Efficient Sampling for Pure Exploration Problems

G Dhinesh Chandran
Department of Electrical Engineering
IIT Madras, Chennai, India
ee22d200@smail.iitm.ac.in

Kota Srinivas Reddy
Department of Artificial Intelligence
IIT Kharagpur Khargpur, India
ksreddy@ai.iitkgp.ac.in

Srikrishna Bhashyam
Department of Electrical Engineering
IIT Madras, Chennai, India
skrishna@ee.iitm.ac.in

***Abstract*—We consider pure exploration problems in stochastic bandits in a fixed-confidence setting. Algorithms for such problems typically consist of a sampling rule, a stopping rule, and a declaration rule, where the sampling rule ensures that the proportions of arm pulls converge to optimal allocations. However, existing approaches require solving an optimization problem at each time step, making them computationally expensive. We propose a computationally efficient sampling rule that extends a heuristic originally developed for the Bandit Online Clustering (BOC) problem to a broader class of pure exploration problems. We provide a theoretical guarantee that the proposed rule ensures convergence to optimal allocations. We obtain the computational complexity for the Best Arm Identification (BAI) problem under our approach and compare it with existing algorithms in the literature. Simulation results demonstrate that our method significantly reduces per-sample runtime compared to existing algorithms for the BAI problem in both unconstrained bandits and linear bandits frameworks, while also yielding comparable expected sample complexity.**

***Index Terms*—Pure Exploration, Efficient Sampling, Complexity Order, Convergence**

I. Introduction

The multi-armed bandit (MAB) is a sequential decision-making framework under uncertainty that involves repeatedly selecting actions (arms) to achieve a specified objective. The MAB problem was first introduced in [1]. MAB problems can be broadly categorized into regret-minimization, where the objective is to maximize cumulative reward by minimizing regret relative to the optimal arm [2], and pure-exploration problems, where the objective is to identify a solution to a specific problem of interest [3]. Pure exploration problems can be studied in the fixed budget setting, where the number of arm pulls is limited, and the goal is to minimize the probability of error [4], or in the fixed confidence setting, where there is an upper bound constraint on the probability of error and the goal is to minimize the total number of arm pulls [5]. Pure exploration problems under the fixed-confidence setting have been extensively studied in both linear bandits, a class of structured bandits, and unstructured bandits, which we refer to as standard bandits. In standard bandits, each arm generates i.i.d. samples from an unknown distribution, whereas in linear bandits, each arm is associated with a feature vector and generates a noisy realization of the dot product between the feature vector and an unknown vector. Our focus is on pure exploration problems in standard and linear bandits under a fixed confidence setting.

This work is supported in part by Qualcomm University Research Grant. Kota Srinivas Reddy was supported by the Department of Science and Technology (DST), Govt. of India, through the INSPIRE faculty fellowship.

Pure exploration problems in standard bandits include Best Arm Identification (BAI) [3], where the objective is to identify the arm with the highest mean; clustering, typically referred to as Bandit Online Clustering (BOC) [6], where the objective is to group arms into clusters based on similarity in their underlying distribution; and sequential multi-hypothesis testing [7], [8], where the objective is to identify the true hypothesis where the parameters of the arms belong. The most-studied pure exploration problem in linear bandits is BAI [9], where the objective is to identify the arm with the highest dot product between the feature vector and an unknown vector. Bandit algorithms designed to solve a pure exploration problem under fixed confidence settings, which we refer to as pure exploration algorithms, are sequential algorithms that consist of a sampling rule to adaptively select an arm at each time step based on previous observations, a stopping rule, and a declaration rule to estimate the solution. The sampling rule of the pure exploration algorithms in [3], [6]–[9] has been proven to ensure that the arm pull proportions converge to the optimal proportions. This convergence guarantee is crucial for proving that those algorithms are asymptotically order-optimal, i.e., the expected sample complexity grows at most by a constant factor times the lower bound slope. However, the sampling rule involves solving an optimization problem in each time, making those algorithms computationally expensive. Recently, computationally efficient sampling rules for pure exploration algorithms in standard bandits have been proposed for BAI [10] and BOC [11] while ensuring convergence to optimal proportions. The key idea is that, instead of solving a full optimization problem, the algorithm takes one step towards the optimum at each time step and eventually converges to it. The runtime for that one-step update is approximately the same as that of the stopping rule. A more efficient heuristic variant is proposed in [11] for the BOC problem in standard bandits settings, which uses the same one-step update idea but with quantities already computed as part of the stopping rule, further reducing runtime, but lacks a convergence guarantee.

The main contributions of this paper are the following. (1) We extend the heuristic sampling rule proposed in [11] to solve a broader class of pure exploration problems in standard and linear bandits, and refer to the resulting algorithm as the Computationally Efficient Pure Exploration (CEPE) Algorithm. (2) We prove that the sampling rule in CEPE ensures that the arm pull proportions converge to the optimal proportions under mild assumptions. (3) We obtain the computational complexity of CEPE for the Best Arm Identification (BAI) problem in both standard and linear bandit settings and

compare it with existing algorithms in the literature [3], [9], [10]. (4) Through simulations, we show the computational efficiency of CEPE by comparing its per-sample runtime with that of existing algorithms proposed for BAI in standard [3], [10] and linear bandits [9], while achieving comparable performance in terms of expected sample complexity.

II. System Model

Consider a multi-armed bandit (MAB) problem with M arms. Each arm m is associated with an n-dimensional known feature vector $\boldsymbol{\alpha}_m \in \mathbb{R}^n$. Each arm $m \in [M]$ generates a sequence of independent and identically distributed (i.i.d.) samples $\{\boldsymbol{X}_s^{(m)} : s \in \mathbb{N}\}$, where $\boldsymbol{X}_s^{(m)}$ is a sample from the arm m at time s. Each sample $\boldsymbol{X}_s^{(m)}$ is generated independently from a vector parametric distribution $P_m(\cdot \mid \boldsymbol{\theta}\boldsymbol{\alpha}_m)$ with the parameter being $\boldsymbol{\theta}\boldsymbol{\alpha}_m \in \mathbb{R}^d$, where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^{d\times n}$ is the unknown weight matrix. We refer to this weight matrix $\boldsymbol{\theta} \in \Theta$ as a problem instance. The following are types of MAB setups that fit into the above-discussed framework and have been widely studied in the literature. *(1) Standard Bandits:* The dimension of the feature vectors n is the same as the number of arms M. The feature vector of arm $m \in [M]$ takes a special form where the m^{th} coordinate is 1 and 0 elsewhere. Now, the columns of the weight matrix $\boldsymbol{\theta} \coloneqq [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M]$ directly correspond to the parameters of the arms, i.e., $\boldsymbol{\theta}_m$ is a d-dimensional parameter vector corresponding to arm m. *(2) Linear Bandits:* The dimension of the parameter is $d = 1$, so the weight matrix $\boldsymbol{\theta}$ reduces to a weight vector of dimensions n. The parameter of arm m is given by the dot product of its feature vector $\boldsymbol{\alpha}_m$ and the weight vector $\boldsymbol{\theta}$.

Let $S \in \mathcal{Z}$ be the solution to an MAB problem, where $\mathcal{Z}$ is the set of all possible solutions to that MAB problem. We use the notation $S_1 \sim S_2$ to indicate that the solutions S_1 and S_2 are equivalent. Let $\mathcal{C} : \Theta \to \mathcal{Z}$ be a relation that takes the weight matrix $\boldsymbol{\theta}$ as input and outputs the solution S, i.e., $S \sim \mathcal{C}(\boldsymbol{\theta})$. To illustrate the notation, consider the following. (1) In the BOC problem [6], the objective is to group the M arms into K clusters based on their similarities. The clustering of arms is represented by the cluster index vector $\boldsymbol{c} \coloneqq [c_1, \ldots, c_M]$, where $c_m = k$ denotes that arm m belongs to cluster k. Here, the solution set $\mathcal{Z} = [K]^M$ is the set of all possible cluster index vectors. Suppose that the number of clusters $K = 2$ and the arm 1 belong to the cluster 1 and the remaining arms belong to the cluster 2, then the cluster index vectors $\boldsymbol{c}_1 = [1, 2, \ldots, 2]$ and $\boldsymbol{c}_2 = [2, 1, \ldots, 1]$ convey the same information and the solutions $S_1 = \boldsymbol{c}_1$ and $S_2 = \boldsymbol{c}_2$ are equivalent. The relation $\mathcal{C}$ is a classical data-point clustering algorithm, such as Single Linkage clustering, which treats the collection of parameters $\boldsymbol{\theta}\boldsymbol{\alpha}_m$ as M points and groups them into K clusters. (2) In the BAI problem [3], the parameters are one-dimensional ($d = 1$) and the objective is to identify the arm with the highest parameter. Since the best arm can be any one of the M arms, the solution set $\mathcal{Z} = \{1, \ldots, M\}$. Suppose that the arm 1 has the highest parameter; then the solution is $S = 1$. The relation $\mathcal{C}$ is a maximizer, $\arg\max_{m\in[M]} \boldsymbol{\theta}\boldsymbol{\alpha}_m$, which identifies the arm with the highest parameter.

We sequentially observe the samples under the bandit framework: at each time t, based on the past $t-1$ observed samples, we can adaptively select an arm $A_t \in [M]$ and observe a sample. Several pure exploration algorithms under standard [3], [6]–[8] and linear bandits [9] have been proposed to estimate the respective solution using as few samples as possible under the constraint that the error probability is upper bounded by $\delta \in (0, 1)$. These sequential algorithms consist of a sampling rule that adaptively selects an arm A_t to sample, a stopping rule that decides whether to stop or continue sampling, and a declaration rule that outputs the estimated solution. However, the sampling rule in these algorithms is typically computationally expensive, as it requires solving an optimization problem at each time step t [3], [6]–[9]. The objective of this work is to design a computationally efficient sampling rule that can be integrated with these existing pure exploration algorithms, with sample complexity performance comparable to that of their original counterparts.

III. Computationally Efficient Pure Exploration (CEPE) Algorithm

We propose the Computationally Efficient Pure Exploration (CEPE) Algorithm, whose pseudocode is presented in Algorithm 1. The CEPE algorithm, like any other pure exploration algorithm, does the following at each time t.

- *Sampling:* Adaptively select an arm A_t to sample.
- *Stopping:* Decide whether to stop or continue.
- *Declaration:* Upon stopping, estimate the solution $\hat{S}$.

Our contribution is in the design and theoretical analysis of an efficient sampling rule, while the stopping and declaration rules are adopted from existing work.

A. Sampling Rule Design

Optimal sample allocation: Define the probability simplex $\mathcal{P}_M \coloneqq \{\boldsymbol{w} \in \mathbb{R}_+^M \sum_{m=1}^M w_m = 1\}$, where $\boldsymbol{w} = [w_1, \ldots, w_M]$. Define the alternative space of $\boldsymbol{\theta}$, denoted by $\text{Alt}(\boldsymbol{\theta})$, as the set of the problem instance $\boldsymbol{\lambda}$ that corresponds to a different solution than that of $\boldsymbol{\theta}$, i.e., $\text{Alt}(\boldsymbol{\theta}) \coloneqq \{\boldsymbol{\lambda} \in \Theta : \mathcal{C}(\theta) \nsim \mathcal{C}(\boldsymbol{\lambda})\}$, where any $\boldsymbol{\lambda} \in \text{Alt}(\boldsymbol{\theta})$ is an alternative problem instance to $\boldsymbol{\theta}$. Let $D_{\text{KL}}^{(m)}(\boldsymbol{\rho}_1, \boldsymbol{\rho}_2)$ denote the KL divergence of the parametric distribution associated with arm m, $P_m(\cdot \mid \boldsymbol{\rho})$, evaluated between the parameters $\boldsymbol{\rho}_1$ and $\boldsymbol{\rho}_2$. Define $\mathcal{S}^*(\boldsymbol{\theta}) \coloneqq \arg\max_{\boldsymbol{w}\in\mathcal{P}_M} \psi(\boldsymbol{w}, \boldsymbol{\theta})$, where $\psi(\boldsymbol{w}, \boldsymbol{\theta}) \coloneqq \inf_{\boldsymbol{\lambda}\in\text{Alt}(\boldsymbol{\theta})} \sum_{m=1}^M w_m D_{\text{KL}}^{(m)}(\boldsymbol{\theta}\boldsymbol{\alpha}_m, \boldsymbol{\lambda}\boldsymbol{\alpha}_m)$. This maximization problem appears in the expression of the expected sample complexity lower bound slope in pure exploration problems [6, Theorem 1] and the optimal set $\mathcal{S}^*(\boldsymbol{\theta})$ has been proven to be convex, compact, and non-empty [6, Section III]. Here, $\boldsymbol{w}$ can be interpreted as the arm pull proportions, i.e., the fraction of times each arm is sampled, and this interpretation stems from the proof of the lower bound [6, Appendix A]. Therefore, $\mathcal{S}^*(\boldsymbol{\theta})$ is the set of all optimal arm pull proportions that maximize the weighted KL-distance between the problem instance and its closest alternative, and the sampling rules of the pure exploration algorithms are designed to match these optimal proportions of arm pulls [6, (4)].

Define $\boldsymbol{N}(t) \coloneqq [N_1(t), \ldots, N_M(t)]$, where $N_m(t)$ is the number of samples observed from arm m until time t. We assume that the weight matrix estimator is consistent[1], i.e., the estimated weight matrix $\hat{\boldsymbol{\theta}}(t)$ converge to the true weight matrix $\boldsymbol{\theta}$, as $t \to \infty$. The objective behind the design of the sampling rule is to make the empirical proportions $\frac{\boldsymbol{N}(t)}{t}$ converge to the optimal arm pull proportions $\boldsymbol{w}^*$, for some $\boldsymbol{w}^* \in \mathcal{S}^*(\boldsymbol{\theta})$. Hence, at each time t, we select an arm to make the empirical proportions $\frac{\boldsymbol{N}(t)}{t}$ move towards the maximizer. That is, at each time t, we select an arm to be pulled at next time A_{t+1} corresponding to the coordinate along which the gradient of $\psi(\hat{\boldsymbol{\theta}}(t), \boldsymbol{w})$ at $\boldsymbol{w} = \frac{\boldsymbol{N}(t)}{t}$ is maximized (Line 7 of Algorithm 1). From Danskin's theorem [12, Proposition A.22], the gradient $\nabla_{\boldsymbol{w}} \psi(\hat{\boldsymbol{\theta}}(t), \boldsymbol{w})\big|_{\boldsymbol{w}=\frac{\boldsymbol{N}(t)}{t}}$ can be computed as

$$\left[D_{\text{KL}}^{(1)}(\hat{\boldsymbol{\theta}}(t)\boldsymbol{\alpha}_1, \boldsymbol{\lambda}^*\boldsymbol{\alpha}_1), \ldots, D_{\text{KL}}^{(M)}(\hat{\boldsymbol{\theta}}(t)\boldsymbol{\alpha}_M, \boldsymbol{\lambda}^*\boldsymbol{\alpha}_M)\right],$$

where $\boldsymbol{\lambda}^* \in \underset{\boldsymbol{\lambda} \in \text{Alt}(\hat{\boldsymbol{\theta}}(t))}{\arg\inf} \sum_{m=1}^{M} \frac{\boldsymbol{N}(t)}{t} D_{\text{KL}}^{(m)}(\hat{\boldsymbol{\theta}}(t)\boldsymbol{\alpha}_m, \boldsymbol{\lambda}\boldsymbol{\alpha}_m)$.

In addition, to ensure the convergence of the estimated weight matrix $\hat{\boldsymbol{\theta}}(t)$ to the true parameter $\boldsymbol{\theta}$, we do forced exploration, where if there exists any arm m whose number of arm pulls $N_m(t)$ is less than $\sqrt{t}$, then the arm with the least number of arm pulls will be sampled (Line 5).

B. Theoretical Analysis

Define $f_t(\cdot) = \psi\left(\hat{\boldsymbol{\theta}}(t), \cdot\right)$, $f(\cdot) = \psi\left(\boldsymbol{\theta}, \cdot\right)$ and $\boldsymbol{w}_t = \frac{\boldsymbol{N}(t)}{t}$. Note that, the empirical proportions $\boldsymbol{w}_t$ are updated as $\boldsymbol{w}_{t+1} = (1-\gamma_t)\boldsymbol{w}_t + \gamma_t \boldsymbol{e}_{A_{t+1}}$, where $\gamma_t = \frac{1}{t+1}$, $\boldsymbol{e}_i$ is the standard basis vector whose i^{th} entry is 1 and all other entries are 0, and

$$A_{t+1} = \begin{cases} \arg\min_{i \in [M]} (\boldsymbol{w}_t)_i & \text{if } \min_{i \in [M]} (\boldsymbol{w}_t)_i < \frac{1}{\sqrt{t}} \\ \arg\max_{i \in [M]} \left(\nabla f_t(\boldsymbol{w}_t)\right)_i & \text{otherwise} \end{cases},$$

with the notation $(\cdot)_i$ indicating i^{th} coordinate of the vector. This corresponds to the Frank–Wolfe update [13] on the probability simplex, where each iterate is a convex combination of the previous iterate and the standard basis vector corresponding to the maximum gradient, with the following differences: (i) f_t varies with t and converges to f; and (ii) gradient-based updates are skipped for $\sqrt{t}$ iterations to model forced exploration.

Assumption 1. $\lim_{T \to \infty} \sum_{t=1}^{T} \max_{w \in \mathcal{P}_M} |f_t(\boldsymbol{w}) - f(\boldsymbol{w})|$ *is finite.*

Assumption 2. $-\lambda_{min}\left(\nabla^2_{\boldsymbol{w}} f(\boldsymbol{w})\right) \leq C_f \quad \forall \boldsymbol{w} \in \mathcal{P}_M$, *where* $\lambda_{min}(\cdot)$ *denotes the minimum eigen value.*

Since $\hat{\boldsymbol{\theta}}(t)$ is a consistent estimator and the function $\psi(\cdot, \cdot)$ is known to be continuous [6, Lemma], we have $f_t(\cdot)$ converges to $f(\cdot)$. However, to tackle the time-varying nature of f_t in our convergence analysis, we require a stronger convergence assumption in Assumption 1. Assumption 2 is a standard condition in the optimization literature [13, Lemma 5]. We use $\|\cdot\|_\infty$ to denote the ℓ-infinity norm, i.e., $\|\boldsymbol{w}\|_\infty = \max_{m \in [M]} |w_m|$. Define $d_\infty(\boldsymbol{w}, \mathcal{S}) \coloneqq \min_{\boldsymbol{w}_1 \in \mathcal{S}} \|\boldsymbol{w} - \boldsymbol{w}_1\|_\infty$.

Theorem 1. *Under Assumptions 1 and 2, the proposed CEPE satisfies* $\lim_{t \to \infty} d_\infty\left(\frac{\boldsymbol{N}(t)}{t}, \mathcal{S}^*(\boldsymbol{\theta})\right) = 0$ *almost surely.*

Proof. Section V. □

Stopping Rule: The stopping rule for most pure exploration algorithms [3], [6], [9] in the literature uses the Generalized Likelihood Ratio (GLR) Test [2]. It has been verified [6, (4)] that log-GLR at time t simplifies to $t\psi(\hat{\boldsymbol{\theta}}(t), \boldsymbol{N}(t)/t)$, and we use this as test statistics, denoted by $Z(t)$. The algorithm stops when the test statistic $Z(t)$ exceeds the threshold $\beta(\delta, t)$ (Line 12). We use the threshold from the literature for the specific problem under consideration. For example, in our simulations (Section IV-B), we use $\beta(\delta, t) = \log(100t^{1.2}/\delta)$ for standard bandits BAI [Proposition 12] [3] and $\beta(\delta, t) = 2\log(\det(\sum_{s=1}^{t} \boldsymbol{\alpha}_{A_s} \boldsymbol{\alpha}_{A_s}^T + I_n)^{1/2}/\delta)$ for linear bandits BAI [9, Proposition 2], where I_n is the identity matrix.

Declaration Rule: Upon stopping, the algorithm uses the estimated weight matrix $\hat{\boldsymbol{\theta}}(t)$ to declare the estimate of the solution, i.e., $\hat{S} = \mathcal{C}(\hat{\boldsymbol{\theta}}(t))$ (Line 13).

Algorithm 1 *CEPE Algorithm*

1: **Input:** δ, M
2: **Initialize:** $t = 0$, $N_m(0) = 0, \forall m \in [M]$.
3: **repeat**
4: **if** $\min_{m \in [M]} N_m(t) < \sqrt{t}$ **then**
5: $A_{t+1} = \arg\min_{m \in [M]} N_m(t)$
6: **else**
7: $A_{t+1} = \arg\max_{m \in [M]} D_{\text{KL}}^{(m)}(\hat{\boldsymbol{\theta}}(t)\boldsymbol{\alpha}_m, \boldsymbol{\lambda}^*\boldsymbol{\alpha}_m)$
8: **end if**
9: Observe a sample from arm A_{t+1}.
10: $t \leftarrow t+1$, $N_{A_t}(t) = N_{A_t}(t-1) + 1$. Update $\hat{\boldsymbol{\theta}}(t)$.
11: Compute $Z(t) = t\psi\left(\frac{\boldsymbol{N}(t)}{t}, \hat{\boldsymbol{\theta}}(t)\right)$ and $\beta(\delta, t)$.
12: **until** $Z(t) \geq \beta(\delta, t)$
13: $\hat{S} = \mathcal{C}(\hat{\boldsymbol{\theta}}(t))$
14: **Output:** $\hat{S}$

Remark 1. *Note that test statistics of the stopping threshold* $Z(t)$ *can be written as* $\frac{\boldsymbol{N}(t)^T}{t} \nabla_{\boldsymbol{w}} \psi\left(\hat{\boldsymbol{\theta}}(t), \boldsymbol{w}\right)\big|_{\boldsymbol{w}=\frac{\boldsymbol{N}(t)}{t}}$. *Hence, the gradient required for the CEPE sampling rule is calculated as part of* $Z(t)$. *This results in a significant reduction in CEPE runtime compared to pure exploration algorithms [3], [6]–[9], which otherwise solve an optimization problem* $\arg\max_{\boldsymbol{w} \in \mathcal{P}_M} \psi\left(\boldsymbol{w}, \hat{\boldsymbol{\theta}}(t)\right)$ *at each time step.*

Remark 2. *For an algorithm to be asymptotically order-optimal, its empirical proportions* $\frac{\boldsymbol{N}(t)}{t}$ *must converge to the optimal proportions set* $\mathcal{S}^*(\boldsymbol{\theta})$. *Hence, the sampling rule of CEPE, with this convergence guarantee (Theorem 1), can replace the sampling rules in the literature [3], [9] and achieve asymptotic order-optimality while remaining computationally efficient. We validate this informal claim of asymptotic order-optimality through simulation results (Figs. 1 and 2).*

[1]Common choices of a consistent estimator are the maximum likelihood estimator [6] and the linear least squares estimator [9].

[2]Modified GLR has been defined and used in the stopping rule in [8], however, our efficient sampling rule can still be integrated with this algorithm.

IV. Comparison with existing methods

In this section, we compare the proposed CEPE algorithm with standard bandits BAI algorithms, namely Track-and-Stop (T&S) [3] and Transportation Cost Balancing (TCB) [10], as well as the linear bandit BAI algorithm Lazy Track-and-Stop (LT&S). First, we provide a comparison of computational complexity. Then, through simulations, we compare expected sample complexity and per-sample runtime.

A. Theoretical-Computational Complexity

Table I presents the computational complexity orders of the various algorithms. T&S and LT&S require solving an optimization problem through an iterative procedure, which typically results in a complexity higher than that of the stopping rule by at least a factor of T_S, where T_S denotes the number of iterations. TCB, an efficient algorithm for standard bandits BAI, performs a one-step update of that optimization problem for the sampling rule, involving gradient computation and yielding a complexity comparable to that of the stopping rule. The proposed CEPE algorithm leverages quantities computed in the stopping rule to identify the highest gradient component, resulting in a sampling rule with complexity lower than that of the stopping rule.

		Sampling	Stopping	Declaration
standard bandits-BAI	T&S	MT_S	M	M
	TCB	M	M	M
	CEPE	1	M	M
Linear bandits-BAI	LT&S	$T_S n^2 M^3$	n^2M^2	nM
	CEPE	nM	n^2M^2	nM

TABLE I
COMPUTATIONAL COMPLEXITY ORDER COMPARISON.

B. Simulation Results

For the expected sample complexity plots (Figs. 1 and 2), the results are averaged over 1000 trials. We numerically evaluate the slope of the expected sample complexity curves by fitting a line to the last four points in the asymptotic regime using linear regression, and use shaded regions around the plots to indicate the standard deviation.

BAI in Standard Bandits: Consider a standard bandits BAI problem with $M = 6$ Gaussian arms with means $1, 2, 3, 5, 6, 7$ and unit variance. From Fig. 1, we can see that our proposed CEPE uses a smaller expected number of samples compared to T&S and TCB for a fixed error probability δ. CEPE has the numerical slope comparable to that of T&S and TCB, and all of them satisfy theoretical lower [3, Theorem 1] and upper bounds [3, Theorem 14]. From Table II, we observe that per-sample runtimes of the sampling rule in CEPE are significantly lower and, as a consequence, the overall per-sample runtimes of CEPE are better than those of T&S and TCB.

BAI in Linear Bandits: Consider a linear bandits BAI problem as in [9] with $M = 4$ arms, whose feature vectors are $[1, 0]$, $[1/\sqrt{2}, 1/\sqrt{2}]$, $[1/\sqrt{2}, -1/\sqrt{2}]$, and $[1/2, \sqrt{3}/2]$, with weight vector $[1, 0]$, and zero-mean Gaussian noise with variance 0.09. From Fig. 2, we observe that the numerical slope of CEPE is lower than that of LT&S and satisfies the theoretical lower [9, Theorem 1] and upper bound [9, Theorem 3]. From Table III, we observe that the sampling rule dominates the per-sample runtime for LT&S. In contrast, the per-sample runtime of CEPE is negligible, leading to a significant improvement in overall per-sample runtime compared to LT&S.

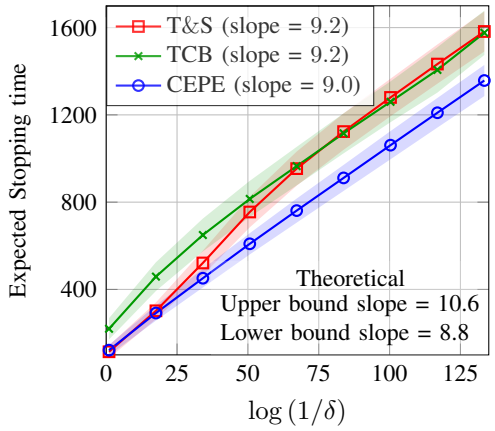

Fig. 1. BAI in standard bandits

Number of arms M		10	100	1000
Sampling Rule (microseconds)	T&S	85.6	93.6	213.4
	TCB	32.6	40.5	69.8
	CEPE	4.9	6.7	7.1
Overall (microseconds)	T&S	107.6	111.4	248.9
	TCB	52.7	61.7	73.3
	CEPE	26.8	27.5	36.8

TABLE II
AVERAGE PER-SAMPLE RUNTIMES FOR STANDARD BANDITS BAI.

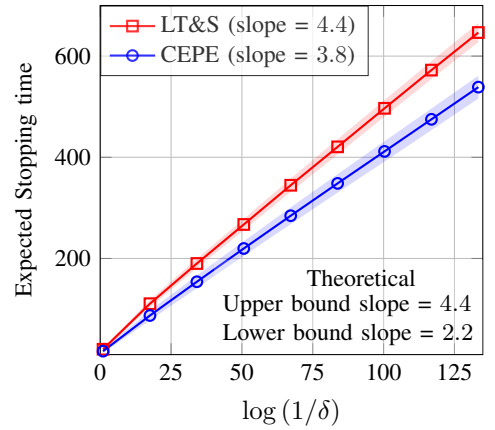

Fig. 2. BAI in Linear Bandits

Number of arms M		4	10	20
Sampling Rule (milliseconds)	LT&S	8.881	12.260	37.083
	ESR	0.038	0.042	0.054
Overall (milliseconds)	LT&S	8.889	12.354	37.237
	ESR	0.139	0.165	0.209

TABLE III
AVERAGE PER-SAMPLE RUNTIMES FOR LINEAR BANDITS BAI

V. Proof of Theorem 1

Proof. Given that

$$\boldsymbol{w}_{t+1} = (1-\gamma_t)\boldsymbol{w}_t + \gamma_t \boldsymbol{e}_{A_{t+1}} = \boldsymbol{w}_t + \gamma_t(\boldsymbol{e}_{A_{t+1}} - \boldsymbol{w}_t).$$

From the second-order mean value theorem, we have

$$f_t(\boldsymbol{w}_{t+1}) = f_t(\boldsymbol{w}_t) + \gamma_t(\boldsymbol{e}_{A_{t+1}} - \boldsymbol{w}_t)^T \nabla f_t(\boldsymbol{w}_t)$$

$$+ \gamma_t^2 (\boldsymbol{e}_{A_{t+1}} - \boldsymbol{w}_t)^T \nabla^2 f_t(\tilde{\boldsymbol{w}}_t)(\boldsymbol{e}_{A_{t+1}} - \boldsymbol{w}_t)$$

for some $\tilde{\boldsymbol{w}}_t$ in the line segment joining $\boldsymbol{w}_t$ and $\boldsymbol{w}_{t+1}$. Note that since $\boldsymbol{e}_{A_{t+1}}, \boldsymbol{w}_t \in \mathcal{P}_M$, we have $\|\boldsymbol{e}_{A_{t+1}} - \boldsymbol{w}_t\|^2 \leq 2$. In addition, using Assumption 2, we have

$$f_t(\boldsymbol{w}_{t+1}) \geq f_t(\boldsymbol{w}_t) + \gamma_t (\boldsymbol{e}_{A_{t+1}} - \boldsymbol{w}_t)^T \nabla f_t(\boldsymbol{w}_t) - \gamma_t^2 C_f. \quad (1)$$

Case 1: At time t, $\min_{i\in[M]}(\boldsymbol{w}_t)_i \geq 1/\sqrt{t}$.
We have $A_{t+1} = \arg\max_{i\in[M]} (\nabla f_t(\boldsymbol{w}_t))_i$. Now, we write

$$(\boldsymbol{e}_{A_{t+1}} - \boldsymbol{w}_t)^T \nabla f_t(\boldsymbol{w}_t) = \max_{i\in[M]} (\nabla f_t(\boldsymbol{w}_t))_i - \boldsymbol{w}_t^T \nabla f_t(\boldsymbol{w}_t)$$

$$\geq \boldsymbol{w}^{*T} \nabla f_t(\boldsymbol{w}_t) - \boldsymbol{w}_t^T \nabla f_t(\boldsymbol{w}_t), \quad (2)$$

for any $\boldsymbol{w}^* \in \mathcal{S}^*(\boldsymbol{\theta})$. From the concavity of $f_t(\cdot)$, we have

$$(\boldsymbol{w}^* - \boldsymbol{w}_t)^T \nabla f_t(\boldsymbol{w}_t) \geq f_t(\boldsymbol{w}^*) - f_t(\boldsymbol{w}_t) \triangleq h_t(\boldsymbol{w}_t). \quad (3)$$

Substituting (2) and (3) in (1), we get

$$f_t(\boldsymbol{w}_{t+1}) \geq f_t(\boldsymbol{w}_t) + \gamma_t h_t(\boldsymbol{w}_t) - \gamma_t^2 C_f. \quad (4)$$

Case 2: At time t, $\min_{i\in[M]}(\boldsymbol{w}_t)_i \leq 1/\sqrt{t}$.
From the concavity of $f_t(\cdot)$ in (1), we get

$$\begin{aligned} f_t(\boldsymbol{w}_{t+1}) &\geq f_t(w_t) + \gamma_t \left[f_t(\boldsymbol{e}_{A_{t+1}}) - f_t(\boldsymbol{w}_t)\right] - \gamma_t^2 C_f \\ &= f_t(w_t) + \gamma_t \left[f_t(\boldsymbol{e}_{A_{t+1}}) - f_t(\boldsymbol{w}^*)\right] - \gamma_t^2 C_f + \gamma_t h_t(\boldsymbol{w}_t) \\ &\geq f_t(w_t) - \gamma_t \left|f_t(\boldsymbol{e}_{A_{t+1}}) - f_t(\boldsymbol{w}^*)\right| - \gamma_t^2 C_f + \gamma_t h_t(\boldsymbol{w}_t). \end{aligned}$$

Since the continuous function $f_t(\cdot)$ on the compact space $\mathcal{P}_M$ is bounded, there exists a finite $D_f > 0$ such that

$$f_t(\boldsymbol{w}_{t+1}) \geq f_t(\boldsymbol{w}_t) - \gamma_t D_f - \gamma_t^2 C_f + \gamma_t h_t(\boldsymbol{w}_t). \quad (5)$$

Define $E_t := \left\{\min_{i\in[M]}(\boldsymbol{w}_t)_i \leq 1/\sqrt{t}\right\}$. On combining (4) and (5), for any time t, we get

$$f_t(\boldsymbol{w}_{t+1}) \geq f_t(\boldsymbol{w}_t) - \gamma_t^2 C_f + \gamma_t h_t(\boldsymbol{w}_t) - \gamma_t D_f \mathbb{1}_{E_t}.$$

On telescoping till T, we get

$$\sum_{t=1}^{T} [f_t(\boldsymbol{w}_{t+1}) - f_t(\boldsymbol{w}_t)] \geq \sum_{t=1}^{T} [\gamma_t h_t(\boldsymbol{w}_t) - \gamma_t^2 C_f - \gamma_t D_f \mathbb{1}_{E_t}].$$

The above expression can be expanded as

$$\sum_{t=1}^{T} [f_t(\boldsymbol{w}_{t+1}) - f(\boldsymbol{w}_{t+1})] + \sum_{t=1}^{T} [f(\boldsymbol{w}_t) - f_t(\boldsymbol{w}_t)] + f(\boldsymbol{w}_{T+1})$$

$$- f(\boldsymbol{w}_1) \geq \sum_{t=1}^{T} \gamma_t h_t(\boldsymbol{w}_t) - \sum_{t=1}^{T} \gamma_t^2 C_f - \sum_{t=1}^{T} \gamma_t D_f \mathbb{1}_{E_t}$$

Let $T \to \infty$. From Assumption 1, the LHS is finite. The second and third RHS terms are finite, since both are sums of inverse squares. Therefore, $\lim_{T\to\infty} \sum_{t=1}^{T} \gamma_t h_t(\boldsymbol{w}_t) < \infty$. We write $h_t(\boldsymbol{w}_t) = [f_t(\boldsymbol{w}^*) - f(\boldsymbol{w}^*)] + [f(\boldsymbol{w}_t) - f_t(\boldsymbol{w}_t)] + [f(\boldsymbol{w}^*) - f(\boldsymbol{w}_t)]$. Now, again using Assumption 1, we get

$$\lim_{T\to\infty} \sum_{t=1}^{T} \gamma_t [f(\boldsymbol{w}^*) - f(\boldsymbol{w}_t)] < \infty.$$

Hence $f(\boldsymbol{w}_t)$ converges to $f(\boldsymbol{w}^*)$. Since the function $f(\cdot)$ is continuous, and the maximizer set $\mathcal{S}^*(\boldsymbol{\theta})$ is the compact, and non-empty, we get $\boldsymbol{w}_t$ converges to $\boldsymbol{w}^*$, for some $\boldsymbol{w}^* \in \mathcal{S}^*$. This proves the theorem. □

VI. Conclusion

We propose a computationally efficient pure exploration (CEPE) algorithm for solving pure exploration problems in standard and linear bandits under the fixed-confidence setting. We prove that the sampling rule of CEPE ensures that the empirical proportions converge to the optimal arm-pull proportions under mild assumptions, using a Frank–Wolfe-based proof technique. Through our computational complexity analysis, we show that, for existing pure exploration algorithms for BAI problems, the computational complexity is dominated by the sampling rule; in contrast, the proposed CEPE algorithm ensures that its sampling rule has a complexity lower than that of the stopping rule, thereby making the algorithm computationally efficient. Simulation results demonstrate that CEPE significantly reduces the per-sample runtime compared to existing algorithms for the Best Arm Identification (BAI) problem in both standard and linear bandits, while also yielding comparable expected sample complexity.

References

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

[2] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, pp. 235–256, 2002.

[3] A. Garivier and E. Kaufmann, "Optimal best arm identification with fixed confidence," in *Conference on Learning Theory*. PMLR, 2016, pp. 998–1027.

[4] J. Yang and V. Tan, "Minimax optimal fixed-budget best arm identification in linear bandits," *Advances in Neural Information Processing Systems*, vol. 35, pp. 12 253–12 266, 2022.

[5] J. Yang, Z. Zhong, and V. Y. Tan, "Optimal clustering with bandit feedback," *The Journal of Machine Learning Research*, vol. 25, pp. 1–54, 2024.

[6] G. D. Chandran, K. S. Reddy, and S. Bhashyam, "Online clustering with bandit information," 2025. [Online]. Available: https://arxiv.org/abs/2501.11421

[7] A. Deshmukh, V. V. Veeravalli, and S. Bhashyam, "Sequential controlled sensing for composite multihypothesis testing," *Sequential Analysis*, vol. 40, no. 2, pp. 259–289, 2021.

[8] G. R. Prabhu, S. Bhashyam, A. Gopalan, and R. Sundaresan, "Sequential multi-hypothesis testing in multi-armed bandit problems: An approach for asymptotic optimality," *IEEE Transactions on Information Theory*, vol. 68, no. 7, pp. 4790–4817, 2022.

[9] Y. Jedra and A. Proutiere, "Optimal best-arm identification in linear bandits," *Advances in Neural Information Processing Systems*, vol. 33, pp. 10 007–10 017, 2020.

[10] A. Mukherjee and A. Tajer, "Efficient best arm identification in stochastic bandits: Beyond β–optimality," *IEEE Transactions on Information Theory*, 2024.

[11] G. D. Chandran, K. S. Reddy, and S. Bhashyam, "Efficient clustering in stochastic bandits," 2026. [Online]. Available: https://arxiv.org/abs/2601.09162

[12] D. P. Bertsekas, "Control of uncertain systems with a set-membership description of the uncertainty." Ph.D. dissertation, Massachusetts Institute of Technology, 1971.

[13] M. Jaggi, "Revisiting frank-wolfe: Projection-free sparse convex optimization," in *International conference on machine learning*. PMLR, 2013, pp. 427–435.