$ .

In the second case, if an action  $a_1 \in \mathcal{A}$  is optimal for multiple parameters, say  $\theta_1, \theta_2 \in \mathcal{O}$ , we can artificially construct an *action label* set  $\mathcal{A}'$  such that two labels,  $a'_1, a'_2 \in \mathcal{A}'$ , are associated with  $a_1$ . For all other actions in  $\mathcal{A} \setminus \{a_1\}$ , there is a corresponding label in  $\mathcal{A}'$ . We denote the mapping between action labels and their corresponding actions using the function  $\rho : \mathcal{A}' \rightarrow \mathcal{A}$ . We can construct a function  $\pi_* : \mathcal{O} \rightarrow \mathcal{A}'$  such that  $\pi_*$  is a one-to-one mapping between the parameters  $\mathcal{O}$  and the action labels  $\mathcal{A}'$ . We define the *optimal action label* as  $A^{*'} = \pi_*(\Theta)$  and the *Thompson Sampling action label* as  $A_t' = \pi_*(\Theta_t)$ . This artificial construction, illustrated in Figure 8, is intended solely for the purposes of our regret analysis and has no impact on the regret of Thompson Sampling. The instant regret of Thompson Sampling at time  $t \in \{1, \dots, T\}$  remains  $R(A^*, \Theta) - R(A_t, \Theta)$ , where  $A^* = \rho(A^{*'})$  is still the optimal action for  $\Theta$ , and  $A_t = \rho(A_t')$  is the action selected by Thompson Sampling for  $\Theta_t$ . In this context, the Thompson Sampling information ratio would be adapted and defined as:

$$\Gamma_t := \frac{\mathbb{E}_t[R(A^*, \Theta) - R(A_t, \Theta)]^2}{\mathbb{I}_t(A^{*'}; R(A_t, \Theta), A_t')},$$

representing the ratio between the current squared regret and the information gathered about the optimal action label. One can verify that the analysis of the information ratio in Section 5, Appendix B, and Appendix C, proceeds the same with this adapted definition and leads to the same upper bound.

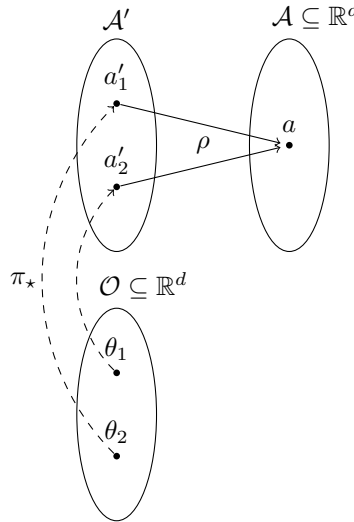


Figure 8: Illustration of the artificial construction of the action label set  $\mathcal{A}'$ .