Incentivizing Exploration with Linear Contexts and Combinatorial Actions

Mark Sellke [1]

Abstract

We advance the study of incentivized bandit exploration, in which arm choices are viewed as recommendations and are required to be Bayesian incentive compatible. Recent work of (Sellke & Slivkins, 2022) has shown that for the special case of independent arms, after collecting enough initial samples, the popular Thompson sampling algorithm becomes incentive compatible. This was generalized to the combinatorial semibandit in (Hu et al., 2022). We give an analog of this result for linear bandits, where the independence of the prior is replaced by a natural convexity condition. This opens up the possibility of efficient and regret-optimal incentivized exploration in high-dimensional action spaces. In the semibandit model, we also improve the sample complexity for the pre-Thompson sampling phase of initial data collection.

1. Introduction

1.1. Problem Formulation and Results

We consider **incentivized** exploration as introduced in (Kremer et al., 2014; Mansour et al., 2020). These are bandit problems motivated by the scenario that the t-th action is a **recommendation** made to the t-th customer. A planner, representing a central service such as a major website, observes outcomes of past customers. By contrast, each customer sees only the recommendation they are given and visits the planner only once. Assuming a common Bayesian prior (and indistinguishability of customers), we would like our recommendations to be trustworthy, in the sense that a rational customer will follow our recommendations. We have the following definition:

Definition 1. Let $\mathcal{A}$ be an action set and μ a prior distribution over the mean reward function $\ell^* : \mathcal{A} \to \mathbb{R}$. A bandit algorithm which recommends a sequence $(A^{(1)}, A^{(2)}, \ldots, A^{(T)})$ of actions in $\mathcal{A}$ is said to be **Bayesian incentive compatible** (BIC) if for each $t \in [T]$:

$$\begin{aligned} &\mathbb{E}[\mu(A) - \mu(A') \mid A^{(t)} = A] \geq 0, \\ &\quad \forall A, A' \in \mathcal{A} \text{ such that } \mathbb{P}[A^{(t)} = A] > 0. \end{aligned} \tag{1.1}$$

We will often refer to the algorithm as the *planner*, and $A^{(t)}$ as a *recommendation* made to an *agent*. Here the agent knows the (source code for the) planner's algorithm, and the value of t (i.e. his place in line) but nothing about the online feedback. (In particular, each agent appears just once and never "comes back".) Thus, (1.1) exactly states that rational agents will follow the planner's recommendation.

This model was formulated in (Kremer et al., 2014), and is a multi-stage generalization of the well-studied Bayesian persuasion problem in information design (Bergemann & Morris, 2019; Kamenica, 2019). Fundamental results on feasible explorability and vanishing regret for BIC algorithms were obtained in (Mansour et al., 2020; 2022). In all cases, the fundamental principle is to use *information asymmetry* to guide agent decisions toward exploration.

Let us remark that averaging over the choice of $A^{(t)}$ shows that for any BIC algorthm $\mathcal{A}$ and fixed time t,

$$\mathbb{E}[\mu(A^{(t)})] \geq \sup_{A \in \mathcal{A}} \mathbb{E}[\mu(A)].$$

That is, the recommended time t action $A^{(t)}$ is always better on average than exploiting according to the prior. Hence as a special case, BIC algorithms are guaranteed to benefit **all** users as compared to naive exploitation without learning. Moreover an agent's knowledge of the exact time t only makes the BIC guarantee stronger.

As mentioned, fundamental results on BIC bandit algorithms were obtained in (Mansour et al., 2020; 2022). For instance, the former work gave "hidden exploration" algorithms which exploit with such high frequency that *any* bandit algorithm is BIC on the remaining time-steps given suitable assumptions on the prior. Unfortunately, all algorithms in these works had to pay exponentially large multiplicative factors in their regret compared to non-BIC bandit algorithms. This "price of incentives" was studied quantitatively in (Sellke & Slivkins, 2022) in the case of independent arms. They proved the classical Thompson sampling

[1]Amazon Core AI. Correspondence to: Mark Sellke <mselke@gmail.com>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

algorithm (Thompson, 1933) is BIC without modification once a mild number of initial samples per arm have been collected, leading to a natural two-phase algorithm. The first phase focused on collecting these initial data in a BIC way, while the second is simply Thompson sampling. Note that Thompson sampling obeys many state-of-the-art regret bounds, both Bayesian (Russo & Van Roy, 2014; Bubeck & Liu, 2013) and frequentist (Agrawal & Goyal, 2012; Kaufmann et al., 2012; Agrawal & Goyal, 2013a). Thus the result of (Sellke & Slivkins, 2022) implies that the additional regret forced by the BIC requirement is essentially additive, and at most the number of rounds needed to collect the needed initial samples in a BIC way (a quantity which (Sellke & Slivkins, 2022) studied separately).

The results of (Sellke & Slivkins, 2022) are limited to the case of independent arms, and it is therefore of interest to expand the range of models under which Thompson sampling leads to provable incentive compatibility. This is the goal of the present paper. Our main focus is on the **linear** bandit in dimension d. Here the action set $\mathcal{A} \subseteq \mathbb{R}^d$ may be infinite, and the reward function $\ell^* : \mathcal{A} \to \mathbb{R}$ is always linear. This is a natural next step and allows for richer correlations between actions. From the practical viewpoint, BIC guarantees are relevant for recommending restaurants, movies, hotels, and doctors (see e.g. (Slivkins, 2019), Chapter 11); in all of these settings it is desirable to leverage contextual information to obtain sample complexity scaling with the ambient data dimension rather than the potentially huge number of total actions. We make no assumptions of independence but instead require natural geometric conditions, e.g. that the prior μ for ℓ^* is uniformly random on a convex body $\mathcal{K}$ with bounded aspect ratio. This covers a fairly broad range of scenarios, including centered Gaussian μ by homogeneity. Indeed we consider the connection with such conditions as a significant contribution of this work; in (Sellke & Slivkins, 2022; Hu et al., 2022) the FKG inequality is used crucially throughout but requires independence properties that are unavaiable with linear contexts. As a regret benchmark, recall that for the Bayesian regret of Thompson sampling is known to be $O(d\sqrt{T \log T})$ by Theorem 2 of (Dong & Van Roy, 2018) (see also (Russo & Van Roy, 2014)). [1]

Relative to this near-optimal guarantee, our first main result Theorem 3.5 shows that the price of incentives is again additive rather than multiplicative. Namely, Thompson sampling is again BIC after obtaining $\mathrm{poly}(d)$ initial samples which are well-spread in a natural spectral sense. Technically, we require action sets be to ε-separated so that the recommendation of a given $A \in \mathcal{A}$ has non-negligible probability; this is a mild condition since any action set can be discretized before running Thompson sampling. Further as shown in Theorem 3.7, the result extends to the **generalized** linear bandit when the link function's derivative is bounded above and below. We finally give two counterexamples. The former shows that Thompson sampling may be BIC at time 1 but not time 2 – this may be surprising as it is impossible for the multi-armed bandit with independent arms (see Lemma 4.9 of (Sellke & Slivkins, 2022)). The latter gives a natural example in which initial data collection provably needs $e^{\Omega(d)}$ samples. In particular, the latter counterexample illustrates that further geometric conditions on the prior and/or action set are necessary for $\mathrm{poly}(d)$ sample complexity of BIC initial exploration. We leave a further study of this interesting problem for future work, but remark that exogenous payments could be used in practice to obtain the required $\mathrm{poly}(d)$ initial samples (see some of the references in Subsection 1.2).

Finally, we give new results for incentivized exploration of the combinatorial semibandit. Here actions consist of subsets $A \subseteq [d]$ of at most d independent atoms which each give separate feedback as well as rewards. This is another natural testing ground for correlated rewards and was the recent focus of (Hu et al., 2022); they showed that Thompson sampling is still BIC with enough initial samples. (Hu et al., 2022) gave initial exploration algorithms to bound the additive regret increase from being BIC, but with exponentially large sample complexity in typical cases. We improve this latter aspect of their work by extending the framework of (Sellke & Slivkins, 2022), linking the initial sample complexity to the minimax value of a two-player zero-sum game.

1.2. Other Relevant Work

Here we mention a few other related works in the broad area of incentivized exploration. (Frazier et al., 2014) studies a similar problem, again in a Bayesian setting. However in their work the full history is made public so there is no information asymmetry. Instead, incentivization is achieved via exogenous payments. (Wang & Huang, 2018; Agrawal & Tulabandhula, 2020; Wang et al., 2021) study incentivized exploration in a non-Bayesian setting, where empirical averages are used instead of posterior means, and again incentives are realized through payments. The latter two works also focus on linear contexts and show $\widetilde{O}(\sqrt{T})$ total payment suffices for incentivization. By contrast Theorem 3.5 implies that Thompson sampling is BIC after a constant (i.e. T-independent) amount of initial exploration. If payments can be used for incentivization, this implies in particular that a constant amount of total payment suffices for incentivized exploration. However as just mentioned, our setting and assumptions differ from the aforementioned works in multiple ways.

[1] To avoid confusion, note the *frequentist* regret of Thompson sampling for linear bandits is the slightly suboptimal $\tilde{\Theta}(d^{3/2}\sqrt{T})$ (Agrawal & Goyal, 2013b; Hamidi & Bayati, 2020).

(Kannan et al., 2017) studied the power of exogenous payments to incentivize related notions of fairness in which better actions ought to be played with higher probability. Their model includes information asymmetry in the opposite direction; agents observe the full history while the planner might not. (Immorlica et al., 2020) proposed incentivization via *selective data disclosure* where agents observe a carefully chosen subset of the history. However their setting is not precisely Bayesian as they assume agents perform naive empirical averaging over the chosen subset rather than taking the planner's algorithm into account. In fact it can be shown by a form of the revelation princple (see (Slivkins, 2019), Chapter 11) that in the perfectly Bayesian setting, general planner-to-agent signals have no more power than simple action recommendations. Finally our work and the ones mentioned above assume agents are identical; a model with heterogenous agents was studied in (Immorlica et al., 2019).

2. Preliminaries

This paper considers two bandit models with correlated rewards. In both cases the unknown optimal action is denoted A^* (with some arbitrary fixed tie-breaking rule). The first of these is the linear bandit, where we assume the action set $\mathcal{A} \subseteq B_1(0) \subseteq \mathbb{R}^d$ is contained inside the unit ball. Moreover, the reward r_t is given by

$$r_t = \langle \ell^*, A^{(t)} \rangle + z_t$$

where z_t is conditionally mean zero and $O(1)$-subgaussian conditioned on $(\ell^*, A^{(t)})$. We assume ℓ^* is such that

$$\mathbb{E}[r] = \langle \ell^*, A \rangle \in [-1, 1]$$

for all $A \in \mathcal{A}$. This includes $r_t = \langle \ell_t, A^{(t)} \rangle + \tilde{z}_t$ with $\mathbb{E}[\ell_t] = \ell^*$, as well as binary rewards $r_t \in \{-1, 1\}$. When the actions $(A_1, \ldots, A^{(t)})$ span $\mathbb{R}^d$, it will be helpful to consider the ordinary least-squares estimator for the reward vector, which we denote by $\hat{\ell}^*$. We denote $\theta_i = \langle \ell^*, A_i \rangle$ the expected reward for $A_i \in \mathcal{A} = (A_1, \ldots, A_{|\mathcal{A}|})$.

The second model we consider is the combinatorial semibandit. Here the action set $\mathcal{A} \subseteq 2^{[d]}$ is a family of subsets; we call $A \in \mathcal{A}$ an *action* and $a_i \in [d]$ an *atom*. After playing $A^{(t)} \in \mathcal{A}$, the player receives a vector of reward feedback $(r_a)_{a \in A^{(t)}} \in \{0, 1\}^{|A^{(t)}|}$ and gains their entrywise sum as a total reward. Each atom a_i gives reward independently with probability θ_i, and following (Hu et al., 2022) we assume that the θ_i are jointly independent under μ. We let $\theta_A = \sum_{i \in A} \theta_i$ for each $A \in \mathcal{A}$. We let $n_t(i)$ denote the number of times atom a_i has been sampled prior to time t, and let $\hat{p}_n(i)$ be the empirical average reward of arm i from its first $n > 0$ samples. For each $j \in [d]$ we set $\mathcal{A}_j$ be the subset of $\mathcal{A}$ consisting of A containing a_j, and $\mathcal{A}_{-j} = \mathcal{A} \backslash \mathcal{A}_j$.

Thompson sampling is a Bayesian bandit algorithm, defined from an initial prior μ over ℓ^*. Let $\mathcal{F}_t$ denote the observed history strictly before time t and set $\mathbb{E}^t[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_t]$, $\mathbb{P}^t[\cdot] = \mathbb{P}[\cdot|\mathcal{F}_t]$. Thompson sampling at time t draws an arm independently at random from the time-t distribution of A^*, so that

$$\mathbb{P}^t[A^{(t)} = A] = \mathbb{P}^t[A^* = A].$$

Given a fixed prior μ, we say a bandit algorithm is BIC if it satisfies (1.1). More leniently, we say it is ε-BIC if for each $t \in [T]$:

$$\begin{aligned} &\mathbb{E}[\mu(A) - \mu(A') \mid A^{(t)} = A] \geq -\varepsilon, \\ &\quad \forall A, A' \in \mathcal{A} \text{ such that } \mathbb{P}[A^{(t)} = A] > 0. \end{aligned} \tag{2.1}$$

We have mentioned regret guarantees for Thompson sampling with linear contexts. For the combinatorial semibandit, (Bubeck & Sellke, 2022) shows among other things that Thompson sampling attains the optimal $\widetilde{O}(d\sqrt{T})$ regret.

We say a mean-zero scalar random variable X is C-subgaussian if for all $t \in \mathbb{R}$,

$$\mathbb{E}[e^{tX}] \leq e^{C^2 t^2/2}. \tag{2.2}$$

The smallest C such that (2.2) holds is the subgaussian norm of X. We say a mean-zero random vector $\vec{X} \in \mathbb{R}^d$ is C-subgaussian if $\langle \vec{X}, \vec{v} \rangle$ is C-subgaussian for all $\vec{v} \in \mathbb{R}^d$ of norm $\|\vec{v}\| \leq 1$.

We use $\preceq$ to denote the positive semidefinite partial order on symmetric matrices, i.e. $M_1 \preceq M_2$ if and only if $M_2 - M_1$ is positive semidefinite.

2.1. Bayesian Chernoff Bounds

We use the following posterior contraction lemma from (Sellke & Slivkins, 2022). When θ is taken to be an empirical average, it yields a Bayesian version of the classical Chernoff bound. The statement below hides some technical specifications for readability, but all quantities lie in $\mathbb{R}^m$ for some m, all functions are Borel measurable, and all probability measures are defined on the Borel sigma algebra.

Lemma 2.1. *Let $\xi \in \mathbb{R}^d$ be an unknown parameter and γ an observed signal with distribution depending on ξ. Suppose there exists an estimator $\theta = \theta(\gamma) \in \mathbb{R}^d$ for ξ depending only on this signal, which satisfies for some deterministic $\varepsilon, \delta > 0$ the concentration inequality*

$$\mathbb{P}\big[\|\theta - \xi\| \geq \varepsilon \mid \xi\big] \leq \delta \quad \forall \xi. \tag{2.3}$$

Further, let $\xi \sim \mu$ be generated according to a prior distribution μ, and let $\hat{\xi}$ be a sample from the posterior distribution $\hat{\mu} = \hat{\mu}(\gamma)$ for ξ conditioned on the observation γ. Then

$$\mathbb{E}^{\xi \sim \mu}\big[\mathbb{P}^{\hat{\xi} \sim \hat{\mu}}\big[\|\hat{\xi} - \xi\| \geq 2\varepsilon\big]\big] \leq 2\delta. \tag{2.4}$$

Proof. The pairs (ξ, γ) and $(\hat{\xi}, \gamma)$ are identically distributed; therefore (ξ, θ) and $(\hat{\xi}, \theta)$ are as well. The result now follows by the triangle inequality. □

Lemma 2.2 (Lemma A.13 of (Sellke & Slivkins, 2022))**.** *Suppose the scalar random variable X is mean zero and $O(1)$-subgaussian and the event E has $\mathbb{P}[E] \leq \delta$. Then*

$$\mathbb{E}[|X \cdot 1_E|] \leq O\big(\delta\sqrt{\log(1/\delta)}\big).$$

Though the statement of Lemma 2.1 is abstract, our uses of it will be very concrete. Namely ξ will be the unknown mean reward and γ the actions and rewards up to some time t. The estimate θ will be obtained simply by an empirical average or linear regression. Then (2.3) amounts to Hoeffding's inequality or a multivariate analog. The conclusion (2.4) for posterior samples will be of great use in the analysis of Thompson sampling.

3. Linear Bandit

In (Sellke & Slivkins, 2022), it was shown that Thompson sampling is BIC once a mild amount of initial data has been collected almost surely at a fixed time. Here we show a qualitatively similar result for the linear bandit. The notion of "amount" of initial data we adopt is that the action vectors $A^{(1)}, \ldots, A^{(t)}$ taken so far satisfy the spectral condition

$$\sum_{s=1}^{t} (A^{(s)})^{\otimes 2} \succeq \gamma I_d. \tag{3.1}$$

If $\mathcal{A} = (e_1, \ldots, e_d)$ forms an orthonormal basis (as in the multi-armed bandit setting), this simply means that each action was sampled at least γ times. We say that γ-spectral exploration has occurred at time t if (3.1) holds. Note that if one is willing to simply purchase initial samples, then γd samples are typically needed to achieve γ-spectral exploration. The next lemma follows from Lemma 2.2 and is proved in the appendix.

Lemma 3.1. *Suppose γ-spectral exploration has occurred almost surely at some (deterministic) time t, and let $v \in \mathbb{R}^d$ be a deterministic vector. Then the random vector $\ell^* - \mathbb{E}^t[\ell^*]$ has zero mean and is $O(\gamma^{-1/2})$-subgaussian.*

Our results will hold for ε-separated action sets $\mathcal{A}$ as defined below. We view this as a generic assumption, e.g. it holds with high probability when $\mathcal{A} \subseteq \mathbb{S}^{d-1}$ consists of $e^{O(d)}$ randomly chosen points. Moreover it is common to discretize infinite action sets for the purposes of analysis. From a technical point of view, requiring discrete $\mathcal{A}$ ensures that the event conditioned on has non-negligible probability.

Definition 2. The set $\mathcal{A} \subseteq \mathbb{S}^{d-1}$ is said to be ε-separated if $\|A_1 - A_2\| \geq \varepsilon$ for any **distinct** $A_1, A_2 \in \mathcal{A}$.

3.1. Convex Geometry

We start with two definitions from convex geometry.

Definition 3. For $\mathcal{K} \subseteq \mathbb{R}^d$ a compact convex set with non-empty interior, let $\mathcal{U}(\mathcal{K})$ denote the uniform measure on $\mathcal{K}$. If $v \in \mathbb{R}^d$ we let

$$\mathrm{width}_v(\mathcal{K}) = \max_{\ell \in \mathcal{K}} \langle \ell, v \rangle - \min_{\ell \in \mathcal{K}} \langle \ell, v \rangle.$$

Definition 4. The convex set $\mathcal{K} \subseteq \mathbb{R}^d$ is r-regular if $B_r(0) \subseteq \mathcal{K} \subseteq B_1(0)$.

The next lemma lets us connect convexity conditions to the BIC property. Its proof follows from standard tools and is given in the Appendix.

Lemma 3.2. *For any convex $\mathcal{K} \subseteq \mathbb{R}^d$,*

$$\mathbb{E}^{\ell \sim \mathcal{U}(\mathcal{K})} |\langle \ell, v \rangle| \geq \Omega(\mathrm{width}_v(\mathcal{K})/d).$$

Our key technical estimate is below. It requires ℓ^* to have uniform prior $\mathcal{U}(\mathcal{K})$ over a regular convex body $\mathcal{K}$. Note that since the left-hand side of (3.3) is 1-homogeneous, it is also fine for ℓ^* to be drawn from a **centered** Gaussian with covariance Σ satisfying $\lambda I_d \preceq \Sigma \preceq \Lambda I_d$.

Corollary 3.3. *Suppose that $\mathcal{A} \subseteq \mathbb{S}^{d-1}$ is ε-separated, the convex set $\mathcal{K} \subseteq \mathbb{R}^d$ is r-regular, and $\ell^* \sim \mu = \mathcal{U}(\mathcal{K})$. Then for each $A_i, A_j \in \mathcal{A}$, we have*

$$\mathbb{P}[A_i = A^*(\ell^*)] \geq (r\varepsilon/4)^d, \tag{3.2}$$

$$\mathbb{E}[\langle \ell^*, A_i - A_j \rangle | A^* = A_i] \geq \Omega\left(\frac{r\varepsilon \|A_i - A_j\|}{d}\right). \tag{3.3}$$

Proof. Fix i and define the convex set $S_i = \{\ell \in B_1 : A_i = A^*(\ell)\}$. To prove (3.2), we first show

$$B_{\varepsilon/2}(A_i) \subseteq S_i. \tag{3.4}$$

Indeed suppose $\ell \in \mathbb{R}^d$ satisfies $\|\ell - A_i\| \leq \frac{\varepsilon}{2}$. Then for any $a \in \mathcal{A}$ different from A_i,

$$\begin{aligned}\langle \ell, A_i - A \rangle &= \langle A_i, A_i \rangle - \langle A_i, a \rangle + \langle \ell - A_i, A_i - A \rangle \\ &\geq 1 - \langle A_i, a \rangle - \frac{\varepsilon}{2}\|A_i - A\|_2 \\ &= \|A_i - A\|_2^2/2 - \frac{\varepsilon}{2}\|A_i - A\|_2 \geq 0.\end{aligned}$$

In the last step we used the fact that $t^2 - \varepsilon t \geq 0$ for $t \geq \varepsilon$ (recall that $\mathcal{A}$ is ε-separated). We conclude that $\ell \in S_i$, hence establishing (3.4). By r-regularity of $\mathcal{K}$ and homogeneity of S_i it follows that

$$B_{r\varepsilon/4}(rA_i/2) \subseteq \mathcal{K} \cap S_i.$$

Since $\mathcal{K} \subseteq B_1(0)$ we find $\mathrm{Vol}(\mathcal{K} \cap S_i)/\mathrm{Vol}(\mathcal{K}) \geq (r\varepsilon/4)^d$, implying (3.2). Next, r-regularity of $\mathcal{K}$ and (3.4) imply

$$\begin{aligned}\mathrm{width}_v(S_i \cap \mathcal{K}) &\geq r \cdot \mathrm{width}_v(S_i) \\ &\geq \Omega(r\varepsilon\|v\|)\end{aligned}$$

for any $v \in \mathbb{R}^d$. Hence by Lemma 3.2, for any fixed v

$$\mathbb{E}^{\ell \sim \mathcal{U}(S_i \cap \mathcal{K})} \left| \langle \ell, v \rangle \right| \geq \Omega(r\varepsilon \|v\|/d).$$

Setting $v = A_i - A_j$, we note that by definition $\langle \ell, A_i - A_j \rangle \geq 0$ for all $\ell \in S_i$. Hence

$$\mathbb{E}^{\ell \sim \mathcal{U}(S_i \cap \mathcal{K})} \langle \ell, A_i - A_j \rangle \geq \Omega(r\varepsilon \|A_i - A_j\|/d)$$

which is equivalent to (3.3). □

3.2. Main Result for Linear Bandit

We now show our main result on Thompson sampling for the linear bandit, namely that γ-spectral exploration suffices for Thompson sampling to be BIC when $\gamma \geq \frac{Cd^3 \log(4/r\varepsilon)}{r^2\varepsilon^2}$ for ε-separated action sets. This roughly means $\widetilde{O}\left(d^4/\varepsilon^2\right)$ "well-dispersed" samples are required for Theorem 3.5 to apply.

Remark 3.4. In fact a more general statement is true. Given an ε_t-discretization $\mathcal{A}^{(t)}$ of $\mathcal{A}$, let us recommend an action $A^{(t)} \in \mathcal{A}^{(t)}$ by performing Thompson sampling on $\mathcal{A}^{(t)}$. Then the proof of Theorem 3.5 shows that a rational user will always choose an action $\tilde{A}^{(t)} \in \mathcal{A}$ within distance ε_t of $A^{(t)}$, since all other actions are inferior to $A^{(t)}$.

Theorem 3.5. *Suppose that $\mathcal{A} \subseteq \mathbb{S}^{d-1}$ is ε-separated, the convex set $\mathcal{K} \subseteq \mathbb{R}^d$ is r-regular, and $\ell^* \sim \mu = \mathcal{U}(\mathcal{K})$. If (3.1) holds almost surely at time t for*

$$\gamma \geq \frac{Cd^3 \log(4/r\varepsilon)}{r^2\varepsilon^2},$$

then Thompson sampling for the linear bandit is BIC at time t.

Proof. As in the proof of Theorem 4.1 in (Sellke & Slivkins, 2022), it suffices to show that

$$\mathbb{E}[1_{A^*=A_i} \cdot \mathbb{E}^{T_0}[\langle \ell^*, A_i - A_j \rangle]] \geq 0$$

for any action $A_j \neq A_i$. Define

$$\delta_i = \mathbb{P}[A_i = A^*(\ell^*)] \overset{(3.2)}{\geq} \left(\frac{r\varepsilon}{4}\right)^d.$$

We have

$$\begin{aligned}
&\mathbb{E}\big[1_{A^*=A_i} \cdot \langle \ell^*, A_i - A_j \rangle\big] \\
&= \mathbb{E}\big[1_{A^*=A_i} \cdot (\langle \ell^*, A_i - A_j \rangle)_+\big] \\
&= \delta_i \cdot \mathbb{E}\big[(\langle \ell^*, A_i - A_j \rangle)_+ \mid A^* = A_i\big] \\
&\overset{Lem.\ 3.2}{\geq} \Omega(\delta_i r\varepsilon \|A_i - A_j\|/d).
\end{aligned}$$

Recall that $\theta_i = \langle \ell^*, A_i \rangle$ denotes the expected reward for action A_i. It remains to upper-bound

$$\mathbb{E}\left[1_{A^*=A_i} \cdot \left|\mathbb{E}^{T_0}[\theta_i - \theta_j] - (\theta_i - \theta_j)\right|\right].$$

By Lemma 2.1 and (3.1), the value $\frac{\gamma^{1/2}\left(\mathbb{E}^{T_0}[\theta_i - \theta_j] - (\theta_i - \theta_j)\right)}{\|A_i - A_j\|}$ is centered and $O(1)$ sub-gaussian (sans conditioning). Using Lemma 2.2, it follows that

$$\begin{aligned}
&\mathbb{E}\left[1_{A^*=A_i} \cdot \left|\mathbb{E}^{T_0}[\theta_i - \theta_j] - (\theta_i - \theta_j)\right|\right] \\
&\leq O\left(\delta_i \|A_i - A_j\| \sqrt{\frac{\log(1/\delta_i)}{\gamma}}\right).
\end{aligned}$$

Since we assumed

$$\gamma \geq \frac{Cd^3 \log(4/r\varepsilon)}{r^2\varepsilon^2}$$

for a large enough absolute constant C, we obtain

$$\begin{aligned}
&\mathbb{E}\left[1_{A^*=A_i} \cdot \left|\mathbb{E}^{T_0}[\theta_i - \theta_j] - (\theta_i - \theta_j)\right|\right] \\
&\leq O\left(\delta_i \|A_i - A_j\| \sqrt{\frac{\log(1/\delta_i)}{\gamma}}\right) \\
&\leq \Omega\left(\frac{\delta_i r\varepsilon \|A_i - A_j\|}{d}\right) \\
&\leq \mathbb{E}\big[1_{A^*=A_i} \cdot \langle \ell^*, A_i - A_j \rangle\big]
\end{aligned}$$

In the second step we used $\log(1/\delta_i) \leq d \log(4/r\varepsilon)$ which follows from (3.2). This concludes the proof. □

Remark 3.6. The action set $\mathcal{A}$ only enters Theorem 3.5 at time t. Thus it holds even if the preceding γ-spectral exploration used actions not in $\mathcal{A}$. In particular Theorem 3.5 extends to the case that $\mathcal{A}^{(t)}$ changes over time as long as $\mathcal{A}^{(t)}$ is ε-separated.

For example, suppose that each $\mathcal{A}^{(t)}$ consists of e^{cd} i.i.d. uniform vectors on the unit sphere for a small constant c, but the sets $\mathcal{A}^{(t)}$ may be arbitrarily dependent. (For example, we might replace the actions gradually.) Each $\mathcal{A}^{(t)}$ is ε-separated with probability $1 - e^{-\Omega(d)}$. Then given a γ-spectral warm start, the algorithm *use Thompson sampling if ε-separation holds, else act greedily* will be BIC and incur expected regret

$$O\left(d\sqrt{T} \log T + e^{-\Omega(d)} T\right).$$

This follows by Proposition 3 of (Russo & Van Roy, 2014), which allows for changing action sets $\mathcal{A}^{(t)}$.

3.3. Extension to Generalized Linear Bandit

Here we give an analogous proof for the generalized linear bandit, which was introduced in (Filippi et al., 2010). In this model, the expected reward is

$$\mathbb{E}[r_t | A^{(t)}] = \chi(\langle A^{(t)}, \ell^* \rangle)$$

for a known strictly increasing **link function** $\chi : \mathbb{R} \to \mathbb{R}$ and unknown vector ℓ^*. A notable example is the logistic

bandit where $\chi(x) = \frac{e^x}{1+e^x}$. Recalling that we assumed $\langle \ell^*, x\rangle \in [-1,1]$ for all $\ell^* \in \text{supp}(\mu)$ and $x \in A$, we follow much of the theoretical literature on this model by fixing the non-negative constants

$$M_\chi = \sup_{x\in[-1,1]} \chi'(x); \qquad m_\chi = \inf_{x\in[-1,1]} \chi'(x).$$

(See e.g. (Neu et al., 2022) for an analysis that avoids such dependence.) Our analysis of the linear bandit turns out to extend to this setting; the proof is given in the Appendix.

Theorem 3.7. *Fix a link function $\mu : \mathbb{R} \to \mathbb{R}$. Suppose that $\mathcal{A} \subseteq \mathbb{S}^{d-1}$ is ε-separated, the convex set $\mathcal{K} \subseteq \mathbb{R}^d$ is r-regular, and $\ell^* \sim \mu = \mathcal{U}(\mathcal{K})$. If (3.1) holds almost surely at time t for*

$$\gamma \geq \frac{CM_\chi^2 d^3 \log(4/r\varepsilon)}{r^2 m_\chi^4 \varepsilon^2}$$

then Thompson sampling for the generalized linear bandit is BIC at time t.

3.4. Two Counterexamples

Theorem 4.7 and Lemma 4.9 of (Sellke & Slivkins, 2022) show that for the multi-armed bandit with an independent product prior over arms, if Thompson sampling is BIC at time s then it is also BIC at any future time $t > s$. (In fact, this holds even if an arbitrary and possibly non-BIC bandit algorithm is used between times s and t.) This is an appealing general "stability" property for Thompson sampling. In particular, the special case $s = 0$ implies that if all arms have the same prior mean reward, then Thompson sampling is BIC with no initial exploration phase.

We give an explicit counterexample (with proof in the appendix) showing that even the $s = 0$ case fails for the linear bandit when $(s, t) = (1, 2)$ with $d = 2$. Precisely:

Proposition 3.8. *Let $\mu \sim \mathcal{N}(0, I_2)$ and*

$$(A_1, A_2, A_3) = \big((1,0),\ (-1,0),\ (1.8, 0.6)\big).$$

*If Thompson sampling recommends action A_1 at time $t = 2$, this recommendation is **not** BIC:*

$$\mathbb{E}[\theta_3 \mid A^{(2)} = A_1] > \mathbb{E}[\theta_1 \mid A^{(2)} = A_1]. \tag{3.5}$$

Note that although A_3 has different Euclidean norm from A_1 and A_2, this set can be put onto the unit circle by a linear change of basis (which just changes μ to a non-spherical Gaussian) in accordance with Section 3.

For our second counterexample, suppose the action set $\mathcal{A}$ is either the polytope

$$\mathcal{P} \equiv \left\{(x_1, \ldots, x_d) \ :\ 10|x_d| + 2\sqrt{d}\max_{i<d}|x_i| \leq 1\right\} \subseteq B_1$$

or its set of extreme points. We show that for a suitable prior μ, *any* initial exploration algorithm which is BIC or even ε-BIC can be made to require $e^{\Omega(d)}$ timesteps to explore in the d-th coordinate.

Note that for $\mathcal{P}$ as above, all extreme points satisfy $x_d = 0$ except for $\pm\hat{A} = \left(0, 0, \ldots, 0, \pm\frac{1}{2\sqrt{d}}\right)$, and so γ-spectral exploration for any $\gamma > 0$ requires exploring these actions. Our prior distribution $\ell^* \sim \mu$ will be uniform over the "biased" convex body $\mathcal{K}/\sqrt{d}$ for

$$\mathcal{K} \equiv [-0.5, 1]^{d-1} \times [-1, 1].$$

Proposition 3.9. *With $\mathcal{A}, \mu$ as above, suppose that an ε-BIC algorithm $\mathbb{A}$ has explored at least one of $\hat{A}$ or $-\hat{A}$ almost surely by time T for $\varepsilon \leq 1/20$. Then $T \geq \exp(\Omega(d))$.*

4. Initial Exploration for Combinatorial Semibandit

It was found in (Hu et al., 2022) that for incentivized exploration, the semibandit problem is well-suited for generalizing (Sellke & Slivkins, 2022). Indeed one can still assume the atoms $i \in [d]$ give rewards independently with probabilities $\theta_1, \ldots, \theta_d$ which are independent under the prior μ; this allowed them to generalize several proofs using independence of atoms rather than full actions. However (Hu et al., 2022) did not extend the initial exploration scheme of (Sellke & Slivkins, 2022), instead proposing several algorithms with exponential sample complexity. Here we find the natural extension of (Sellke & Slivkins, 2022), again connecting the sample complexity of BIC initial exploration to the value of a two-player zero sum game.

Following (Hu et al., 2022), we make some non-degeneracy assumptions. Namely we require that all atom rewards are almost surely in $[0, 1]$ and their distributions satisfy:

1. $\mathbb{E}[\theta_\ell] \geq \tau$.
2. $\text{Var}[\theta_\ell] \geq \sigma^2$.
3. $\mathbb{P}[\theta_\ell < x] \geq \exp(-x^{-\alpha})$.

for fixed constants $\tau, \sigma, \alpha > 0$.

Under these conditions, we give an initial exploration algorithm in the style of (Sellke & Slivkins, 2022). We will explore the arms in a deterministic order which is defined via the lemmas below. Throughout, we say that an atom has been N-explored if it has been sampled at least N times. Moreover we define the event

$$\texttt{ZEROS}_{<j,N} = \{\hat{p}_N(i) = 0\ \forall i < j\}$$

that each $i < j$ receives no reward in its first N samples, and also set

$$\varepsilon_j = \mathbb{P}[\texttt{ZEROS}_{<j,N}]. \tag{4.1}$$

It follows from the assumptions above that for all $j \in [d]$,

$$\varepsilon_j \geq (1-p)^{dN} \exp(-dp^{-\alpha}).$$

In particular setting $p = \frac{1}{dN}$ we find:

$$\log(1/\varepsilon_j) \leq \Omega\left(d^{1+\alpha} N^{\alpha}\right). \tag{4.2}$$

Lemma 4.1. *Let $S_{j-1} = \{a_1, \ldots, a_{j-1}\} \subseteq [d]$ and assume that all atoms in S_{j-1} have been N-explored almost surely at time T_{j-1}, for*

$$N \geq \left(20d/\tau\right)^{1+\alpha} \log(20d/\tau). \tag{4.3}$$

Then conditioned on the event $\texttt{ZEROS}_{<j,N}$, the action with highest posterior mean contains an action in $[d] \backslash S_{j-1}$. More generally, let

$$x_j = \max\left(1_{\texttt{ZEROS}_{<j,N}}, b_j\right)$$

for $y_j = b_j \sim \mathsf{Ber}\left(q_j\right)$ an independent Bernoulli random variable. Then the conclusion remains true conditionally on the event $x_j = 1$ whenever $q_j \leq \varepsilon_j = \mathbb{P}[\texttt{ZEROS}_{<j,N}]$.

Proof. We claim that for any actions $A, A' \in \mathcal{A}$ with $A \subseteq S_{j-1}$ and $A' \not\subseteq S_{j-1}$, we have

$$\mathbb{E}[\theta_A \mid \texttt{ZEROS}_{<j,N}] \leq \mathbb{E}[\theta_{A'} \mid \texttt{ZEROS}_{<j,N}]/2.$$

Thanks to the factor of two, this implies that

$$\arg\max_{A' \in \mathcal{A}} \mathbb{E}[\theta_{A'} \mid x_j = 1] \not\subseteq S_{j-1}$$

for $q_j \leq \varepsilon_j$. (Note that the left-hand side is a deterministic subset of $[d]$ because $\{x_j = 1\}$ is just a single event; we do not condition on all the information of $\mathcal{F}_{T_{j-1}}$.)

First, if $a \notin S_{j-1}$, then $\mathbb{E}[\theta_a \mid \texttt{ZEROS}_{<j,N}] = \mathbb{E}[\theta_a] \geq \tau$ by assumption. It remains to show that

$$\mathbb{E}[\theta_a \mid \texttt{ZEROS}_{<j,N}] \overset{?}{<} \frac{\tau}{2d}, \quad \forall\, a \in S_{j-1}. \tag{4.4}$$

We will show that in fact

$$\mathbb{P}\left[\theta_a \geq \frac{\tau}{5d} \mid \texttt{ZEROS}_{<j,N}\right] \leq \frac{\tau}{5d} \tag{4.5}$$

which implies (4.4) since $\theta_a \leq 1$ almost surely. To see (4.5) we use likelihood ratios to bound probabilities. Note that for any $a \in S_{j-1}$

$$\begin{aligned}
\mathbb{P}\left[\theta_a \geq \frac{\tau}{3d} \mid \texttt{ZEROS}_{<j,N}\right] &\leq \frac{\mathbb{P}[\theta_a \geq \frac{\tau}{3d} \mid \texttt{ZEROS}_{<j,N}]}{\mathbb{P}[\theta_a \leq \frac{\tau}{6d} \mid \texttt{ZEROS}_{<j,N}]} \\
&\leq \left(\frac{1-\frac{\tau}{3d}}{1-\frac{\tau}{6d}}\right)^N \frac{\mathbb{P}[\theta_a \geq \frac{\tau}{3d}]}{\mathbb{P}[\theta_a \leq \frac{\tau}{6d}]} \\
&\leq \frac{e^{-(N\tau/10d)}}{\mathbb{P}[\theta_a \leq \frac{\tau}{6d}]} \\
&\leq \exp\left(\left(10d/\tau\right)^{\alpha} - \frac{N\tau}{10d}\right).
\end{aligned}$$

Since N is at least

$$\left(20d/\tau\right)^{1+\alpha} \log(20d/\tau) \geq 2\left(10d/\tau\right)^{1+\alpha} \log(10d/\tau)$$

the upper bound on $\mathbb{P}\left[\theta_a \geq \frac{\tau}{3d} \mid \texttt{ZEROS}_{<j,N}\right]$ is at most

$$\exp(-\left(10d/\tau\right)^{\alpha} \log(10d/\tau)) \leq (\tau/10d)^{\left(10d/\tau\right)^{\alpha}} \leq \frac{\tau}{3d}.$$

This concludes the proof. □

4.1. Algorithm 1

In light of Lemma 4.1, we can assume without loss of generality that the atoms are ordered such that a_j is BIC conditioned on the event $1 = \max(1_{\texttt{ZEROS}_{<j,N}}, b_j)$, for $b_j \sim \mathsf{Ber}(q_j)$ as in the lemma. Then we will explore the atoms in increasing order. For each $j \in [d]$, our high-level strategy is as follows:

1. On the event $\texttt{ZEROS}_{<j,N}$, obtain N samples of a_j.
2. Exponentially grow the probability to have N-explored a_j.

Similarly to (Sellke & Slivkins, 2022), the main insight is that it is possible to achieve an exponential growth rate. This enables sample efficiency even when the probability from the first phase is exponentially small. We will assume N is large enough to satisfy (4.3); if not, one simply increases N and obtains some additional samples in using the algorithm.

Pseudo-code for our Algorithm 1 is given below; note that when we say to "exploit" given a signal, we simply mean the planner should recommend the greedy action given some signal he generated; this is BIC by definition. The policies π_j are important for the exponential growth strategy and are non-obvious to construct. However much of the algorithm can be understood more easily. First, in the j-th iteration of the outer loop, we collect N samples of atom a_j almost surely, this having been accomplished for all $i < j$ in the previous iterations. The starting point for each loop is to sample a_j on the event $\texttt{ZEROS}_{<j,N}$ that all atoms $i < j$ received no reward during their first N samples. While $\texttt{ZEROS}_{<j,N}$ has tiny probability (denoted ε_j), we grow the probability to sample a_j exponentially using π_j. The sample complexity is largely dictated by the optimal such exponential growth rate, which is the minimax value of a two-player zero sum game; therefore we take π_j to be an approximately optimal strategy for said game.

4.2. The Policy π_j

The policy π_j is constructed via a j-recommendation game, a two-player zero sum game which generalizes the one considered in (Sellke & Slivkins, 2022). In particular, the optimal exponential growth rate of exploration is dictated by its

Algorithm 1: `Semibandit BIC Exploration`

1 **Parameters:** Number of desired samples N satisfying (4.3), Minimax parameter $\lambda = \underline{\lambda}/2d$.
2 **Given:** recommendation policies $\pi_2 \cdots \pi_d$ for PADDED PHASE.
3 **Initialize:** EXPLORATION PHASE for arm 1
4 **for** *each arm* $j = a_1, \ldots, a_d$ **do**
  `// each arm` a_i `for` $i < j$ `has been sampled at least` N `times`
5 Generate $y_j = b_j \sim \mathsf{Ber}(\varepsilon_j)$ and set $x_j = \max(1_{\texttt{ZEROS}_{<j,N}}, b_j)$.
6 EXPLOITATION PHASE with depth N, conditionally on the signal x_j.
  `// If` $x_j = 1$`, then this explores` a_j
7 $p_j \leftarrow \varepsilon_j$.
  `// main loop: exponentially grow the exploration probability`
8 **while** $p_j < 1$ **do**
    `// Try to increase` b_j `to 1, and maintain` $y_j = \mathbb{P}[b_j = 1]$.
9 Generate $b_j \sim \mathsf{Ber}\big(\min(1, \frac{p_j \lambda}{1-p_j})\big)$ and set $z_j = \max(b_j, y_j)$.
10 **if** $z_j = 1$ **then**
11 PADDED PHASE: use policy π_j for N time-steps
12 Otherwise, exploit for N time-steps
13 Update $p_j \leftarrow \min(1, p_j(1+\lambda))$.

minimax value in the perfect information (infinite-sample) setting. In the Appendix, we treat finite-sample approximations to it which are required for constructing π_j.

Definition 5. The *infinite sample j-recommendation game* is a two-player zero sum game played between a *planner* and an *agent*. The players share a common independent prior over the true mean rewards $(\theta_i)_{i \le j}$, and the planner gains access to the values of $(\theta_i)_{i \le j}$. The planner then either: 1) *Recommends an arm $A_j \in \mathcal{A}_j$ containing a_j*, or 2) *Does nothing*.

In the first case, the agent observes $A_j \in \mathcal{A}_j$ and chooses a response A_{-j} from a mixed strategy on $\mathcal{A}_{-j}$ which can depend on A_j. The payoff for the planner is zero if the planner did nothing, and otherwise $\theta_{A_j} - \theta_{A_{-j}}$. Let the minimax value of the game for the planner be $\lambda_j \ge 0$ and

$$\underline{\lambda} \equiv \min_{j \in [d]} \lambda_j.$$

It is clear that $\lambda_j \ge 0$ since the first player can always do nothing. In general, the value of this game measures the planner's ability to convince the agent to try arm j. As in (Sellke & Slivkins, 2022), this can be much larger than the expected regret from not being able to play arm j. The importance of the value $\underline{\lambda}$ is explained by the following lemma, which ensures that with some care, it is possible to increase the j-exploration probability by a factor $1 + \Omega(\underline{\lambda})$ at each stage.

Lemma 4.2. *Let $\lambda = \underline{\lambda}/2d$, so in particular $\lambda \le \lambda_j/2$. For $N \ge \Omega\left(\frac{d^2 \log|\mathcal{A}|}{\lambda^2}\right)$, let z be a signal such that conditioned on $z = 1$, with probability at least $\frac{1}{1+\lambda}$ the first j arms have all been N-explored. Moreover, suppose that the event $z = 1$ is independent of the true mean reward vector $\vec{\mu}$. Then there is a BIC policy which always plays an action in $\mathcal{A}_j$ when $z = 1$.*

Theorem 4.3. *Algorithm 1 is BIC and almost surely obtains N samples of each atom within T timesteps for*

$$T = O\left(\frac{d^{3+\alpha} N^{1+\alpha}}{\underline{\lambda}}\right). \tag{4.6}$$

Proof. For the j-th of the $d-1$ iterations of the outer loop, in the first step Lemma 4.1 implies that if $x_j = 1$, then Line 6 explores a new action a_k for $k \ge j$, which is without loss of generality a_j.

For the inner loop, note that we do not consider the event $\texttt{ZEROS}_{<j,N}$ and only use the "clean" data coming from the event $b_j = 1$. This prevents the data-dependent trajectory of the algorithm from causing unwanted correlations. In particular Line 11 is then BIC by the defining property of π_j since the required independence holds. Other steps of the algorithm are BIC by virtue of being exploitation conditionally on some signal.

To complete the analysis, we have $p_j = 1$ in the final iteration of the inner loop, and so $b_j = 1$ almost surely here. Hence some action in $\mathcal{A}_i$ is chosen each of these times by definition of π_j. This holds for each j, so N samples of each atom are obtained. For sample complexity, the inner loop requires $O\left(\frac{dN \log(1/\varepsilon_j)}{\lambda}\right)$ steps; recalling the bound (4.2) now yields (4.6). The last line in the algorithm contributes a lower order term $O(dN/\underline{\lambda})$. □

4.3. Lower Bound

We now give a sample complexity lower bound for each a_j. The lower bound is inversely proportional to the minimax value $\overline{\lambda}_j \ge \lambda_{j,\infty}$ of the infinite-sample **easy j-recommendation game** in which all values $(\theta_i)_{i \ne j}$ are known exactly to player 1.

Proposition 4.4. *For any $j \in [d]$, the sample complexity to almost surely BIC-explore arm a_j is at least $\Omega\left(\frac{\sigma^2}{d\overline{\lambda}_j}\right) - 2$.*

The question of which values θ_i should be treated as known when proving a lower bound on a_j leads to a mismatch with our upper bound, i.e. we may have $\underline{\lambda} < \overline{\lambda}$ in general. This issue was not present in the multi-armed bandit case

of (Sellke & Slivkins, 2022); without combinatorial action sets, one can actually explore the arms in decreasing order of prior mean. This was justified using the FKG inequality, see Appendix B and Definition 3 therein. However with combinatorial action sets there is no canonical ordering of atoms, which leads to the mismatch above.

References

Agrawal, P. and Tulabandhula, T. Incentivising exploration and recommendations for contextual bandits with payments. In *Multi-Agent Systems and Agreement Technologies: 17th European Conference, EUMAS 2020, and 7th International Conference, AT 2020, Thessaloniki, Greece, September 14-15, 2020, Revised Selected Papers 17*, pp. 159–170. Springer, 2020.

Agrawal, S. and Goyal, N. Analysis of Thompson Sampling for the multi-armed bandit problem. In *25nd*, 2012.

Agrawal, S. and Goyal, N. Further optimal regret bounds for thompson sampling. In *16th*, pp. 99–107, 2013a.

Agrawal, S. and Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In *30th*, pp. 127–135, 2013b.

Bergemann, D. and Morris, S. Information Design: A Unified Perspective. *Journal of Economic Literature*, 57(1): 44–95, 2019.

Bubeck, S. and Liu, C.-Y. Prior-Free and Prior-Dependent Regret Bounds for Thompson Sampling. In *26th*, pp. 638–646, 2013.

Bubeck, S. and Sellke, M. First-Order Bayesian Regret Analysis of Thompson Sampling. *IEEE Transactions on Information Theory*, 2022.

Dong, S. and Van Roy, B. An Information-Theoretic Analysis for Thompson Sampling with many Actions. *Advances in Neural Information Processing Systems*, 31, 2018.

Filippi, S., Cappé, O., Garivier, A., and Szepesvári, C. Parametric Bandits: The Generalized Linear Case. In *24th*, pp. 586–594, 2010.

Frazier, P., Kempe, D., Kleinberg, J., and Kleinberg, R. Incentivizing exploration. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pp. 5–22, 2014.

Guédon, O., Nayar, P., and Tkocz, T. Concentration inequalities and geometry of convex bodies. *Analytical and probabilistic methods in the geometry of convex bodies*, 2:9–86, 2014.

Hamidi, N. and Bayati, M. On Worst-Case Regret of Linear Thompson Sampling. *arXiv preprint arXiv:2006.06790*, 2020.

Hu, X., Ngo, D. D., Slivkins, A., and Wu, Z. S. Incentivizing Combinatorial Bandit Exploration. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Immorlica, N., Mao, J., Slivkins, A., and Wu, Z. S. Bayesian exploration with heterogeneous agents. In *The world wide web conference*, pp. 751–761, 2019.

Immorlica, N., Mao, J., Slivkins, A., and Wu, Z. S. Incentivizing Exploration with Selective Data Disclosure. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pp. 647–648, 2020.

Kamenica, E. Bayesian Persuasion and Information Design. *Annual Review of Economics*, 11:249–272, 2019.

Kannan, R., Lovász, L., and Simonovits, M. Isoperimetric problems for convex bodies and a localization lemma. *Discrete & Computational Geometry*, 13(3):541–559, 1995.

Kannan, S., Kearns, M., Morgenstern, J., Pai, M., Roth, A., Vohra, R., and Wu, Z. S. Fairness incentives for myopic agents. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pp. 369–386, 2017.

Kaufmann, E., Korda, N., and Munos, R. Thompson Sampling: An Asymptotically Optimal Finite-Time Analysis. In *23rd*, pp. 199–213, 2012.

Kremer, I., Mansour, Y., and Perry, M. Implementing the "Wisdom of the Crowd". *Journal of Political Economy*, 122(5):988–1012, 2014.

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, Cambridge, UK, 2020. Versions available at `https://banditalgs.com/` since 2018.

Li, L., Lu, Y., and Zhou, D. Provably optimal algorithms for generalized linear contextual bandits. In *International Conference on Machine Learning*, pp. 2071–2080. PMLR, 2017.

Mansour, Y., Slivkins, A., and Syrgkanis, V. Bayesian Incentive-Compatible Bandit Exploration. *Operations Research*, 68(4):1132–1161, 2020. Preliminary version in *ACM EC 2015*.

Mansour, Y., Slivkins, A., Syrgkanis, V., and Wu, S. Bayesian exploration: Incentivizing exploration in Bayesian games. *Operations Research*, 70(2), 2022. Preliminary version in *ACM EC 2016*.

Neu, G., Olkhovskaya, J., Papini, M., and Schwartz, L. Lifting the Information Ratio: An Information-Theoretic Analysis of Thompson Sampling for Contextual Bandits. *arXiv preprint arXiv:2205.13924*, 2022.

Russo, D. and Van Roy, B. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39 (4):1221–1243, 2014.

Sellke, M. and Slivkins, A. The Price of Incentivizing Exploration: A Characterization via Thompson Sampling and Sample Complexity. *Operations Research*, 2022.

Slivkins, A. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, November 2019. Published with *Now Publishers* (Boston, MA, USA). Also available at `https://arxiv.org/abs/1904.07272`. Latest online revision: Jan 2022.

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

Wang, H., Xu, H., Li, C., Liu, Z., and Wang, H. Incentivizing exploration in linear bandits under information gap. *arXiv preprint arXiv:2104.03860*, 2021.

Wang, S. and Huang, L. Multi-armed bandits with compensation. *Advances in Neural Information Processing Systems*, 31, 2018.

A. Proofs for Section 3

Proof of Lemma 3.1. The mean zero property just says $\mathbb{E}[\mathbb{E}^t[\ell^*]] = \mathbb{E}[\ell^*]$, which follows by the tower rule for conditional expectations. Let $\hat{\ell}$ be the natural ordinary least squares estimator. Theorem 20.5 of (Lattimore & Szepesvári, 2020) implies that for any unit vector $v \in \mathbb{R}^d$, $\ell^* - \hat{\ell}^*$ is $O(\gamma^{-1/2})$-subgaussian (recall that $\hat{\ell}^*$ denotes the ordinary least squares estimator for ℓ^*). From Lemma 2.1 we find that $\ell^* - \tilde{\ell}^*$ is also $O(\gamma^{-1/2})$-subgaussian, for $\tilde{\ell}^*$ a posterior sample for ℓ^* at time t. The result now follows since the sum of two arbitrarily coupled $O(\gamma^{-1/2})$-subgaussian vectors is still $O(\gamma^{-1/2})$-subgaussian (by applying the AM-GM inequality to (2.2)). □

Proof of Lemma 3.2. By Theorem 4.1 of (Kannan et al., 1995),

$$\sqrt{\mathrm{Var}_{\ell\sim\mathcal{U}(\mathcal{K})}\langle \ell, v\rangle} \geq \Omega(\mathrm{width}_v(\mathcal{K})/d).$$

Since $\mathcal{U}(\mathcal{K})$ is a log-concave distribution, Theorem 5.1 of (Guédon et al., 2014) implies

$$\mathbb{E}^{\ell\sim\mathcal{U}(\mathcal{K})}|\langle \ell, v\rangle| \geq \Omega\left(\sqrt{\mathrm{Var}_{\ell\sim\mathcal{U}(\mathcal{K})}\langle \ell, v\rangle}\right).$$

Combining yields the desired estimate. □

A.1. Proofs for Generalized Linear Bandit

As explained in (Filippi et al., 2010), the maximum likelihood estimator (MLE) $\hat{\ell}$ at time t is unique, and is in fact the solution to

$$\sum_{s=1}^{t} \big(R_s - \chi(\langle A_s, \hat{\ell}\rangle)\big) A_s = 0. \tag{A.1}$$

The next result essentially shows that γ-spectral exploration for

$$\gamma \geq \frac{CM_\chi^2 d^2}{m_\chi^4}$$

suffices for the estimation error of the MLE to be subgaussian.

Theorem A.1 (Theorem 1 of (Li et al., 2017))**.** *Suppose γ-spectral exploration has occurred for*

$$\gamma \geq \frac{CM_\chi^2}{m_\chi^4}\left(d^2 + \log(1/\delta)\right).$$

Then the MLE estimate $\hat{\ell}$ has error $(\hat{\ell} - \ell^*)$ satisfying for any fixed $v \in \mathbb{R}^d$:

$$\mathbb{P}\left[|\langle \hat{\ell} - \ell^*, v\rangle| \leq \frac{\|v\|}{m_\chi\sqrt{\gamma}}\right] \geq 1 - \delta.$$

The following result follows directly from Corollary 3.3.

Lemma A.2. *Fix a link function χ. Suppose that $\mathcal{A} \subseteq \mathbb{S}^{d-1}$ is ε-separated, the convex set $\mathcal{K} \subseteq \mathbb{R}^d$ is r-regular, and $\ell^* \sim \mu = \mathcal{U}(\mathcal{K})$. Then for each $A_i, A_j \in \mathcal{A}$, we have*

$$\begin{aligned}
&\mathbb{P}[A_i = A^*(\ell^*)] \geq \left(\frac{\varepsilon}{2}\right)^d, \\
&\mathbb{E}\big[\chi(\langle \ell^*, A_i\rangle) - \chi(\langle \ell^*, A_j\rangle) \mid A^* = A_i\big] \\
&\quad \geq \Omega(r\varepsilon\|A_i - A_j\| m_\chi/d).
\end{aligned}$$

We now extend Theorem 3.5 to the generalized linear bandit model.

Proof of Theorem 3.7. As before it suffices to show that

$$\mathbb{E}\Big[1_{A^*=A_i} \cdot \mathbb{E}^{T_0}\big[\chi(\langle \ell^*, A_i\rangle) - \chi(\langle \ell^*, A_j\rangle)\big]\Big] \geq 0$$

for any action $A_j \neq A_i$. Define

$$\delta_i = \mathbb{P}[A_i = A^*(\ell^*)] \overset{(3.2)}{\geq} \left(\frac{r\varepsilon}{4}\right)^d.$$

We have

$$\begin{aligned}
&\mathbb{E}\Big[1_{A^*=A_i} \cdot \big[\chi(\langle \ell^*, A_i\rangle) - \chi(\langle \ell^*, A_j\rangle)\big]\Big] \\
&= \mathbb{E}\big[1_{A^*=A_i} \cdot \big[\chi(\langle \ell^*, A_i\rangle) - \chi(\langle \ell^*, A_j\rangle)\big]_+\big] \\
&= \delta_i \cdot \mathbb{E}\big[\big[\chi(\langle \ell^*, A_i\rangle) - \chi(\langle \ell^*, A_j\rangle)\big]_+ \mid A^* = A_i\big] \\
&\overset{Lem.\ 3.2}{\geq} \Omega\left(\frac{\delta_i r\varepsilon m_\chi \|A_i - A_j\|}{d}\right).
\end{aligned}$$

We now use $\chi_i = \chi(\langle \ell^*, A_i \rangle)$ to denote the expected reward for action A_i. It remains to upper-bound

$$\mathbb{E}\left[1_{A^*=A_i} \cdot \left|\mathbb{E}^{T_0}[\chi_i - \chi_j] - (\chi_i - \chi_j)\right|\right].$$

By Theorem A.1 and Lemma 2.1, the value $\frac{m_\chi \gamma^{1/2}\left(\mathbb{E}^{T_0}[\chi_i - \chi_j] - (\chi_i - \chi_j)\right)}{\|A_i - A_j\|}$ is centered and $O(1)$-subgaussian (again sans conditioning).[2] Using Lemma 2.2 and the fact that χ is M_χ-Lipschitz, we find

$$\begin{aligned}&\mathbb{E}\left[1_{A^*=A_i} \cdot \left|\mathbb{E}^{T_0}[\chi_i - \chi_j] - (\chi_i - \chi_j)\right|\right] \\ &\leq O\left(\frac{\delta_i M_\chi \|A_i - A_j\|}{m_\chi}\sqrt{\frac{\log(1/\delta_i)}{\gamma}}\right).\end{aligned}$$

Since we assumed

$$\gamma \geq \frac{C M_\chi^2 d^3 \log(4/r\varepsilon)}{r^2 m_\chi^4 \varepsilon^2}$$

for a large enough absolute constant C, we obtain

$$\begin{aligned}&\mathbb{E}\left[1_{A^*=A_i} \cdot \left|\mathbb{E}^{T_0}[\chi_i - \chi_j] - (\chi_i - \chi_j)\right|\right] \\ &\leq O\left(\frac{\delta_i M_\chi \|A_i - A_j\|}{m_\chi}\sqrt{\frac{\log(1/\delta_i)}{\gamma}}\right) \\ &\leq \Omega\left(\frac{\delta_i r \varepsilon m_\chi \|A_i - A_j\|}{d}\right) \\ &\leq \mathbb{E}\Big[1_{A^*=A_i} \cdot \left[\chi(\langle \ell^*, A_i \rangle) - \chi(\langle \ell^*, A_j \rangle)\right]\Big]\end{aligned}$$

In the second step we used $\log(1/\delta_i) \leq d \log(4/r\varepsilon)$ which follows from (3.2). This concludes the proof. □

B. Proofs for Counterexamples

Proof of Proposition 3.8. We show that (3.5) holds conditionally on $A^{(1)} \neq A_3$, and that equality holds conditionally on $A^{(1)} = A_3$. Together these imply the result.

First if $A^{(1)} = A_1$ or $A^{(2)} = A_2$, then at time 2 we have learned the first coordinate of ℓ^*. So in this case, if $\mathbb{P}^2[A^* = A_2]$ then $\mathbb{E}^2[\ell^* = (\mu_1, 0)]$ for some $\mu_1 > 0$ almost surely. In particular, on this event

$$\mathbb{E}^2[\langle \ell^*, A_3 - A_1 \rangle] > 0$$

so we conclude that

$$\mathbb{E}[\mu_3 \mid (A^{(1)}, A^{(2)}) = (A_1, A_1)] > \mathbb{E}[\mu_1 \mid (A^{(1)}, A^{(2)}) = (A_1, A_2)],$$
$$\mathbb{E}[\mu_3 \mid (A^{(1)}, A^{(2)}) = (A_2, A_1)] > \mathbb{E}[\mu_1 \mid (A^{(1)}, A^{(2)}) = (A_2, A_1)].$$

Next if $A^{(1)} = A_3$, then at time 2 we have learned the value $\mu_3 = \langle \ell^*, A_3 \rangle$. We claim that for any μ_3,

$$\mathbb{P}^2[A^* = A_1 \mid \mu_3] = \mathbb{P}^2[A^* = A_1 \mid -\mu_3]. \tag{B.1}$$

Indeed it is easy to see that $A^* = A_1$ if and only if ℓ^* lies in the angles spanned by the outward normal cone to A_1 for the triangle $A_1 A_2 A_3$. In other words, we must have

$$\arg(\ell^*) \in \left[-\pi/2, \tan^{-1}(-4/3)\right].$$

By construction, $\|A_1\| = \|A_1 - A_3\|$ and so the angle bisector to this normal cone is orthogonal to A_3 by elementary geometry. It is easy to see that (B.1) now follows, and implies

$$\begin{aligned}&\mathbb{E}[\mu_3 \mid (A^{(1)}, A^{(2)}) = (A_3, A_1)] \\ &\quad = \mathbb{E}[\mu_1 \mid (A^{(1)}, A^{(2)}) = (A_3, A_1)].\end{aligned}$$

Averaging over the conditioning on $A^{(1)}$ now yields the result (since $\mathbb{P}[A^{(1)} = A_1] > 0$, say). □

B.1. Exponential Lower Bound for Initial Exploration

First, we observe that setting $\mathcal{A} = \mathcal{E}(\mathcal{P})$ to be the set of extreme points of $\mathcal{P}$ is essentially equivalent to taking $\mathcal{A} = \mathcal{P}$ itself. The proof is easily seen to generalize to any convex polytope. This means that our lower bound below applies to the classes of convex action sets as well as separated action sets.

Proposition B.1. *Suppose there exists a BIC algorithm $\mathcal{A}$ which explores in $\mathcal{P}$ and achieves γ-spectral exploration almost surely within T timesteps. Then there exists a BIC algorithm $\mathcal{A}'$ which explores in $\mathcal{A} = \mathcal{E}(\mathcal{P})$ and achieves γ-spectral exploration almost surely within Td timesteps.*

Proof. We show how to simulate each step of $\mathcal{A}$ using exactly d steps of $\mathcal{A}'$, in a BIC way. For each $t \in \{0, 1, \ldots, T-1\}$, at the start of time-step $td + 1$ we will have a simulated version of $\mathcal{A}$ which has received some artifical feedback and recommends an action $A^{(t)} \in \mathcal{P}$. To construct $\mathcal{A}'$ we do the following:

1. Write $A^{(t)} = \sum_{i=1}^d p_{t,i} \widetilde{A}_{t,i}$ as a convex combination of d not-necessarily-distinct $\widetilde{A}_{t,i} \in \mathcal{A}$, with $p_{t,i} \geq 0$ and $\sum_{i=1}^d p_{t,i} = 1$.
2. At timestep $td + i$ for $i \in \{1, 2, \ldots, d\}$, play action $\widetilde{A}_{t,i}$ and receive reward $\widetilde{r}_{t,i} \in \{0, 1\}$.
3. After timestep $(t+1)d$, choose $r_t = \widetilde{r}_{t,i}$ with probability $p_{t,i}$ independently of the past, and take this as the feedback for $\mathcal{A}$.

[2] Technically we need to do the estimates more carefully, but basically the $d^2 + \log(1/\delta)$ in Theorem A.1 is just $O(d^2)$ unless $\delta \leq e^{-d^2}$. And we should only need $\delta \approx e^{-d}$.

It is easy to see that r_t has the correct expected value by linearity of rewards for the linear bandit problem. It then follows that after receiving $(r_1, \ldots, r_{t-1})$, the recommendation $A^{(t)} \in \mathcal{P}$ is BIC within $\mathcal{P}$ by the BIC property of $\mathcal{A}$. By linearity it follows that conditioned on observing $A^{(t)}$, each extreme point $\widetilde{A}_{t,i}$ is also a BIC recommendation, since they all must have the same conditional mean reward.

It remains to show that $\mathcal{A}'$ also achieves γ-spectral exploration. Indeed

$$\sum_{i=1}^{d} p_{t,i} \widetilde{A}_{t,i}^{\otimes 2} \succeq (A^{(t)})^{\otimes 2}$$

follows by testing against any $v^{\otimes 2}$ and using Cauchy-Schwarz. This completes the proof. □

Based on Proposition B.1, we state Proposition 3.9 below in the discrete action set setting $\mathcal{A} = \mathcal{E}(\mathcal{P})$, with the understanding that it extends also to the convex action set $\mathcal{A} = \mathcal{P}$. Note that for $\mathcal{P}$ as above, all extreme points satisfy $x_d = 0$ except for $\pm\hat{A} = \left(0, 0, \ldots, 0, \pm\frac{1}{2\sqrt{d}}\right)$, and so γ-spectral exploration for any $\gamma > 0$ requires exploring these actions. Our prior distribution $\ell^* \sim \mu$ will be uniform over the "biased" convex body $\mathcal{K}/\sqrt{d}$ for

$$\mathcal{K} \equiv [-0.5, 1]^{d-1} \times [-1, 1].$$

Proof of Proposition 3.9. Consider a modification $\mathbb{A}'$ of $\mathbb{A}$ which changes all plays of $\pm\hat{A}$ to the prior-optimal action $A_1 = \frac{(1,1,\ldots,1,0)}{2\sqrt{d}}$. If $\mathbb{A}'$ and $\mathbb{A}$ play different actions at m times in expectation, then by definition the expected total rewards satisfy

$$R_T(\mathbb{A}') + \varepsilon m \geq R_T(\mathbb{A}). \tag{B.2}$$

Recall that one of $\pm\hat{A}$ must be explored at least once regardless of ℓ^*. Moreover at the first such modification time (denoted by τ),

$$\mathbb{E}^\tau[\langle \ell^*, \hat{A}\rangle] = \mathbb{E}^\tau[\langle \ell^*, -\hat{A}\rangle] = 0.$$

This implies

$$\begin{aligned}\mathbb{E}[r_\tau(\mathbb{A}') + \varepsilon - r_\tau(\mathbb{A})] &\leq \varepsilon - \mathbb{E}[\langle \ell^*, A_1\rangle] \\ &\leq \varepsilon - 1/3 \leq -1/6.\end{aligned}$$

Moreover $\mathbb{A}'$ makes at most $T-1$ additional modifications; at each time s that such a modification occurs,

$$\mathbb{E}^s[r_s(\mathbb{A}') + \varepsilon - r_s(\mathbb{A})] \geq \left(|\mathbb{E}^s[\langle \ell^*, \hat{A}\rangle]| + \varepsilon - \mathbb{E}^s[\langle \ell^*, A_1\rangle]\right)_+.$$

Summing over times s and applying Jensen in the second inequality, we find

$$\begin{aligned}&R_T(\mathbb{A}') + \varepsilon m - R_T(\mathbb{A}) \\ &\leq -1/6 + \sum_{s=1}^{T} \mathbb{E}\left[\left(|\mathbb{E}^s[\langle \ell^*, \hat{A}\rangle]| + \varepsilon - \mathbb{E}^s[\langle \ell^*, A_1\rangle]\right)_+\right] \\ &\leq -1/6 + T \cdot \mathbb{E}\left[\left(|\langle \ell^*, \hat{A}\rangle| + \varepsilon - \langle \ell^*, A_1\rangle\right)_+\right].\end{aligned}$$

A simple Chernoff estimate shows that

$$\begin{aligned}&\mathbb{E}\left[\left(|\langle \ell^*, \hat{A}\rangle| + \varepsilon - \langle \ell^*, A_1\rangle\right)_+\right] \\ &\leq \mathbb{E}\left[\left(\frac{1}{10} - \langle \ell^*, A_1\rangle\right)_+\right] \leq e^{-\Omega(d)}.\end{aligned} \tag{B.3}$$

Recalling (B.2), we conclude $T \geq e^{\Omega(d)}$ as desired. □

C. j-Recommendation Game

We first formally define the j-recommendation game. We proceed more generally than in the main body, defining it relative to any d-tuple $(N_1, \ldots, N_d)$. We recall the definition of the static σ-algebra $\mathcal{G}_{N_1,\ldots,N_d}$ which is generated by N_i samples of each atom i. When considering arm j, we will always have $N_k = 0$ for all $k > j$. If $N_i = N$ for all i we recover the (j, N)-recommendation game as defined in the main body. The definition of (j, N)-informed generalizes readily to $(j, \mathcal{G})$-informed.

Definition 6. The *j-recommendation game* is a two-player zero sum game played between a *planner* and an *agent*. The players share a common independent prior over the true mean rewards $(\theta_i)_{i \leq j}$, and the planner gains access to the static σ-algebra $\mathcal{G} = \mathcal{G}_{(N_1,\ldots,N_j)}$. The planner then either:

1. Recommends an arm $A_j \in \mathcal{A}_j$ containing A_j.
2. Does nothing.

In the first case, the agent observes $A_j \in \mathcal{A}_j$ and chooses a response A_{-j} from a mixed strategy on $\mathcal{A}_{-j}$ which can depend on A_j. The payoff for the planner is zero if the planner did nothing, and otherwise $\theta_{A_j} - \theta_{A_{-j}}$.

Definition 7. A planner strategy π for the j-recommendation game is said to be (j, λ)-padded, for a function $\lambda : \mathcal{A}_j \to \mathbb{R}_{\geq 0}$, if for each $A_j \in \mathcal{A}_j$ and $A_{-j} \in \mathcal{A}_{-j}$:

$$\min_{A_{-j} \in \mathcal{A}_{-j}} \mathbb{E}[(\theta_{A_j} - \theta_{A_{-j}}) \cdot 1_{\pi = A_j}] = \lambda_{A_j}. \tag{C.1}$$

Such a strategy has total j-padding at least

$$\lambda_j \equiv \sum_{A_j \in \mathcal{A}_j} \lambda_{A_j}.$$

Note that since $\lambda_{A_j} \geq 0$, a planner strategy need not be (j, λ)-padded for any λ, and hence need not have a total padding value.

We can without loss of generality view all j-recommendation game strategies as depending only on the posterior means of each arm conditional on $\mathcal{G}$. In the below, we set $\widetilde{\theta}_i = \mathbb{E}[\theta_i|\mathcal{G}]$ for the relevant static σ-algebra $\mathcal{G}$. Given a planner strategy in the j-recommendation game, we naturally obtain a corresponding $(j, \mathcal{G})$-informed policy for our original problem in which we recommend arm A_j when the planner would, and recommend the $\mathcal{G}$-conditional-expectation-maximizing arm otherwise. The key point of this game is as follows.

Lemma C.1. *If a strategy π for the planner in the j-recommendation game has minimax value λ_j, then its total j-padding value equals λ_j.*

Proof. Simply note that the left-hand side of (C.1) is the contribution of playing A_j to the minimax value of π for the planner. □

Lemma C.2. *Suppose there exists a planner strategy π in the j-recommendation game with total j-padding at least λ_j. Let $p \geq d/(d+\lambda_j)$ and let $b_j \sim \mathsf{Ber}(p)$ be independent of the signal.*

Consider a modified game where the planner observes $\mathcal{G}$ if and only if $b_j = 1$ (and where the agent never observes b_j). There exists a BIC planner strategy $\widetilde{\pi}$ in this game such that if $b_j = 0$, then $\widetilde{\pi}$ always recommends an action in $\mathcal{A}_j$.

Proof. The planner follows π if $b_j = 1$. Conditioned on $b_j = 0$, the planner plays from the probability distribution q on $\mathcal{A}_j$ given by

$$q(A_j) = \frac{\lambda_{A_j}}{\lambda_j}.$$

We call this modified strategy $\hat{\pi}$. Noting that $\mathbb{E}[\theta_{A_j} - \theta_{A_{-j}}] > -d$, we find that if the agent observes $\hat{\pi} = A_j$, then A_j has higher conditional mean reward than any $A_{-j} \in \mathcal{A}_{-j}$:

$$\begin{aligned}
&\mathbb{E}[(\theta_{A_j} - \theta_{A_{-j}}) \cdot 1_{\hat{\pi}=A_j}] \\
&= \mathbb{E}[(\theta_{A_j} - \theta_{A_{-j}}) \cdot 1_{b_j=1,\pi=A_j}] \\
&\qquad + (1-p)q(A_j) \cdot \mathbb{E}[\theta_{A_j} - \theta_{A_{-j}}] \\
&> p\lambda_{A_j} - jd(1-p)q(A_j).
\end{aligned}$$

Since $p \geq d/(d+\lambda_j)$, it easily follows that the last expression is non-negative.

Therefore the greedy strategy conditioned on observing $\hat{\pi} = A_j$ is to choose some $A'_j \in \mathcal{A}_j$ (any $A_{-j} \in \mathcal{A}_{-j}$ is inferior to A_j hence suboptimal). The claimed BIC strategy $\widetilde{\pi}$ is given by playing A'_j when $\hat{\pi} = A_j$. Note that by definition, $\widetilde{\pi}$ exploits conditioned on the signal from $\hat{\pi}$, hence is BIC. □

Next we upper bound the number of samples required in $\mathcal{G}$ to ensure $\lambda \geq \Omega(G_{\mathrm{pad}})$. The point is that the game has an $N \to \infty$ limit that can be approximated by coupling. To this end, we let $\lambda_{j,\infty}$ be the minimax value for the $N = \infty$ version where the planner observes the exact values of $\theta_1, \ldots, \theta_j$. We note that in the simpler setting of (Sellke & Slivkins, 2022), $\lambda_{j,\infty} = \inf_{q \in \Delta_j} \mathbb{E}[(\theta_j - \theta_q)_+]$ has a simple interpretation by using the minimax theorem to choose the agent's strategy first. It is similarly possible to apply the minimax theorem here but the result is not particularly interpretable since the planner chooses a different mixed strategy for each $A_j \in \mathcal{A}_j$.

Lemma C.3. *For $N \geq \Omega\left(\frac{d^2 \log|\mathcal{A}|}{\lambda_{j,\infty}^2}\right)$, there is an N-informed policy with minimax value at least $\lambda_{j,\infty}/2$.*

Proof. Given the signal, first generate a posterior sample $(\hat{\theta}_1, \ldots, \hat{\theta}_j)$ for $\theta_1, \ldots, \theta_j$ and then consider the infinite-sample policy π_∞. We claim that π_∞ has minimax value at least $\lambda_{j,\infty}/2$

To find the minimax value of this policy, we use Lemmas 2.1 and 2.2. In particular they imply that for each (A, A_j),

$$\mathbb{E}[|\hat{\theta}_A - \theta_A| \cdot 1_{\pi=A_j}] \leq CdN^{-1/2} p_\pi(A_j) \sqrt{\log(1/p_\pi(A_j))}.$$

Taking $A = A_j$ and $A = A_{-j}$, we find that the minimax value of this policy is at least

$$\begin{aligned}
&\lambda_{j,\infty} - O(dN^{-1/2}) \cdot \sum_A p_\pi(A) \sqrt{\log(1/p_\pi(A))} \\
&\geq \lambda_{j,\infty} - O(dN^{-1/2}\sqrt{\log|\mathcal{A}|}).
\end{aligned}$$

The last expression is at least $\lambda_{j,\infty}/2$ for $N \geq \Omega\left(\frac{d^2 \log|\mathcal{A}|}{\lambda_{j,\infty}^2}\right)$ as desired. □

Proof of Lemma 4.2. By Lemma C.3, conditioned on $z = 1$ the minimax value of the game is at least $\lambda_{j,\infty}/2$. We now apply Lemma C.2, with b_j the signal z. The policy $\widetilde{\pi}$ is the desired one, proving the lemma. (Note that by definition $\lambda = \underline{\lambda}/2d$ in Lemma 4.2.) □

Proof of Proposition 4.4. We assume $T < T_0$ and consider two agents. The first, an **obedient agent**, simply obeys recommendations. The second, a **j-avoiding agent**, when recommended to play some $A_j \in \mathcal{A}_j$, instead plays from cA_{-j} using the minimax optimal response for the easy j-recommendation game. We claim that if $T < T_0$ and almost surely some $A_j \in \mathcal{A}_j$ must be recommended, then the j-avoiding agent attains larger expected reward than the

obedient one against any planner. This will contradict the BIC property. First note that each time-step, by definition the alternative agent loses at most $\overline{\lambda}_j$ per round compared to an obedient agent. To get the lower bound, we will show that the alternative agent does better the first time t_j that A_j is recommended (i.e. the random value t_j is minimal such that $A_j \in A_{t_j}$).

At this time, note that

$$\theta_{A_j} - \mathbb{E}[\theta_{A_{-j}} \mid A_j]$$

has standard deviation at least σ for each $A_j \in \mathcal{A}_j$ and any agent strategy. (A_j can depend on $\theta_1, \ldots, \theta_{j-1}$ but is independent of θ_j.) In particular, we have

$$\begin{aligned}
&\mathbb{E}\left[\left(\theta_{A_{t_j}} - \mathbb{E}[\theta_{A_{-j}} \mid A_{t_j}]\right)_+\right] \\
&\quad + \mathbb{E}\left[\left(\theta_{A_{t_j}} - \mathbb{E}[\theta_{A_{-j}} \mid A_{t_j}]\right)_-\right] \\
&= \mathbb{E}\left[\left|\theta_{A_{t_j}} - \mathbb{E}[\theta_{A_{-j}} \mid A_{t_j}]\right|\right] \\
&\geq \mathbb{E}\left[\left(\theta_{A_{t_j}} - \mathbb{E}[\theta_{A_{-j}} \mid A_{t_j}]\right)^2\right]/d \\
&\geq \sigma^2/d.
\end{aligned}$$

The first term is at most $\overline{\lambda}_j$ for any strategy, so

$$\begin{aligned}
&\mathbb{E}\left[\theta_{A_{t_j}} - \mathbb{E}[\theta_{A_{-j}} \mid A_{t_j}]\right] \\
&= \mathbb{E}\left[\left(\theta_{A_{t_j}} - \mathbb{E}[\theta_{A_{-j}} \mid A_{t_j}]\right)_+\right] \\
&\qquad - \mathbb{E}\left[\left(\theta_{A_{t_j}} - \mathbb{E}[\theta_{A_{-j}} \mid A_{t_j}]\right)_-\right] \\
&\leq 2\overline{\lambda}_j - \frac{\sigma^2}{d}.
\end{aligned}$$

Thus in total, the j-avoiding agent outperforms the obedient one by

$$\frac{\sigma^2}{d} - (T+2)\overline{\lambda}_j,$$

which is non-positive if recommendations are BIC. This implies the claimed estimate. □