Gamification of Pure Exploration for Linear Bandits

Rémy Degenne [*1] **Pierre Ménard** [*2] **Xuedong Shang** [3] **Michal Valko** [4]

Abstract

We investigate an active *pure-exploration* setting, that includes *best-arm identification*, in the context of *linear stochastic bandits*. While asymptotically optimal algorithms exist for standard *multi-arm bandits*, the existence of such algorithms for the best-arm identification in linear bandits has been elusive despite several attempts to address it. First, we provide a thorough comparison and new insight over different notions of optimality in the linear case, including G-optimality, transductive optimality from optimal experimental design and asymptotic optimality. Second, we design the first asymptotically optimal algorithm for fixed-confidence pure exploration in linear bandits. As a consequence, our algorithm naturally bypasses the pitfall caused by a simple but difficult instance, that most prior algorithms had to be engineered to deal with explicitly. Finally, we avoid the need to fully solve an optimal design problem by providing an approach that entails an efficient implementation.

1. Introduction

Multi-armed bandits (MAB) probe fundamental *exploration-exploitation* trade-offs in sequential decision learning. We study the pure exploration framework, from among different MAB models, which is subject to the maximization of information gain after an exploration phase. We are particularly interested in the case where noisy linear payoffs depending on some regression parameter θ are assumed. Inspired by Degenne et al. (2019), we treat the problem as a *two-player zero-sum* game between the agent and the nature (in a sense described in Section 2), and we search for algorithms that are able to output a correct answer with high confidence to a given query using as few samples as possible.

Since the early work of Robbins (1952), a great amount of literature explores MAB in their standard stochastic setting with its numerous extensions and variants. Even-Dar et al. (2002) and Bubeck et al. (2009) are among the first to study the pure exploration setting for stochastic bandits. A non-exhaustive list of pure exploration game includes best-arm identification (BAI), top-m identification (Kalyanakrishnan & Stone, 2010), threshold bandits (Locatelli et al., 2016), minimum threshold (Kaufmann et al., 2018), signed bandits (Ménard, 2019), pure exploration combinatorial bandits (Chen et al., 2014), Monte-Carlo tree search (Teraoka et al., 2014), etc.

In this work, we consider a general pure-exploration setting (see Appendix D for details). Nevertheless, for the sake of simplicity, in the main text we primarily focus on BAI. For stochastic bandits, BAI has been studied within two major theoretical frameworks. The first one, *fixed-budget* BAI, aims at minimizing the probability of misidentifying the optimal arm within a given number of pulls (Audibert & Bubeck, 2010). In this work, we consider another setting, *fixed-confidence* BAI, introduced by Even-dar et al. (2003). Its goal is to ensure that the algorithm returns a wrong arm with probability less than a given risk level, while using a small total number of samples before making the decision. Existing fixed-confidence algorithms are either elimination-based such as `SuccessiveElimination` (Karnin et al., 2013), rely on confidence intervals such as `UGapE` (Gabillon et al., 2012), or follow plug-in estimates of the optimal pulling proportions by a lower bound such as `Track-and-Stop` (Garivier & Kaufmann, 2016). We pay particular attention to the first two since they have been extended to the linear setting, which is the focus of this paper. In particular, a natural extension of pure exploration to linear bandits. Linear bandits were first investigated by Auer (2002) in the stochastic setting for *regret minimization* and later considered for fixed-confidence BAI problems by Soare et al. (2014).

Linear bandits. We consider a *finite-arm linear bandit* problem, where the collection of arms $\mathcal{A} \subset \mathbb{R}^d$ is given with $|\mathcal{A}| = A$, and spans $\mathbb{R}^d$. We assume that $\forall a \in \mathcal{A}, \|a\| \leq L$, where $\|a\|$ denotes the Euclidean norm of the vector a. The

[*]Equal contribution [1]INRIA - DIENS - PSL Research University, Paris, France [2]INRIA [3]INRIA - Université de Lille [4]DeepMind Paris. Correspondence to: Rémy Degenne <remydegenne@gmail.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

learning protocol goes as follows: for each round $1 \leq t \leq T$, the agent chooses an arm $a_t \in \mathcal{A}$ and observes a noisy sample

$$Y_t = \langle \theta, a_t \rangle + \eta_t \,,$$

where $\eta_t \sim \mathcal{N}(0, \sigma^2)$ is conditionally independent from the past and θ is some unknown regression parameter. For the sake of simplicity, we use $\sigma^2 = 1$ in the rest of this paper.

Pure exploration for linear bandits. We assume that θ belongs to some set $\mathcal{M} \subset \mathbb{R}^d$ known to the agent. For each parameter a *unique correct answer* is given by the function $i^\star : \mathcal{M} \to \mathcal{I}$ among the $I = |\mathcal{I}|$ possible ones (the extension of pure exploration to multiple correct answers is studied by Degenne & Koolen 2019). Given a parameter θ, the agent then aims to find the correct answer $i^\star(\theta)$ by interacting with the finite-armed linear bandit environment parameterized by θ.

In particular, we detail the setting of BAI for which the objective is to identify the arm with the largest mean. That is, the correct answer is given by $i^\star(\theta) = a^\star(\theta) := \operatorname{argmax}_{a \in \mathcal{A}} \langle \theta, a \rangle$ for $\theta \in \mathcal{M} = \mathbb{R}^d$ and the set of possible correct answers is $\mathcal{I} = \mathcal{A}$. We provide other pure-exploration examples in Appendix D.

Algorithm. Let $\mathcal{F}_t = \sigma(a_1, Y_1, \dots, a_t, Y_t)$ be the information available to the agent after t round. A deterministic pure-exploration algorithm under the fixed-confidence setting is given by three components: (1) a *sampling rule* $(a_t)_{t \geq 1}$, where $a_t \in \mathcal{A}$ is $\mathcal{F}_{t-1}$-measurable, (2) a *stopping rule* τ_δ, a stopping time for the filtration $(\mathcal{F}_t)_{t \geq 1}$, and (3) a *decision rule* $\hat{\imath} \in \mathcal{I}$ which is $\mathcal{F}_{\tau_\delta}$-measurable. Non-deterministic algorithms could also be considered by allowing the rules to depend on additional internal randomization. The algorithms we present are deterministic.

δ-correctness and fixed-confidence objective. An algorithm is δ-correct if it predicts the correct answer with probability at least $1 - \delta$, precisely if $\mathbb{P}_\theta\big(\hat{\imath} \neq i^\star(\theta)\big) \leq \delta$ and $\tau_\delta < +\infty$ almost surely for all $\theta \in \mathcal{M}$. Our goal is to find a δ-correct algorithm that minimizes the *sample complexity*, that is, $\mathbb{E}_\theta[\tau_\delta]$ the expected number of sample needed to predict an answer.

Pure exploration (in particular BAI) for linear bandits has been previously studied by Soare et al. (2014); Tao et al. (2018); Xu et al. (2018); Zaki et al. (2019); Fiez et al. (2019); Kazerouni & Wein (2019). They all consider the fixed-confidence setting. To the best of our knowledge, only Hoffman et al. (2014) study the problem with a fixed-budget.

Beside studying fixed-confidence sample complexity, Garivier & Kaufmann (2016) and some subsequent works (Qin et al., 2017; Shang et al., 2020) investigate a general criterion of judging the optimality of a BAI sampling rule: Algorithms that achieve the minimal sample complexity when δ tends to zero are called asymptotically optimal. Ménard (2019) and Degenne et al. (2019) further study the problem in a game theoretical point of view, and extend the asymptotic optimality to the general pure exploration for structured bandits. Note that a naive adaptation of the algorithm proposed by Degenne et al. (2019) may not work smoothly in our setting. In this paper we use some different confidence intervals that benefit better from the linear structure.

Contributions. **1)** We provide new insights on the complexity of linear pure exploration bandits. In particular, we relate the asymptotic complexity of the BAI problem and other measures of complexity inspired by optimal design theory, which were used in prior work. **2)** We develop a saddle-point approach to the lower bound optimization problem, which also guides the design of our algorithms. In particular we highlight a new insight on a convex formulation of that problem. It leads to an algorithm with a more direct analysis than previous lower-bound inspired methods. **3)** We obtain two algorithms for linear pure exploration bandits in the fixed-confidence regime. Their sample complexity is asymptotically optimal and their empirical performance is competitive with the best existing algorithms.

2. Asymptotic Optimality

In this section we extend the lower bound of Garivier & Kaufmann (2016), to hold for *pure exploration in finite-armed linear bandit* problems.

Alternative. For any answer $i \in \mathcal{I}$ we define *the alternative to i*, denoted by $\neg i$ the set of parameters where the answer i is not correct, i.e. $\neg i := \{\theta \in \mathcal{M} : \; i \neq i^\star(\theta)\}$.

We also define, for any $w \in (\mathbb{R}^+)^A$, the design matrix

$$V_w := \sum_{a \in \mathcal{A}} w^a a a^\intercal \,.$$

Further, we define $\|x\|_V := \sqrt{x^\intercal V x}$ for $x \in \mathbb{R}^d$ and a symmetric positive matrix $V \in \mathbb{R}^{d \times d}$. Note that it is a norm only if V is positive definite. We also denote by Σ_K the probability simplex of dimension $K - 1$ for all $K \geq 2$.

Lower bound. We have the following non-asymptotic lower bound, proved in Appendix C, on the sample complexity of any δ-correct algorithm. This bound was already proved by Soare et al. (2014) for the BAI example.

Theorem 1. *For all δ-correct algorithms, for all $\theta \in \mathcal{M}$,*

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_\theta[\tau_\delta]}{\log(1/\delta)} \geq T^\star(\theta) \,,$$

where the characteristic time $T^\star(\theta)$ *is defined by*

$$T^\star(\theta)^{-1} := \max_{w\in\Sigma_A} \inf_{\lambda\in\neg i^\star(\theta)} \frac{1}{2}\|\theta-\lambda\|^2_{V_w}\,.$$

In particular, we say that a δ-correct algorithm is asymptotically optimal if for all $\theta\in\mathcal{M}$,

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_\theta[\tau_\delta]}{\log(1/\delta)} \le T^\star(\theta)\,.$$

As noted in the seminal work of Chernoff (1959), the complexity $T^\star(\theta)^{-1}$ is the value of a fictitious zero-sum game between the agent choosing an optimal proportion of allocation of pulls w and a second player, the nature, that tries to fool the agent by choosing the most confusing alternative λ leading to an incorrect answer.

Minimax theorems. Using Sion's minimax theorem we can invert the order of the players if we allow nature to play mixed strategies

$$\begin{aligned} T^\star(\theta)^{-1} &= \max_{w\in\Sigma_A} \inf_{\lambda\in\neg i^\star(\theta)} \frac{1}{2}\|\theta-\lambda\|^2_{V_w} \qquad (1)\\ &= \inf_{q\in\mathcal{P}(\neg i^\star(\theta))} \max_{a\in\mathcal{A}} \frac{1}{2}\mathbb{E}_{\lambda\sim q}\|\theta-\lambda\|^2_{aa^\intercal}\,, \end{aligned}$$

where $\mathcal{P}(\mathcal{X})$ denotes the set of probability distributions over the set $\mathcal{X}$. The annoying part in this formulation of the characteristic time is that the set $\neg i^\star(\theta)$ where the nature plays is a priori unknown (as the parameter is unknown to the agent). Indeed, to find an asymptotically optimal algorithm one should somehow solve this minimax game. But it is easy to remove this dependency noting that $\inf_{\lambda\in\neg i}\|\theta-\lambda\|=0$ for all $i\neq i^\star(\theta)$,

$$T^\star(\theta)^{-1} = \max_{i\in\mathcal{I}} \max_{w\in\Sigma_A} \inf_{\lambda\in\neg i} \frac{1}{2}\|\theta-\lambda\|^2_{V_w}\,.$$

Now we can see the characteristic time $T^\star(\theta)^{-1}$ as the value of an other game where the agent plays a proportion of allocation of pulls w *and* an answer i. The agent could also use mixed strategies for the answer which leads to

$$\begin{aligned} T^\star(\theta)^{-1} &= \max_{\rho\in\Sigma_I} \max_{w\in\Sigma_A} \frac{1}{2}\sum_{i\in\mathcal{I}} \inf_{\lambda^i\in\neg i} \rho_i\|\theta-\lambda^i\|^2_{V_w}\\ &= \max_{\rho\in\Sigma_I} \max_{w\in\Sigma_A} \inf_{\widetilde{\lambda}\in\prod_i(\neg i)} \frac{1}{2}\sum_{i\in\mathcal{I}} \rho_i\|\theta-\widetilde{\lambda}^i\|^2_{V_w}\,, \end{aligned}$$

where $\prod_{i\in\mathcal{I}}(\neg i)$ denotes the Cartesian product of the alternative sets $\neg i$. But the function that appears in the value of the new game is not anymore convex in (w,ρ) and Sion's minimax theorem does not apply anymore. We can however convexify the problem by letting the agent to play a distribution $\widetilde{w}\in\Sigma_{AI}$ over the arm-answer pairs (a,i), see Lemma 1 below proved in Appendix C.

Lemma 1. *For all* $\theta\in\mathcal{M}$,

$$\begin{aligned} T^\star(\theta)^{-1} &= \max_{\widetilde{w}\in\Sigma_{AI}} \inf_{\widetilde{\lambda}\in\prod_i(\neg i)} \frac{1}{2}\sum_{(a,i)\in\mathcal{A}\times\mathcal{I}} \widetilde{w}^{a,i}\|\theta-\widetilde{\lambda}^i\|^2_{aa^\intercal}\\ &= \inf_{\widetilde{q}\in\prod_i\mathcal{P}(\neg i)} \frac{1}{2}\max_{(a,i)\in\mathcal{A}\times\mathcal{B}} \mathbb{E}_{\widetilde{\lambda}^i\sim\widetilde{q}^i}\|\theta-\widetilde{\lambda}^i\|^2_{aa^\intercal}\,. \end{aligned}$$

Thus in this formulation the characteristic time is the value of a fictitious zero-sum game where the agent plays a distribution $\widetilde{w}\in\Sigma_{AI}$ over the arm-answer pairs $(a,i)\in\mathcal{A}\times\mathcal{I}$ and nature chooses an alternative $\widetilde{\lambda}^i\in\neg i$ for all the answers $i\in\mathcal{I}$. The algorithm `LinGame-C` that we propose in Section 3 is based on this formulation of the characteristic time whereas algorithm `LinGame` is based on the formulation of Theorem 1.

Best-arm identification complexity. The inverse of the characteristic time of Theorem 1 specializes to

$$T^\star(\theta)^{-1} = \max_{w\in\Sigma_A} \min_{a\neq a^\star(\theta)} \frac{\big\langle\theta, a^\star(\theta)-a\big\rangle^2}{2\|a^\star(\theta)-a\|^2_{V_w^{-1}}}$$

for BAI (see Appendix D.1 for a proof). It is also possible to explicit the characteristic time

$$T^\star(\theta) = \min_{w\in\Sigma_A} \max_{a\neq a^\star(\theta)} \frac{2\|a^\star(\theta)-a\|^2_{V_w^{-1}}}{\big\langle\theta, a^\star(\theta)-a\big\rangle^2}\,.$$

Since the characteristic time involves many problem dependent quantities that are unknown to the agent, previous papers target loose problem-independent upper bounds on the characteristic time. Soare et al. (2014) (see also Tao et al. 2018, Fiez et al. 2019) introduce the G-complexity (denoted by $\mathcal{AA}$) which coincides with the G-optimal design of experimental design theory (see Pukelsheim 2006) and the $\mathcal{AB}_{\texttt{dir}}$-complexity[1] (denoted by $\mathcal{AB}_{\texttt{dir}}$) inspired by the transductive experimental design theory (Yu et al., 2006),

$$\begin{aligned} \mathcal{AA} &= \min_{w\in\Sigma_A} \max_{a\in\mathcal{A}} \|a\|^2_{V_w^{-1}}\,,\\ \mathcal{AB}_{\texttt{dir}} &= \min_{w\in\Sigma_A} \max_{b\in\mathcal{B}_{\texttt{dir}}} \|b\|^2_{V_w^{-1}}\,, \end{aligned}$$

where $\mathcal{B}_{\texttt{dir}} := \{a-a' : (a,a')\in\mathcal{A}\times\mathcal{A}\}$. For the G-optimal complexity we seek for a proportion of pulls w that explores *uniformly* the means of the arms, since the statistical uncertainty for estimating $\langle\theta,a\rangle$ scales roughly with $\|a\|_{V_w^{-1}}$. In the $\mathcal{AB}$-complexity we try to estimate *uniformly* all the *directions* $a-a'$. On the contrary in this paper we try to maximize directly the characteristic times, that is try to estimate all the *directions* $a^\star(\theta)-a$ scaled by the squared gaps $\langle\theta,a^\star(\theta)-a\rangle$. Note that the characteristic time can also be seen as a particular optimal transductive design. Indeed for

[1] This complexity is denoted as $\mathcal{XY}$ by Soare et al. (2014).

$\mathcal{B}^\star := \big\{(a^\star(\theta)-a)/\left|\langle\theta, a^\star(\theta)-a\rangle\right| : a \in \mathcal{A}/\{a^\star(\theta)\}\big\}$, it holds

$$T^\star(\theta) = 2\mathcal{A}\mathcal{B}^\star(\theta) := 2\min_{w\in\Sigma_\mathcal{A}} \max_{b\in\mathcal{B}^\star(\theta)} \|b\|^2_{V_w^{-1}}\,.$$

We have the following ordering on the complexities

$$T^\star(\theta) \leq 2\frac{\mathcal{A}\mathcal{B}_{\texttt{dir}}}{\Delta_{\min}(\theta)^2} \leq 8\frac{\mathcal{A}\mathcal{A}}{\Delta_{\min}(\theta)^2} = \frac{8d}{\Delta_{\min}(\theta)^2}\,, \quad (2)$$

where $\Delta_{\min} = \min_{a\neq a^\star(\theta)}\langle\theta, a^\star(\theta)-a\rangle$ and the last equality follows from the Kiefer-Wolfowitz equivalence theorem (Kiefer & Wolfowitz, 1959). Conversely the $\mathcal{A}\mathcal{A}$-complexity and the $\mathcal{A}\mathcal{B}_{\texttt{dir}}$-complexity are linked to an *other* pure exploration problem, the thresholding bandits (see Appendix D.2).

Remark 1. *In order to compute all these complexities, it is sufficient to solve the following generic optimal transductive design problem: for $\mathcal{B}$ a finite set of elements in $\mathbb{R}^d$,*

$$\mathcal{A}\mathcal{B} = \min_{w\in\Sigma_K} \max_{b\in\mathcal{B}} \|b\|^2_{V_w^{-1}}\,.$$

When $\mathcal{B}=\mathcal{A}$ we can use an algorithm inspired by Frank-Wolfe (Frank & Wolfe, 1956) which possesses convergence guarantees (Atwood, 1969; Ahipasaoglu et al., 2008). But in the general case, up to our knowledge, there is no algorithm with the same kind of guarantees. Previous works used an heuristic based on a straightforward adaptation of the aforementioned algorithm for general sets $\mathcal{B}$ but it seems to not converge on particular instances, see Appendix I. We instead propose in the same appendix an algorithm based on Saddle point Frank-Wolfe algorithm that seems to converge on the different instances we tested.

3. Algorithm

We present two asymptotically optimal algorithms for the general pure-exploration problem. We also make the additional assumption that the set of parameter is bounded, that is we know $M > 0$ such that for all $\theta \in \mathcal{M}$, $\|\theta\| \leq M$. This assumption is shared by most of the works on linear bandits (e.g. Abbasi-Yadkori et al. 2011; Soare et al. 2014).

We describe primarily `LinGame-C`, detailed in Algorithm 2. The principle behind `LinGame`, detailed in Algorithm 1, is similar and significant differences will be highlighted.

3.1. Notations

Counts. At each round t the algorithms will play an arm a_t and choose (fictitiously) an answer i_t. We denote by $N_t^{a,i} := \sum_{s=1}^t \mathbb{1}_{\{(a_t,i_t)=(a,i)\}}$ the number of times the pair (a,i) is chosen up to and including time t, and by $N_t^a = \sum_{i\in\mathcal{I}} N_s^{a,i}$ and $N_t^i = \sum_{a\in\mathcal{I}} N_s^{a,i}$ the partial sums. The vectors of counts at time t is denoted by $N_t := (N_t^a)_{a\in\mathcal{A}}$ and when it is clear from the context we will also denote by $N_t^a = (N_t^{a,i})_{i\in\mathcal{I}}$ and $N_t^i = (N_t^{a,i})_{i\in\mathcal{I}}$ the vectors of partial counts.

Algorithm 1 `LinGame`

Input: Agent learners for each answers $(\mathcal{L}_w^i)_{i\in\mathcal{I}}$, threshold $\beta(\cdot,\delta)$
for t = 1 … **do**
  // Stopping rule
  if $\max_{i\in\mathcal{I}} \inf_{\lambda\in\neg i} \frac{1}{2}\|\hat\theta_{t-1}-\lambda\|^2_{V_{N_{t-1}}} \geq \beta(t-1,\delta)$ **then**
    stop and **return** $\hat\imath = i^\star(\hat\theta_{t-1})$
  end if
  // Best answer
  $i_t = i^\star(\hat\theta_{t-1})$
  // Agent plays first
  Get w_t from $\mathcal{L}_w^{i_t}$ and update $W_t = W_{t-1} + w_t$
  // Best response for the nature
  $\lambda_t \in \operatorname{argmin}_{\lambda\in\neg i_t} \|\hat\theta_{t-1}-\lambda\|^2_{V_{w_t}}$
  // Feed optimistic gains
  Feed learner $\mathcal{L}_w^{i_t}$ with $g_t(w) = \sum_{a\in\mathcal{A}} w^a U_t^a/2$
  // Track the weights
  Pull $a_t \in \operatorname{argmin}_{a\in\mathcal{A}} N_{t-1}^a - W_t^a$
end for

Regularized least square estimator. We fix a regularization parameter $\eta > 0$. The regularized least square estimator for the parameter $\theta \in \mathcal{M}$ at time t is

$$\hat\theta_t = (V_{N_t} + \eta I_d)^{-1} \sum_{s=1}^t Y_s a_s\,,$$

where I_d is the identity matrix. By convention $\hat\theta_0 = 0$.

3.2. Algorithms

Stopping rule. Our algorithms share the same stopping rule. Following Garivier & Kaufmann (2016), our algorithms stop if a generalized likelihood ratio exceeds a threshold. It stops if

$$\max_{i\in\mathcal{I}} \inf_{\lambda_i\in\neg i} \frac{1}{2}\|\hat\theta_t - \lambda_i\|^2_{V_{N_t}} > \beta(t,\delta)\,, \quad (3)$$

and return $i_t^\star \in \operatorname{argmax}_{i\in\mathcal{I}} \inf_{\lambda_i\in\neg i} \|\hat\theta_t-\lambda_i\|^2_{V_{N_t}}/2$. This stopping and decision rules ensures that the algorithms `LinGame` and `LinGame-C` are δ-correct regardless of the sampling rule used, see lemma below[2] proved in Appendix F.

Lemma 2. *Regardless of the sampling rule, the stopping rule (3) with the threshold*

$$\beta(t,\delta) = \left(\sqrt{\log\left(\frac{1}{\delta}\right) + \frac{d}{2}\log\left(1+\frac{tL^2}{\eta d}\right)} + \sqrt{\frac{\eta}{2}}M\right)^2, \quad (4)$$

[2] The fact that $\tau_\delta < +\infty$ is a consequence of our analysis, see Appendix E.

Algorithm 2 LinGame-C

Input: Agent learner $\mathcal{L}_{\widetilde{w}}$, threshold $\beta(\cdot,\delta)$
for t = 1 ... **do**
  // Stopping rule
  if $\max_{i\in\mathcal{I}} \inf_{\lambda\in\neg i} \frac{1}{2}\|\hat{\theta}_{t-1}-\lambda\|^2_{V_{N_{t-1}}} \geq \beta(t-1,\delta)$ **then**
    stop and **return** $\hat{\imath} = i^\star(\hat{\theta}_{t-1})$.
  end if
  // Agent plays first
  Get $\widetilde{w}_t$ from $\mathcal{L}_{\widetilde{w}}$ and update $\widetilde{W}_t = \widetilde{W}_{t-1} + \widetilde{w}_t$
  // Best response for the nature
  For all $i \in \mathcal{I}$, $\widetilde{\lambda}^i_t \in \operatorname{argmin}_{\lambda\in\neg i}\|\hat{\theta}_{t-1} - \lambda\|^2_{V_{\widetilde{w}^i_t}}$
  // Feed optimistic gains
  Feed learner $\mathcal{L}_{\widetilde{w}}$ with $g_t(\widetilde{w}) = \sum_{(a,i)\in\mathcal{A}\times\mathcal{I}} \widetilde{w}^{a,i} U^{a,i}_t/2$
  // Track the weights
  Pull $(a_t, i_t) \in \operatorname{argmin}_{(a,i)\in\mathcal{A}\times\mathcal{I}} N^{a,i}_{t-1} - \widetilde{W}^{a,i}_t$
end for

satisfy $\mathbb{P}_\theta\big(\tau_\delta < \infty \wedge i^\star_{\tau_\delta} \neq i^\star(\theta)\big) \leq \delta$.

Our contribution is a sampling rule that minimizes the sample complexity when combined with these stopping and decision rules. We now explain our sampling strategy to ensure that the stopping threshold is reached as soon as possible.

Saddle point computation. Suppose in this paragraph, for simplicity, that the parameter vector θ is known. By the definition of the stopping rule and the generalized likelihood ratio, as long as the algorithm does not stop,

$$\beta(t,\delta) \geq \inf_{\lambda\in\neg i^\star(\theta)} \sum_{a\in\mathcal{A}} N^a_t \|\theta-\lambda\|^2_{aa^\intercal}/2\,.$$

If we manage to have $N_t \approx t w^\star(\theta)$ (the optimal pulling proportions at θ), then this leads to $\beta(t,\delta) \geq t T^\star(\theta)^{-1}$ and, solving that equation, we have asymptotic optimality.

Since there is only one correct answer, the parameter θ belongs to all sets $\neg i$ for $i \neq i^\star(\theta)$. Hence

$$\begin{aligned}&\inf_{\lambda\in\neg i^\star(\theta)} \frac{1}{2}\sum_{a\in\mathcal{A}} N^a_t \|\theta-\lambda\|^2_{aa^\intercal}\\ &\geq \inf_{\widetilde{\lambda}_t\in\prod_i(\neg i)} \frac{1}{2}\sum_{(a,i)\in\mathcal{A}\times\mathcal{I}} N^{a,i}_t \|\theta-\widetilde{\lambda}^i\|^2_{aa^\intercal}.\end{aligned}$$

Introducing the sum removes the dependence in the unknown $i^*(\theta)$. LinGame-C then uses an agent playing weights w in $\Sigma_{\mathcal{A}\mathcal{I}}$. LinGame does not use that sum over answers, but uses a guess for $i^\star(\theta)$. Its analysis involves proving that the guess is wrong only finitely many times in expectation.

Our sampling rule implements the lower bound game between an agent, playing at each stage s a weight vector $\widetilde{w}_s$ in the probability simplex $\Sigma_{A\times I}$, and nature, who computes at each stage a point $\lambda^i_s \in \neg i$ for all $i \in \mathcal{I}$. We additionally ensure that $N^{a,i}_t \approx \sum_{s=1}^t \widetilde{w}^{a,i}_s$. Suppose that the sampling rule is such that at stage t, a ε_t-approximate saddle point is reached for the lower bound game, see Lemma 1. That is,

$$\begin{aligned}&\inf_{\widetilde{\lambda}\in\prod_i(\neg i)} \sum_{s=1}^t \sum_{(a,i)\in\mathcal{A}\times\mathcal{I}} \widetilde{w}^{i,a}_s \|\theta-\widetilde{\lambda}^i\|^2_{aa^\intercal}/2 + \varepsilon_t\\ &\geq \sum_{s=1}^t \sum_{(a,i)\in\mathcal{A}\times\mathcal{I}} \widetilde{w}^{i,a}_s \|\theta-\widetilde{\lambda}^i_s\|^2_{aa^\intercal}/2\\ &\geq \max_{(a,i)\in\mathcal{A}\times\mathcal{I}} \sum_{s=1}^t \|\theta-\widetilde{\lambda}^i_s\|^2_{aa^\intercal}/2 - \varepsilon_t\,.\end{aligned}$$

Then if the algorithm did not stop, it verifies, using Lemma 1,

$$\begin{aligned}\beta(t,\delta) &\geq t \max_{(a,i)\in\mathcal{A}\times\mathcal{I}} \frac{1}{t}\sum_{s=1}^t \|\theta-\widetilde{\lambda}^i_s\|^2_{aa^\intercal}/2 - 2\varepsilon_t\\ &\geq t \inf_{\widetilde{q}\in\prod_{i\in\mathcal{I}}\mathcal{P}(\neg i)} \max_{(a,i)\in\mathcal{A}\times\mathcal{I}} \mathbb{E}_{\lambda^i\sim q^i}\|\theta-\widetilde{\lambda}^i\|^2_{aa^\intercal}/2 - 2\varepsilon_t\\ &= tT^\star(\theta)^{-1} - 2\varepsilon_t\,.\end{aligned}$$

Solving that equation, we get asymptotically the wanted $t \lesssim T^\star(\theta)\log(1/\delta)$.

We implement the saddle point algorithm by using AdaHedge for the agent (a regret minimizing algorithm of the exponential weights family), and using best-response for the nature, which plays after the agent. Precisely the learner $\mathcal{L}_w$ for LinGame-C is AdaHedge on Σ_{AI} with the gains

$$g^\theta_t(\widetilde{w}) = \frac{1}{2}\sum_{(a,i)\in\mathcal{A}\times\mathcal{I}} \widetilde{w}^{a,i}\|\theta-\widetilde{\lambda}^i_s\|^2_{aa^\intercal}\,.$$

Whereas LinGame uses I learners $\mathcal{L}^i_w$, one for each possible guess of $i^\star(\theta)$ with the gains. For $i \in \mathcal{I}$, the learner $\mathcal{L}^i_w$ is also AdaHedge but only on Σ_A with the gains (when the guess is i)

$$g^\theta_t(w) = \frac{1}{2}\sum_{a\in\mathcal{A}} w^a\|\theta-\lambda^i_s\|^2_{aa^\intercal}\,.$$

ε_t is then the sum of the regrets of the two players. Best-response has regret 0, while the regret of AdaHedge is $O(\sqrt{t})$ for bounded gains, as seen in the following lemma, taken from de Rooij et al. (2014).

Lemma 3. *On the online learning problem with K arms and gains $g_s(w) = \sum_{k\in[K]} w^k U^k_s$ for $s \in [t]$, AdaHedge,*

predicting $(w_s)_{s\in[t]}$*, has regret*

$$R_t := \max_{w\in\Sigma_K} \sum_{s=1}^{t} g_s(w) - g_s(w_s)$$
$$\leq 2\sigma\sqrt{t\log(K)} + 16\sigma(2+\log(K)/3)\,,$$
$$\text{where } \sigma = \max_{s\leq t}(\max_{k\in[K]} U_s^k - \min_{k\in[K]} U_s^k)\,.$$

Other combinations of algorithms are possible, as long as the sum of their regrets is sufficiently small. At each stage $t \in \mathbb{N}$, both algorithms advance only by one iteration and as time progresses, the quality of the saddle point approximation improves. This is in contrast with `Track-and-Stop` (Garivier & Kaufmann, 2016), in which an exact saddle point is computed at each stage, at a potentially much greater computational cost.

Optimism. The above saddle point argument would be correct for a known game, while our algorithm is confronted to a game depending on the unknown parameter θ. Following a long tradition of stochastic bandit algorithms, we use the principle of Optimism in Face of Uncertainty. Given an estimate $\hat{\theta}_{t-1}$, we compute upper bounds for the gain of the agent at θ, and feed these optimistic gains to the learner. Precisely, given the best response $\lambda_t^i \in \neg i$ we define,

$$U_t^{a,i} = \begin{cases} \max_\xi & \min\left(\|\xi-\lambda_t^i\|^2_{aa^\intercal}, 4L^2M^2\right) \\ \text{s.t.} & \|\hat{\theta}_{t-1}-\xi\|^2_{V_{N_{t-1}}+\eta I_d} \leq 2h(t) \end{cases},$$

where $h(t) = \beta(t, 1/t^3)$ is some exploration function. We clipped the values, using that $\mathcal{M}$ and $\mathcal{A}$ are bounded to ensure bounded gains for the learners. Under the event that the true parameter verifies $\|\hat{\theta}_{t-1} - \theta\|^2 \leq 2h(t)$, this is indeed an optimistic estimate of $\|\theta - \lambda_t^i\|^2_{aa^\intercal}$. Note that $U_t^{a,i}$ has a closed form expression, see Appendix E. The optimistic gain is then, for `LinGame-C` (see Algorithm 1 for the one of `LinGame`),

$$g_t(\widetilde{w}) = \frac{1}{2}\sum_{(a,i)\in\mathcal{A}\times\mathcal{I}} \widetilde{w}^{a,i} U_t^{a,i}\,.$$

Tracking. In both Algorithm 1 and 2, the agent plays weight vectors in a simplex. Since the bandit procedure allows only to pull one arm at each stage, our algorithm needs a procedure to transcribe weights into pulls. This is what we call tracking, following Garivier & Kaufmann (2016). The choice of arm (or arm and answer) is

$$a_{t+1} \in \operatorname*{argmin}_{a\in\mathcal{A}} N_t^a - W_{t+1}^a \qquad \text{for Algorithm 1,}$$
$$(a_{t+1}, i_{t+1}) \in \operatorname*{argmin}_{(a,i)\in\mathcal{A}\times\mathcal{I}} N_t^{a,i} - \widetilde{W}_{t+1}^{a,i} \qquad \text{for Algorithm 2.}$$

This procedure guarantees that for all $t \in \mathbb{N}, u \in \mathcal{U}$, with $\mathcal{U} = \mathcal{A}$ (resp. $\mathcal{U} = \mathcal{I}\times\mathcal{A}$) for Algo. 1 (resp. Algo. 2), $-\log(|\mathcal{U}|) \leq N_t^u - W_t^u \leq 1$. That result is due to Degenne et al. (2020) and its proof is reproduced in Appendix G.

Theorem 2. *For a regularization parameter*[3] $\eta \geq 2(1+\log(A))AL^2 + M^2$*, for the threshold* $\beta(t,\delta)$ *given by* (4)*, for an exploration function* $h(t) = \beta(t, 1/t^3)$*,* `LinGame` *and* `LinGame-C` *are* δ*-correct and asymptotically optimal. That is, they verify for all* $\theta \in \mathcal{M}$*,*

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_\theta[\tau_\delta]}{\log 1/\delta} \leq T^\star(\theta)\,.$$

The main ideas used in the proof are explained above. The full proof is in appendix E with finite δ upper bounds.

3.3. Bounded parameters

We provide a bounded version of different examples (e.g. BAI) in Appendix D where we add the assumption that the parameter set $\mathcal{M}$ is bounded. In particular we show how it affects the lower bound of Theorem 1: the characteristic time $T^\star(\theta)$ is reduced (or equivalently $T^\star(\theta)^{-1}$ increases). This is not surprising since we add a new constraint in the optimization problem. This means that the algorithm should stop earlier. The counterpart of this improvement is that it is often difficult to compute the best response for nature. Indeed, for example, in BAI, there is an explicit expression of the best response, see Appendix D.1. When the constraint $\|\lambda\| \leq M$ is added there is no explicit expression anymore and one needs to solve an uni-dimensional optimization problem, see Lemma 5. To devise an asymptotically optimal algorithm without the boundedness assumption remains an open problem.

Note that in the proof of Theorem 2 we only use two times the boundedness assumption, first in the definition of the threshold $\beta(t,\delta)$ (see Theorem 3) to handle the bias induced by the regularization. Second, since the regret of AdaHedge is proportional to the maximum of the upper confidence bounds $U_s^{i,a}$, we need to ensure that they are bounded.

4. Related Work

We survey previous work on linear BAI. The major focus is put on sampling rules in this section. We stress that all the stopping rules employed in the linear BAI literature are equivalent up to the choice of their exploration rate (More discussion given in Appendix H). As aforementioned, existing sampling rules are either based on `SuccessiveElimination` or `UGapE`. Elimination-based sampling rules usually operate in phases and progressively

[3] This condition is a simple technical trick to simplify the analysis. An η independent of A,L,M will lead to the same results up to minor adaptations of the proof.

discard sub-optimal directions. Gap-based sampling rules always play the most informative arm that reduces the uncertainty of the gaps between the empirical best arm and the others.

$\mathcal{XY}$-Static and $\mathcal{XY}$-Adaptive. Soare et al. (2014) first propose a static allocation design $\mathcal{XY}$-Static that aims at reducing the uncertainty of the gaps of all arms. More precisely, it requires to either solve the $\mathcal{AB}_{\texttt{dir}}$-complexity or use a *greedy* version that pulls the arm $\operatorname{argmin}_{a\in\mathcal{A}} \max_{b\in\mathcal{B}_{\texttt{dir}}} \|b\|^2_{V_w^{-1}}$ at the cost of having no guarantees. An elimination-like alternative called $\mathcal{XY}$-Adaptive is proposed then to overcome that issue. We say elimination-like since $\mathcal{XY}$-Adaptive does not discard arms once and for all, but reset the active arm set at each phase. $\mathcal{XY}$-Adaptive and $\mathcal{XY}$-Static are the first algorithms being linked to $\mathcal{AA}$-optimality, but are not asymptotically optimal.

ALBA. ALBA is also an eliminations-based algorithm designed by Tao et al. (2018) that improves over $\mathcal{XY}$-Adaptive by a factor of d in the sample complexity using a tighter elimination criterion.

RAGE. Fiez et al. (2019) extend $\mathcal{XY}$-Static and $\mathcal{XY}$-Adaptive to a more general transductive bandits setting. RAGE is also elimination-based and requires the computation of $\mathcal{AB}_{\texttt{dir}}$-complexity at each phase.

LinGapE and variants. LinGapE (Xu et al., 2018) is the first gap-based sampling rule for linear BAI. LinGapE is inspired by UGapE (Gabillon et al., 2012). It is, however, not clear whether LinGapE is asymptotically optimal or not. Similar to $\mathcal{XY}$-Static, LinGapE either requires to solve a time-consuming optimization problem at each step, or can use a greedy version that pulls arm $\operatorname{argmin}_{a\in\mathcal{A}} \|a_{i_t} - a_{j_t}\|^2_{(V_w + aa^\intercal)^{-1}}$ instead, again at the cost of losing guarantees. Here $i_t = i^\star(\hat{\theta}_t)$ and a_{j_t} is the most ambiguous arm w.r.t. a_{i_t}, i.e. $\operatorname{argmax}_{j\neq i_t} \langle \hat{\theta}_t, a_j - a^\star(\hat{\theta}_t)\rangle + \|a^\star(\hat{\theta}_t) - a_{j_t}\|_{V_{N_t}^{-1}} \sqrt{2\beta(t,\delta)}$. On the other hand, Zaki et al. (2019) propose a new algorithm based on LUCB. With a careful examination, we note that the sampling rule of GLUCB is equivalent to that of the greedy LinGapE using the Sherman-Morrison formula. Later, Kazerouni & Wein (2019) provide a natural extension of LinGapE to the *generalized linear bandits* setting, where the rewards depend on a strictly increasing *inverse link function*. GLGapE reduces to LinGapE when the inverse link function is the identity function.

Note that all the sampling rules presented here depend on δ (except $\mathcal{XY}$-Static), while our sampling rules have a δ-free property which is appealing for applications as argued by Jun & Nowak (2016). Also all the guarantees in the literature are of the form $C\log(\delta) + O\big(\log(1/\delta)\big)$ for a constant C that is strictly larger than $T^\star(\theta)^{-1}$.

5. Experiments

Besides our algorithms, we implement the following algorithms, all using the same stopping rule (more discussion given in Appendix H): uniform sampling, the greedy version of $\mathcal{XY}$-Static (including $\mathcal{AA}$-allocation and $\mathcal{AB}_{\texttt{dir}}$-allocation), $\mathcal{XY}$-Adaptive, and the greedy version of LinGapE. We skip GLUCB/GLGapE since they are more or less equivalent to LinGapE in the scope of this paper.

The usual hard instance. Usual sampling rules for classical BAI may not work for the linear case. This can be understood on a well-studied instance already discussed by Soare et al. (2014); Xu et al. (2018), which encapsulates the difficulty of BAI in a linear bandit, and thus is the first instance on which we test our algorithms. In this instance, arms are the canonical basis $a_1 = e_1, a_2 = e_2, a_d = e_d$, plus an additional disturbing arm $a_{d+1} = (\cos(\alpha), \sin(\alpha), 0, \dots, 0)^\intercal$, and a true regression parameter θ equal to e_1. In this problem, the best arm is always a_1, but when the angle α is small, the disturbing arm a_{d+1} is hard to discriminate from a_1. As already argued by Soare et al. (2014), an efficient sampling rule for this problem instance would rather pull a_2 in order to reduce the uncertainty in the direction $a_1 - a_{d+1}$. Naive adaptation of classical BAI algorithms cannot deal with that situation naturally. We further use a simple set of experiments to justify that intuition. We run our two algorithms and the one of Degenne et al. (2019) that we call DKM over the problem instance whence $d = 2$, $\delta = 0.01$ and $\alpha = 0.1$. We show the number of pulls for each arm averaged over 100 replications of experiments in Table 1. We see that, indeed, DKM pulls too much a_3, while our algorithms focus mostly on a_2.

	LinGame	LinGame-C	DKM
a_1	1912	1959	1943
a_2	5119	4818	4987
a_3	104	77	1775
Total	7135	**6854**	8705

Table 1. Average number of pulls of each arm.

Comparison of different complexities. We use the previous setting to illustrate various complexities differ in practice. In Table 2 we compare the different complexities mentioned in this paper: the characteristic time $T^\star(\theta)$ and its associated optimal weights $w^\star_{\mathcal{AB}^\star(\theta)}$, the $\mathcal{AB}_{\texttt{dir}}$-complexity and its associated optimal design $w^\star_{\mathcal{AB}_{\texttt{dir}}}$, the G-optimal complexity $\mathcal{AA}$ and its associated optimal design $w^\star_{\mathcal{AA}}$. For each weight vector $w \in \{w^\star_{\mathcal{AB}^\star(\theta)}, w_{\mathcal{AB}_{\texttt{dir}}}, w_{\mathcal{AA}}\}$, we also provide the lower bound T_w given by Therorem 1, i.e.

$$T_w = \max_{a\neq a^\star(\theta)} \frac{\big\langle \theta, a^\star(\theta) - a\big\rangle^2}{2\|a^\star(\theta) - a\|^2_{V_w^{-1}}} \log(1/\delta).$$

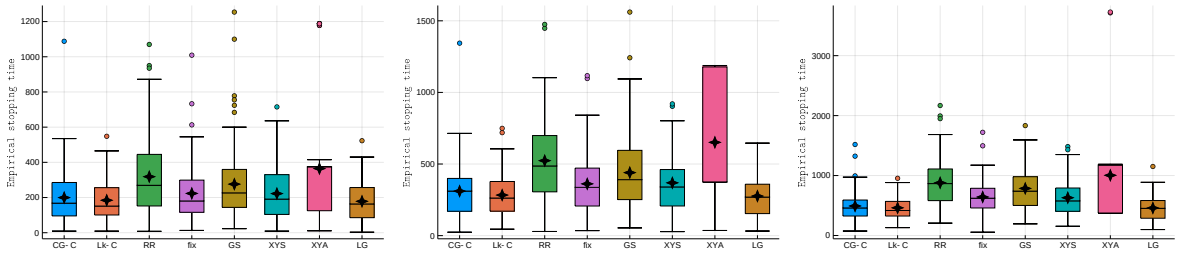

Figure 1. Sample complexity over the usual counter-example with $\delta = 0.1, 0.01, 0.0001$ respectively. CG = `LinGame-C`, Lk = `LinGame`, RR = uniform sampling, fix = tracking the fixed weights, GS = $\mathcal{XY}$-`Static` with $\mathcal{AA}$-allocation, XYS = $\mathcal{XY}$-`Static` with $\mathcal{AB}_{\texttt{dir}}$-allocation, LG = `LinGapE`. The mean stopping time is represented by a black cross.

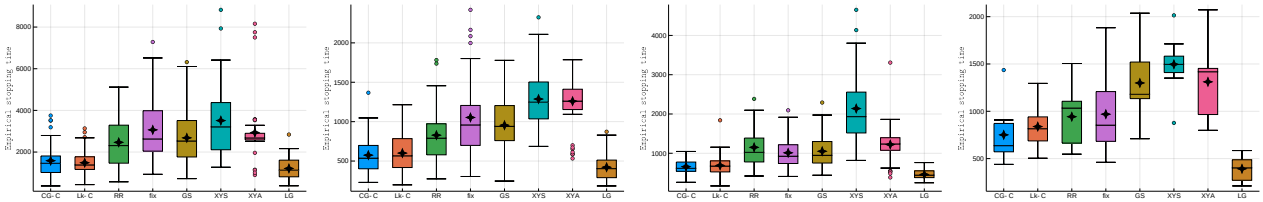

Figure 2. Sample complexity over random unit sphere vectors with $d = 6, 8, 10, 12$ from left to right.

In particular we notice that targeting the proportions of pulls $w_{\mathcal{AB}_{\texttt{dir}}}, w_{\mathcal{AA}}$ leads to a much larger lower bound than the one obtained with the optimal weights.

	$w^\star_{\mathcal{AB}^\star}$	$w^\star_{\mathcal{AB}_{\texttt{dir}}}$	$w^\star_{\mathcal{AA}}$
a_1	0.047599	0.499983	0.499983
a_2	0.952354	0.499983	0.499983
a_3	0.000047	0.000033	0.000033
T_w	369	2882	2882
	$T^\star(\theta)$	$2\mathcal{AB}_{\texttt{dir}}/\Delta^2_{\min}$	$8\mathcal{AA}/\Delta^2_{\min}$
Complexity	0.124607	32.0469	64.0939

Table 2. Optimal w for various complexities ($\Delta_{\min} = 0.0049958$).

Comparison with other algorithms. Finally we benchmark our two sampling rules against others from the literature. We test over two synthetic problem instances, with the first being the previous counter-example. We set $d = 2$, $\alpha = \pi/6$. Fig. 1 shows the empirical stopping time of each algorithms averaged over 100 runs, with a confidence level $\delta = 0.1, 0.01, 0.0001$ from left to right. Our two algorithms show competitive performance (the two leftmost boxes on each plot), and are only slightly worse than `LinGapE`.

For the second instance, we consider 20 arms randomly generated from the unit sphere $\mathbb{S}^{d-1} := \{a \in \mathbb{R}^d; \|a\|_2 = 1\}$. We choose the two closest arms a, a' and we set $\theta = a + 0.01(a' - a)$ so that a is the best arm. This setting has already been considered by Tao et al. (2018). We report the same box plots over 100 replications as before with increasing dimension in Fig. 2. More precisely, we set $d = 6, 8, 10, 12$ respectively, and always keep a same $\delta = 0.01$. Our algorithms consistently show strong performances compared to other algorithms apart from `LinGapE`. Moreover, we can see that in these random examples, `LinGame-C` works better than the non-confexified one, and is even competitive compared to `LinGapE`.

We stress that although the main focus of this paper is theoretical, with algorithms that are asymptotically optimal, our methods are also competitive with earlier algorithms experimentally.

6. Conclusion

In this paper, we designed the first practically usable asymptotically optimal sampling rules for the pure exploration game for finite-arm linear bandits. Should the boundedness assumption be necessary to have optimal algorithm remains an open question.

Another concern about the current sampling rules could be their computational complexity. In BAI, the one step complexity of `LinGame-C` (or `LinGame`) is dominated by the computation of the best response for nature, which requires a full matrix inversion. Alternatives that involve rank-1 updates should be considered.

More generally, however, the part of fixed-confidence pure exploration algorithms that needs an improvement the most is the stopping rule. While the one we used guarantees δ-correctness, it is very conservative. Indeed, the experimental error rates of algorithms using that stopping rule are orders of magnitude below δ. This means that the concentration

inequality does not reflect the query we seek to answer. It quantifies deviations of the d-dimensional estimate in all directions (morally, along 2^d directions). However, for the usual BAI setting with d arms in an orthogonal basis, it would be sufficient to control the deviation of that estimator in $d-1$ directions to make sure that $i^*(\theta) = i^*(\hat{\theta}_t)$.

Finally, the good performance of `LinGapE` raises the natural question of whether it could be proven to have similar asymptotic optimality.

Acknowledgements

The research presented was supported by European CHIST-ERA project DELTA, French Ministry of Higher Education and Research, Nord-Pas-de-Calais Regional Council, and by French National Research Agency as part of the "Investissements d'avenir" program, reference ANR19-P3IA-0001 (PRAIRIE 3IA Institute) and the project BOLD, reference ANR19-CE23-0026-04.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24 (NIPS)*, 2011. ISBN 9781618395993. URL https://sites.ualberta.ca/{~}szepesva/papers/linear-bandits-NIPS2011.pdf.

Ahipasaoglu, S. D., Sun, P., and Todd, M. J. Linear convergence of a modified Frank-Wolfe algorithm for computing minimum-volume enclosing ellipsoids. *Optimization Methods and Software*, 23(1):5–19, 2008. ISSN 10556788. doi: 10.1080/10556780701589669. URL https://www.tandfonline.com/doi/full/10.1080/10556780701589669.

Atwood, C. L. Optimal and efficient designs of experiments. *The Annals of Mathematical Statistics*, 40(5):1570–1602, 1969. ISSN 0003-4851. doi: 10.1214/aoms/1177697374.

Audibert, J.-Y. and Bubeck, S. Best arm identification in multi-armed bandits. In *Proceedings of the 23rd Annual Conference on Learning Theory (CoLT)*, 2010. ISBN 9780982252925. URL https://hal-enpc.archives-ouvertes.fr/hal-00654404/document.

Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002. URL https://www.jmlr.org/papers/volume3/auer02a/auer02a.pdf.

Bubeck, S., Munos, R., and Stoltz, G. Pure exploration in multi-armed bandits problems. In *Proceedings of the 20th International Conference on Algorithmic Learning Theory (ALT)*, pp. 23–37, 2009. ISBN 3642044131. doi: 10.1007/978-3-642-04414-4_7. URL https://arxiv.org/pdf/0802.2655.pdf.

Chen, S., Lin, T., King, I., Lyu, M. R., and Chen, W. Combinatorial pure exploration of multi-armed bandits. In *Advances in Neural Information Processing Systems 27 (NIPS)*, pp. 379–387, 2014. URL https://papers.nips.cc/paper/5433-combinatorial-pure-exploration-of-multi-armed-bandits.pdf.

Chernoff, H. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3):755–770, 1959. ISSN 0003-4851. doi: 10.1214/aoms/1177706205. URL https://projecteuclid.org/download/pdf{_}1/euclid.aoms/1177706205.

de Rooij, S., Van Erven, T., Grünwald, P. D., and Koolen, W. M. Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15:1281–1316, 2014. ISSN 15337928. URL https://arxiv.org/pdf/1301.0534.pdf.

Degenne, R. and Koolen, W. M. Pure exploration with multiple correct answers. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019. URL https://arxiv.org/pdf/1902.03475.pdf.

Degenne, R., Koolen, W., and Ménard, P. Non-asymptotic pure exploration by solving games. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019. URL http://arxiv.org/pdf/1906.10431.pdf.

Degenne, R., Shao, H., and Koolen, W. M. Structure Adaptive Algorithms for Stochastic Bandits. *International Conference on Machine Learning*, 2020.

Even-Dar, E., Mannor, S., and Mansour, Y. PAC bounds for Multi-armed Bandit and Markov Decision Processes. In *Proceedings of the15th Annual Conference on Learning Theory (CoLT)*, volume 2375, pp. 255–270, 2002. ISBN 354043836X. doi: 10.1007/3-540-45435-7_18.

Even-dar, E., Mannor, S., and Mansour, Y. Action elimination and stopping conditions for reinforcement learning. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pp. 162–169, 2003. ISBN 1577351894. URL https://www.aaai.org/Papers/ICML/2003/ICML03-024.pdf.

Fiez, T., Jain, L., Jamieson, K., and Ratliff, L. Sequential experimental design for transductive linear bandits. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019. URL http://arxiv.org/pdf/1906.08399.pdf.

Frank, M. and Wolfe, P. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2): 95–110, 1956. URL https://onlinelibrary.wiley.com/doi/epdf/10.1002/nav.3800030109.

Gabillon, V., Ghavamzadeh, M., and Lazaric, A. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems 25 (NIPS)*, pp. 3212–3220, 2012. ISBN 9781627480031. URL http://papers.nips.cc/paper/4640-best-arm-identification-a-unified-approach-to-fixed-budget-and-fixed-confidence.pdf.

Garivier, A. and Kaufmann, E. Optimal best arm identification with fixed confidence. In *Proceedings of the 29th Annual Conference on Learning Theory (CoLT)*, 2016. URL http://arxiv.org/pdf/1602.04589.pdf.

Garivier, A., Ménard, P., and Stoltz, G. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2018. ISSN 15265471. doi: 10.1287/moor.2017.0928. URL https://pubsonline.informs.org/doi/pdf/10.1287/moor.2017.0928.

Hoffman, M. W., Shahriari, B., and de Freitas, N. On correlation and budget constraints in model-based bandit optimization with application to automatic machine learning. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AIStats)*, pp. 365–374, 2014. URL http://proceedings.mlr.press/v33/hoffman14.pdf.

Jun, K.-S. and Nowak, R. Anytime exploration for multi-armed bandits using confidence information. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 974–982, 2016. ISBN 9781510829008. URL http://proceedings.mlr.press/v48/jun16.pdf.

Kalyanakrishnan, S. and Stone, P. Efficient selection of multiple bandit arms: Theory and practice. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pp. 511–518, 2010. ISBN 9781605589077. URL https://www.cs.utexas.edu/users/pstone/Papers/bib2html-links/ICML10-kalyanakrishnan.pdf.

Karnin, Z., Koren, T., and Somekh, O. Almost optimal exploration in multi-armed bandits. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pp. 1238–1246, 2013. URL http://proceedings.mlr.press/v28/karnin13.pdf.

Kaufmann, E., Koolen, W. M., and Garivier, A. Sequential test for the lowest mean: From Thompson to Murphy sampling. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pp. 6332–6342, 2018. URL https://papers.nips.cc/paper/7870-sequential-test-for-the-lowest-mean-from-thompson-to-murphy-sampling.pdf.

Kazerouni, A. and Wein, L. M. Best arm identification in generalized linear bandits. *arXiv preprint arXiv:1905.08224*, 2019. URL http://arxiv.org/pdf/1905.08224.pdf.

Kiefer, J. and Wolfowitz, J. Optimum designs in regression problems. *The Annals of Mathematical Statistics*, 30(2):271–294, 1959. doi: 10.1007/978-1-4615-6660-1_2. URL https://projecteuclid.org/download/pdf{_}1/euclid.aoms/1177706252.

Lattimore, T. and Szepesvari, C. *Bandit Algorithms*. Cambridge University Press, 2018. URL http://downloads.tor-lattimore.com/book.pdf.

Locatelli, A., Gutzeit, M., and Carpentier, A. An optimal algorithm for the thresholding bandit problem. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 2539–2554, 2016. ISBN 9781510829008. URL http://proceedings.mlr.press/v48/locatelli16.pdf.

Lu, H., Freund, R. M., and Nesterov, Y. Relatively-smooth convex optimization by first-order methods , and applications. *SIAM Journal of Optimization*, 28(1):333–354, 2018. URL https://epubs.siam.org/doi/pdf/10.1137/16M1099546.

Ménard, P. Gradient ascent for active exploration in bandit problems. *arXiv preprint arXiv:1905.08165*, 2019. URL http://arxiv.org/pdf/1905.08165.pdf.

Pukelsheim, F. *Optimal Design of Experiments*. Society for Industrial and Applied Mathematics, 2006. doi: 10.1002/0471667196.ess3056.pub2. URL https://epubs.siam.org/doi/pdf/10.1137/1.9780898719109.bm.

Qin, C., Klabjan, D., and Russo, D. Improving the expected improvement algorithm. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pp. 5381–5391, 2017. URL http://arxiv.org/pdf/1705.10033.pdf.

Robbins, H. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematics Society*, 58(5):527–535, 1952. doi: 10.1090/S0002-9904-1952-09620-8. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.335.3232{&}rep=rep1{&}type=pdf.

Shang, X., de Heide, R., Kaufmann, E., Ménard, P., and Valko, M. Fixed-confidence guarantees for Bayesian best-arm identification. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AIStats)*, 2020. URL http://arxiv.org/pdf/1910.10945.pdf.

Soare, M., Lazaric, A., and Munos, R. Best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems 26 (NIPS)*, pp. 828–836, 2014. URL https://arxiv.org/pdf/1409.6110.pdf.

Tao, C., Blanco, S. A., and Zhou, Y. Best arm identification in linear bandits with linear dimension dependency. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 7773–7786, 2018. ISBN 9781510867963. URL http://proceedings.mlr.press/v80/tao18a/tao18a.pdf.

Teraoka, K., Hatano, K., and Takimoto, E. Efficient sampling method for monte carlo tree search problem. *IEICE Transactions on Information and Systems*, E97-D(3):392–398, 2014. ISSN 17451361. doi: 10.1587/transinf.E97.D.392. URL https://pdfs.semanticscholar.org/30c9/85d7eb0deb9e40fa63ec5a7df818efc85952.pdf.

Xu, L., Honda, J., and Sugiyama, M. A fully adaptive algorithm for pure exploration in linear bandits. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AIStats)*, pp. 843–851, 2018. URL http://proceedings.mlr.press/v84/xu18d/xu18d.pdf.

Yu, K., Bi, J., and Tresp, V. Active learning via transductive experimental design. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pp. 1081–1088, 2006. ISBN 1595933832. URL https://dl.acm.org/doi/pdf/10.1145/1143844.1143980?download=true.

Zaki, M., Mohan, A., and Gopalan, A. Towards optimal and efficient best arm identification in linear bandits. In *Workshop on Machine Learning at Neural Information Processing Systems (NeurIPS-CausalML)*, 2019. URL https://arxiv.org/pdf/1911.01695.pdf.