informs
http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0030-364X, EISSN 1526-5463

Submitted to *Operations Research*

# Pure Exploration Linear Bandit Drug Discovery

**(Authors' names blinded for peer review)**



**Abstract.**

**Key words:**

## 1. Introduction

The process of drug discovery stands as one of the most challenging and high-stakes endeavors in contemporary science and industry, primarily characterized by its exorbitant costs and precarious outcomes (Lou and Wu 2021). It can span up to a decade of intensive research and development, coupled with a financial commitment ranging from $2 to $3 billion, with a low likelihood of achieving success at around a mere 1%. (Das et al. 2021a). Adding to this complexity, the expenses associated with developing new drugs have continued to rise over the years, a trend known as Eroom's law or anti-Moore's law, where the rate of new drug approvals per billion dollars spent on research and development halves approximately every nine years since the 1950s (Ringel et al. 2020).

Traditionally, the drug discovery process revolves around the high-throughput screening of active molecules. However, this extensive screening process is incredibly demanding in terms of both time and resources. An enormous amount of effort goes into screening a vast array of potential molecules, and painstaking analysis is required to identify and select the most promising candidates from an exponentially expanding pool of molecules.

One of the primary approaches in sequential experimentation involves randomly selecting molecules for evaluation, aiming to ensure an unbiased sample. However, while this method maintains impartiality, it is riddled with inefficiencies and resource-intensive challenges. Its effectiveness is hampered by the sheer magnitude of possibilities, making it highly improbable to successfully identify effective molecules through random selection alone (Powell and Ryzhov 2012). An alternative strategy is to test molecules with structural resemblances to previously discovered target molecules. This approach exhibits a more promising initial success rate and a streamlined focus. Nonetheless, it comes with certain limitations, particularly regarding its restricted exploration. Depending solely on this method might lead to the oversight of potentially effective molecules with distinct structures from known targets, ultimately leading to prolonged inefficiencies (Ravina 2011). Another avenue in drug discovery involves harnessing expert knowledge to guide molecular design. This approach empowers decision-makers with informed choices and offers the flexibility to customize molecular structures based on accumulated expertise. However, it is susceptible to expert bias, potentially resulting in the inadvertent dismissal of novel molecules that lie beyond the scope of current expert understanding. Lastly, modern drug discovery benefits from the integration of data-driven algorithms, which have shown promise in efficiently identifying potential drug candidates. These algorithms possess the advantage of adaptability over time as they accumulate exposure to increasingly extensive datasets. Nonetheless, their performance hinges critically on the availability and quality of training data. In instances where the data proves inadequate or tainted by bias, these algorithms run the risk of yielding poor results, a scenario with profound consequences, including the identification of ineffective or misleading drug candidates (Cai et al. 2020).

In this paper, we aim to identify all $\varepsilon$-best arms. $\varepsilon$-best is an intuitive and robust description in the context of the bandit, meaning that we want to identify the arms whose gaps from the best arm fall within a certain range. The concept of all $\varepsilon$-best arms is particularly beneficial in scenarios where a wide exploration of nearly optimal options can lead to significant breakthroughs or efficiencies. For example, in drug discovery, drug development is costly and risky. By advancing multiple promising candidates, the strategy distributes risk and increases the chances that at least one candidate will successfully progress through clinical trials and regulatory approval. In addition, other compounds slightly less effective in preliminary tests might exhibit better properties later, such as fewer side effects, easier manufacturing, or lower resistance rates.

For some other examples, like e-commerce or streaming services, all $\varepsilon$-best (e.g., products, movies, or songs) can enhance user satisfaction. This approach allows for a diversity of recommendations that cater to slightly varying tastes or preferences, rather than recommending the same top items to everyone. In manufacturing, different processes might yield products of nearly identical quality. Identifying many process settings can help in optimizing production lines for cost, speed, and quality, providing flexibility in response to supply chain or demand changes. In investment portfolio management, identifying all $\varepsilon$-best investment opportunities (e.g., stocks, bonds, or other assets) can help in diversifying investments to manage risk and

optimize returns. This strategy ensures that portfolio managers are not overly reliant on a small number of assets that appear optimal under current conditions but might be volatile or risky.

~~Then, we may start discussing the design of experiments, pure exploration in multi-armed bandits, best arm identification and Top-$K$ identification, structures bandits including linear bandits, and adaptive sequential experiments.~~ Then, we may start discussing the drug discovery sequential experiment design, pure exploration in multi-armed bandits, and structured bandits including linear bandits.

(ZK L: To be polished.)

### 1.1. Main Contributions

This paper makes both algorithmic and theoretical contributions. Considering the algorithmic aspect, we develop a phase-based semi-adaptive algorithm. The newly proposed Linear Fast Arm Classification with Threshold Estimation (LinFACTE) algorithm could have a direct impact on various application areas where a linear structure is assumed and the objective is to screen out multiple candidates. On account of the feasibility and high efficiency, the proposed policy can be easily adapted to solve problems beyond the scope of this paper's focus on drug discovery, such as streaming service, manufacturing, and investment portfolio management.

This paper also makes several theoretical contributions. On the one hand, we provide an explicit description of the problem complexity of the All $\varepsilon$-Best Arms Identification problem in the linear bandit. This lower bound, to the best of our knowledge, is the first result in the literature up to now. We also graphically explain how the problem can be interpreted, which can be of broad interest to provide general guidance on understanding the pure exploration problem in the linear bandit. On the other hand, we give three different upper bounds on the expected sample complexity of the LinFACTE algorithm, providing some insights into the theoretical analysis of our algorithm and the distinction between different optimal design criteria. The last upper bound is proved to be instance-optimal up to some logarithmic term, closing the gap.

TM: Our experiment shows that our theoretically guaranteed algorithm is superior to former algorithms in time complexity and correctly proposing all $\epsilon$-best arms.

(ZK L: To be polished.)

### 1.2. Related Literature

(ZK L: The following "delete" operations are just for the 12-page version of our paper.)

**1.2.1. Drug Discovery Sequential Experiment Design.** In recent years, the field of drug discovery has witnessed a significant acceleration due to the application of statistical and computational methods, a phenomenon often referred to as virtual or in silico screening in the medical domain (Negoescu et al. 2011). While many researchers have focused on accurately predicting molecular properties, such as chemical and biochemical attributes, employing machine learning and deep learning algorithms (Vamathevan

et al. 2019), the process of drug design involves a crucial element, the selection of molecules for synthesis and subsequent testing, constituting a sequential decision problem.

One particular line of research approaches this challenge by framing it as a ranking and selection problem, aiming to determine the sequence of experiments that maximizes the likelihood of identifying the best or most promising molecules within a predefined experimental budget. In this context, a seminal work by Negoescu et al. (2011) introduced the concept of the knowledge gradient, a Bayesian optimization technique. This approach has been instrumental in the pursuit of finding the single most promising compound while significantly reducing the number of required tests. Notably, this technique's utility extends beyond drug discovery, as it has been applied to identify targeted regions in RNA molecules (Li et al. 2018) and optimize the appropriate dosage in drug development (Nasrollahzadeh and Khademi 2022).

However, the knowledge gradient method has many problems. There is no theoretical guarantee for the sample complexity and convergence. In fact, based on computational results, the algorithm might not converge in some simple cases. Also, while Negoescu et al. (2011) take a linear model into consideration, the information structure is not leveraged in the knowledge gradient method.

~~Réda et al. (2020) (ZK L: To be finished.)~~

~~Recently, data-driven algorithms have been emerging for drug discovery (Negoescu et al. 2011, Negoescu et al. 2019, Negoescu et al. 2021). However, current algorithm-assisted drug optimization suffers from data hungriness in real-world drug discovery tasks (Cai et al. 2020), considering there is only limited availability of labeled data for new drug discovery.~~

~~In addition, these algorithms usually focus on a single property, ignoring the fact that multiple requirements need to be met (Millan 2006).~~

~~Bertsimas and Zhuo (2020) propose a data-driven optimization incorporating machine learning to search for precision cancer medicine.~~

~~Tallorin et al. (2018) use basic Bayesian optimization techniques from Powell and Ryzhov (2012) to guide the drug optimization for properties that cannot be biologically examined directly.~~

~~Among all ML algorithms for drug discovery, Bayesian optimization is a well-suited one for searching the potential candidates and fitting models on a large dataset simultaneously (Frazier and Wang 2016). Bayesian optimization aims to construct models for the objective and quantifies the uncertainty at the same time based on a Bayesian learning technique, with which researchers can sequentially and adaptive choose desired samples for evaluations Frazier (2018). Yoshida et al. (2018) achieves a 160-fold increase in the searching speed by using Bayesian optimization techniques in screening potent antimicrobial peptides and (ZK L: To be finished.)~~

~~Das et al. (2021b) Anishchenko et al. (2021) Heuristic methods to generate. (ZK L: To be finished.)~~

Our research setting, however, distinguishes itself from Negoescu et al. (2011) in several key aspects. First, Second... (ZK L: To be finished.)

**1.2.2. Pure Exploration Problem.** The stochastic multi-armed bandits (MABs) model has stood as a predominant and effective paradigm for delineating the classical exploration-exploitation trade-off since its inception by Thompson (1933) in the scope of medical trials. MABs essentially pose a sequential decision-making challenge (Robbins 1952), where the agent pulls the arm sequentially and observes the stochastic rewards drawn from the distribution associated with the chosen arm. While a significant body of research emphasizes the minimization of the cumulative regret (Bubeck et al. 2012, Lattimore and Szepesvári 2020), our focus lies on the pure exploration setting (Koenig and Law 1985), which aims at selecting a subset from $K$ arms through a two-stage sampling procedure based on some specific criterion.

The origins of pure exploration problems can be traced back to the 1950s within the context of the Ranking and Selection (R&S) problem, as initially addressed by Bechhofer (1954). Numerous methodologies have been proposed to tackle the canonical R&S problem since then. These include elimination-type algorithms (Kim and Nelson 2001, Bubeck et al. 2013, Fan et al. 2016), Optimal Computing Budget Allocation (OCBA) (Chen et al. 2000), knowledge-gradient algorithms (Frazier et al. 2008), UCB-type algorithms (Kalyanakrishnan et al. 2012, Kaufmann and Kalyanakrishnan 2013), and unified gap-based exploration (UGapE) algorithm (Gabillon et al. 2012). Additionally, it is noteworthy that Hong et al. (2021) provides a thorough overview of the R&S problem.

The general framework of pure exploration comprises various exploration tasks (Qin and You 2023), such as Best Arm Identification (BAI) (Mannor and Tsitsiklis 2004), Top-$K$(Kalyanakrishnan and Stone 2010, Kalyanakrishnan et al. 2012), Thresholding Bandit (Locatelli et al. 2016, Abernethy et al. 2016), and All $\varepsilon$-Best Arms Identification (Mason et al. 2020), and two classical performance criteria, i.e., the Fixed-Budget (FB) setting and the Fixed-Confidence (FC) setting. ~~The exploration task specifies the recommended decision following the sampling phase with a stopping rule designed based on some performance criterion.~~

In the fixed-budget setting, the number of draws is predetermined, and the algorithm aims to optimize the experiment budget with a pre-established confidence level. ~~This becomes especially relevant in scenarios involving a preliminary exploration phase, where costs are measured not in terms of rewards but rather in relation to the resources allocated within a limited budget (e.g. the number of drugs in the preclinical phase in the drug discovery setting (Réda et al. 2020)).~~ Whereas in the fixed-confidence setting, the objective is to identify the optimal answer set for the exploration task with a guaranteed level of confidence, all while minimizing the sample complexity. ~~Originating from the 1950s, fixed-confidence problems stem from the seminal work of Chernoff (1959), which focuses on the design of sequential experiments for binary hypothesis testing. In comparison to the fixed-budget setting and from an algorithm design perspective, the fixed-confidence setting introduces a fundamental difference in the sense of the stopping rule, a classical procedure in statistics (Wald 1992, Wald 1985, Wald 1970). The primary complexity in this sequential setting lies in the fact that the time taken to provide a recommendation arm in the fixed-confidence setting, depending on observed data, is a stopping time (Doob and Joseph 1990), rather than a fixed,~~

~~data-independent time. This necessitates that the hypothesis testing maintains a confidence level that is uniformly controlled throughout the testing procedure. In other words, the preset confidence level remains unaffected by any kind of stopping rules.~~

In the context of bandit literature, Best Arm Identification (BAI) stands out as a distinctive form of exploration task, extensively studied in pure exploration problem (Mannor and Tsitsiklis 2004, Even-Dar et al. 2006, Bubeck et al. 2009, Audibert et al. 2010, Garivier and Kaufmann 2021, Jourdan et al. 2024). Bubeck et al. (2011) exhibit that any sampling strategy designed for regret minimization falls short of expectations regarding simple regret in the fixed-budget best arm identification problem. Consequently, the algorithm for the pure exploration setting needs to be constructed from a fresh perspective. The concept of pure exploration gained traction under the concept of R&S problem, the work of which is summarized by Bechhofer et al. (1968) with the basic uniform sampling strategy and the fixed-confidence setting. In the fixed-confidence best arm identification problem, Paulson (1964) first introduced the idea of arm elimination, proving a substantial improvement on the sample complexity. This methodology was later promoted and further researched by Mannor and Tsitsiklis (2004) who establish the matching upper and lower bound in the form of $\Theta\left(\left(\frac{K}{\varepsilon^2}\right)\log\left(\frac{1}{\delta}\right)\right)$ on the maximum sample complexity among all sample paths. Besides, Even-Dar et al. (2006) proposed the worst optimal Median Elimination strategy which proves to match the minimax lower bound in Mannor and Tsitsiklis (2004). Following the rekindled research by Bubeck et al. (2009), Audibert et al. (2010), a noteworthy contribution in the development stream of pure exploration was made by Garivier and Kaufmann (2016). In the fixed-confidence setting, Garivier and Kaufmann (2016) first introduced the Track-and-Stop type algorithm, establishing matching lower bound and upper bounds for the BAI problem. In subsequent years, this algorithm pathway garnered some other attention, leading to further research aimed at enhancing its generality and numerical efficiency (Juneja and Krishnasamy 2019, Degenne et al. 2019). (CH: We may take this paragraph out of the 12-page version)

The conceptualization of exploration tasks aligns with a behavioral model rooted in rational decision-making (Simon 1955). Simon's seminal work broadened the classical economic notion of rationality by introducing the term 'satisficing', a combination of the words 'satisfy' and 'suffice', to refer to the thresholding concept and the confidence principle. The satisficing model seeks to make the decision-making process compatible with access to information and computational capacities. Garnering extensive attention, the satisficing problem has been systematically reviewed, with Michel et al. (2023) providing a well-organized summarization of related works. Notably, Reverdy et al. (2016) made a significant contribution by presenting a comprehensive classification method for the satisficing model within a general Bayesian framework, encompassing a wide range of bandit problems as special cases within this overarching structure.(ZK L: It seems that the content in this paragraph is a little strange. The content was introduced to explain related conceptions from a higher view) (CH: We may take this paragraph out of the 12-page version)

~~Then in the realm of specific exploration tasks, quite a lot of other exploration tasks have been proposed to cater to diverse applications except for the most common BAI problem.~~ The problem of BAI was first generalized to identify $m$ arms out of $K$ arms such that the expected reward of every chosen arm is at least $\mu_m - \varepsilon$, where $\mu_m$ is the $m^{th}$ highest expected reward among the arms. This variant is referred to as the Explore-m problem (thus, the special case studied by Even-Dar et al. (2006) is the Explore-1 problem). Under the Explore-m or Top-$K$ setting, Kalyanakrishnan and Stone (2010) extend the median elimination algorithm proposed by Even-Dar et al. (2006) to the Halving algorithm for Explore-m that conforms to the PAC condition similar to that of Explore-1, achieving a maximum sample complexity of $O\left(\left(\frac{K}{\varepsilon^2}\right)\log\left(\frac{m}{\delta}\right)\right)$. Additionally, other algorithms such as the LUCB algorithm have been proposed (Kalyanakrishnan et al. 2012), exploiting confidence bound to formulate its sampling strategy and stopping rule. Beyond methods that select arms by the index, there is a kind of algorithm aiming at identifying the set of arms with means above a specified threshold, up to a given precision (Locatelli et al. 2016, Abernethy et al. 2016). (CH: We may take this paragraph out of the 12-page version)

To be more general, characterized by a decision class, the Explore-$m$ problem and the Thresholding Bandit Problem are actually special cases of the Combinatorial Pure Exploration (CPE) problem (Chen et al. 2014). Chen et al. (2014) provides the Combinatorial Lower-Upper Confidence Bound (CLUCB) algorithm for the fixed-confidence setting and the Combinatorial Successive Accept Reject (CSAR) algorithm for the fixed-budget setting to reveal the non-trivial combinatorial essence of a series of problem, recovering the best available upper bounds for the Explore-m problem and the Multi-Bandit Best Arm Identification problem (Gabillon et al. 2011) up to some constant factors. (CH: We may take this paragraph out of the 12-page version)

As stated above, Drug Discovery is a natural and robust objective, in which pharmacologists aim to identify a set of highly potent drug candidates from potentially millions of compounds using various in vitro and in silico assays. Only the selected candidates undergo the following more extensive testing (Christmann-Franck et al. 2016). Since the final clinical efficacy is uncertain, it is important to identify multiple candidates at a time. Given the potential cost of performing these assays, there is a desire to employ an adaptive, sequential experiment design that requires fewer experiments compared to a fixed design.

To obtain multiple alternatives, both the objectives of finding the top-$K$ performing drugs and identifying all drugs above a threshold can cause failure. In the Top-$K$ approach, selecting a small $m$ may overlook potent compounds, while opting for a large $m$ may result in numerous ineffective compounds, necessitating an excessively high number of experiments. Setting a threshold faces similar issues, with the additional concern that setting it too high may lead to no drug discoveries. Moreover, both the Top-$K$ and thresholding problems necessitate some prior knowledge about the distribution of the arms to ensure a good performance of the recommended decision, and the time-varying nature of the arms distribution can add even more challenges in real applications. In contrast, the all $\varepsilon$-best objective of identifying all arms whose potency is

within 20% of the best avoids these concerns by providing a robust and natural guarantee: no significantly suboptimal arms will be returned, ensuring meaningful discoveries. However, the problem of finding all $\varepsilon$-best arms has been historically neglected in the past, while this exploration task can arguably be deemed the most reasonable objective in many scenarios. Mason et al. (2020) first introduced the complexity of All $\varepsilon$-Best Arms Identification and provided two lower bounds on the sample complexity: the first based on a classical change-of-measure argument exhibits the sample complexity behavior in the low confidence regime ($\delta \to 0$). The second bound, employing the Simulator technique (Simchowitz et al. 2017) combined with an algorithmic reduction to BAI, shows the sample complexity's dependency on the number of arms $K$ for moderate values of $\delta$. Subsequent research by Al Marjani et al. (2022) provided two corresponding tighter lower bounds compared to those in Mason et al. (2020).

**1.2.3. Pure Exploration for Linear Bandits.** Introduced by Abe and Long (1999), the linear bandit (LB) problem is a notable extension of MABs considering the structure among different arms. ~~The first work considering the algorithms based on the optimism principle for linear bandit was developed by Auer (2002).~~ In the context of pure exploration, based on the classical lower bound established by Garivier and Kaufmann (2016), Fiez et al. (2019) extend it in the setting of linear bandit using the same transportation inequalities ~~and standard tricks with respect to the linear bandit~~. ~~From this complexity description, it can be concluded that the optimal allocation strategy should be in proportion with the optimal distribution of budgets among different arms which forms the lower bound.~~ However, the calculation of optimal distribution $\boldsymbol{p}^*$ relies on the knowledge of $\boldsymbol{\mu}^*$ (which represents the structure of the bandit instance), which is unknown a prior, thus the optimal allocation is not a realizable strategy.

Short of $\boldsymbol{p}^*$, it is natural to introduce the concept of online learning to learn the structure of the bandit sequentially as the rewards are observed. The first work that considers the linear bandit problem in pure exploration was by Hoffman et al. (2014). They studied the BAI problem in the fixed-budget setting factoring correlation among different arm distributions and devised an algorithm called BayesGap, a Bayesian version of a gap-based exploration algorithm (Gabillon et al. 2012). (CH: Shall we move this to the previous paragraph?)(ZK L: Placing it here can maintain logical coherence.)

Although BayesGap outperformed algorithms that ignore the correlation and intrinsic structure, the drawback that it never pulls arms identified as sub-optimal significantly harms the performance of the algorithm in the linear bandit pure exploration setting. The essential difference between the stochastic MABs and the linear bandit strategies stems from the fact that in MABs an arm is no longer pulled as soon as its sub-optimality is evident (with high probability), while in the linear bandit setting, even a sub-optimal arm may offer valuable information about the parameter vector $\boldsymbol{\mu}^*$ and thus improve the confidence of the estimation in discarding among near-optimal arms. (CH: We may mention this in the experiment section.)(ZK L: Got it.)

To tackle this correlation essence, the introduction of optimal linear experiment design becomes a significant framework (Borkowski and Pukelsheim 1994, Chaloner and Verdinelli 1995, Soare et al. 2014, Allen-Zhu et al. 2021) and Soare et al. (2014) is the first work taking this nature into account. They studied the fixed-confidence setting and introduced the definition of $G$-optimal allocation strategy and $\mathcal{XY}$-optimal strategy. ~~The process of finding the $G$-optimal confidence set in the hyperspace can be interpreted as forming the Minimum Volume Enclosing Ellipsoid (MVEE), a dual process of the $G$-optimal design (Silvey and Sibson 1972, Silvey and Sibson 2016). However, the $G$-optimal allocation can be understood as estimating the linear structure $\boldsymbol{\theta} \subset R^d$ ($d$ is the character dimension of the feature) uniformly well over all the arms, as a result of which, the bound can be rather loose since it does not utilize gaps between different arms efficiently. To solve this problem, Soare et al. (2014) also defines an alternative static allocation algorithm based on the transductive experimental design (Yu et al. 2006), called $\mathcal{XY}$-static allocation. The utilization of differences between different arms allows the $\mathcal{XY}$-static to improve the estimate $\boldsymbol{\theta}$ along the dimensions which can provide more information in discriminating against sub-optimal arms.~~

However, the above ~~two~~ static algorithms fix all allocations before observing any reward and cannot act adaptively in the face of historical data, making their performance quite sub-optimal compared to the oracle allocation derived from the confidence set of $\mathcal{XY}$-static setting. Thus typically, it is necessary to pull arms considering the past observed data. The challenge in formulating an adaptive strategy lies in the fact that a confidence bound for statically selected arms may not always be applicable when arms are chosen adaptively. Specifically, the confidence bound for an adaptive strategy, as proposed by Abbasi-Yadkori et al. (2011), is looser than the bound for a static strategy derived from Azuma's inequality (Azuma 1967) by a factor of $\sqrt{d}$ in certain cases. Moreover, Lattimore and Szepesvári (2020) elucidates that it is impossible to get rid of the additional dependency on $d$ if the algorithm tries to acquire a confidence set for the whole parameter vector or achieve the sequential design.

Continuing within the framework of fixed-confidence setting, several phase-based algorithms have been proposed in the face of this dimension dependency. One such algorithm is a semi-adaptive algorithm called the $\mathcal{XY}$-adaptive allocation (Soare et al. 2014). Subsequently, fully adaptive algorithms were developed, the representatives of which are Linear Gap-based Exploration (LinGapE) algorithm (Xu et al. 2018), a linear variant of UGapE algorithm (Gabillon et al. 2012), and Adaptive Linear Best Arm (ALBA) algorithm proposed by Tao et al. (2018). ~~While the LinGapE algorithm's sample complexity is generally not instance-optimal and may scale with the total number of arms $K$, Tao et al. (2018) took a different path of constructing the new estimators instead of ordinary least squares. This led to an algorithm achieving the well-known sum-of-inverse-gaps sample complexity observed in standard bandits. However, it is not optimal for general linear bandits. In another development, Fiez et al. (2019) gave a phase-based elimination algorithm named Randomized Adaptive Gap Elimination (RAGE) achieving the ideal information-theoretic sample complexity but with minimax oracle access and an additional rounding~~

~~operation. From a performance standpoint, it remains unclear how and to what accuracy this optimization problem should be practically solved. Addressing this problem, Zaki et al. (2020) provided the Phased Elimination Linear Exploration Game (PELEG) algorithm to achieve rapid shrinkage of the confidence ellipsoid along the most confusing directions (i.e., close to but not optimal) by framing the problem as a two-player zero-sum game.~~

Within the framework of the fixed-budget setting is another stream of research development. Following the inception of the first fixed-budget scheme, the BayesGap algorithm proposed by (Hoffman et al. 2014), several subsequent algorithms have been introduced by various researchers(Hoffman et al. 2014, Katz-Samuels et al. 2020, Yang and Tan 2022, Azizi et al. 2021). Besides, it needs to be noted that to the best of our knowledge, there is currently no matching upper and lower bound for fixed-budget BAI in any setting, as pointed out by Carpentier and Locatelli (2016). There still remain two important open problems in the context of fixed-budget BAI, which have been systematically outlined by Qin (2022): firstly, the exploration of matching upper bounds and lower bounds in a similar fashion as in the fixed-confidence setting; and secondly, the study into whether there exists an algorithm other than uniform sampling itself, that consistently performs no worse than uniform sampling in the fixed-budget setting. (ZK L: We may remove this paragraph since it is mainly about the FB setting.)

The Peace algorithm (Katz-Samuels et al. 2020) is mainly a fixed-budget BAI algorithm adapted from a fixed-budget strategy. However, it poses a challenge due to its intractability stemming from the lack of a closed form for the Gaussian width, making it computationally expensive to minimize. A cumulative regret minimization algorithm based on the $G$-optimal design was introduced for linear bandit with a finite number of arms in the book of Lattimore and Szepesvári (2020), which can be easily adjusted to achieve pure exploration performance. Another noteworthy approach is the Optimal Design-based Linear Best Arm Identification (OD-LinBAI), which also utilizes $G$-optimal design within a sequential elimination framework for the fixed-budget BAI problem. Additionally, the Generalized Successive Elimination (GSE) algorithm is the first paper considering both the linear model and the generalized linear model (GLM), demonstrating competitive error guarantees compared to prior works and performing at least as well empirically. (ZK L: We may remove this paragraph since it is mainly about the FB setting.)

Unfortunately, the strong guarantees provided by the aforementioned algorithms are applicable only when the expected rewards are in perfect linear relationship with the given features, a property that is often violated in real-world applications. In fact, when using linear models with real data, one inevitably faces the problem of misspecification, i.e., the situation in which the data deviates from linearity. A misspecified linear bandit model is often described as a linear bandit model with an additive term to encode deviation from linearity. Due to its importance, this problem has recently gained increasing attention in the bandit community for cumulative regret minimization (Ghosh et al. 2017, Foster et al. 2020, Pacchiano et al. 2020, Takemura et al. 2021, Zhang et al. 2023). However, it was rarely addressed in the context of pure

exploration. To the best of our knowledge, the most related work is contributed by Réda et al. (2021). In this work, they first derived a tractable lower bound on the sample complexity of any $\delta$-*correct* algorithm for the general Top-$K$ identification problem with a misspecified linear model. They also derived an upper bound to its sample complexity which confirms this adaptivity and that matches the lower bound when $\delta \to 0$.(ZK L: This paragraph is related to the misspecification and can be removed.)

**1.2.4. Classical Ranking and Selection.** (ZK L: To be finished (or removed).) (CH: I don't think we need it in the 12-page version)

Hong et al. (2021)

## 2. Problem Formulation

Throughout the paper, we let $[N]$ represent the set of integers up to $N$, i.e., $[N] = \{1, \dots, N\}$. We also use boldface variables to represent vectors and matrices. We use $\langle \cdot, \cdot \rangle$ to define the inner product of two vectors. We also let the matrix norm $\|\boldsymbol{x}\|_{\boldsymbol{A}}$ be defined as $\|\boldsymbol{x}\|_{\boldsymbol{A}} = \sqrt{\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}$, where the definite matrix $\boldsymbol{A}$ serves the function of weighting and scaling the norm.

### 2.1. Linear Bandits

We consider the linear bandit model with $K$ arms. The reward distribution of each arm $i \in [K]$, assumed to be independent and identically distributed, has a mean value $\mu_i$, which remains fixed but unknown and can only be estimated through the bandit feedback in the sense that only rewards from the selected arms are observable. We let $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_K)$ be the vector containing all arms. Without loss of generality, we assume that $\mu_1 > \mu_2 \geq \ldots \geq \mu_K$ and thus arm 1 is optimal. In the linear context, the mean value depends on an unknown parameter $\boldsymbol{\theta}$ that is shared among all arms. Let $\boldsymbol{a}_i \in \mathbb{R}^d$ be the feature vector of arm $i$, and we assume that $\max_{i \in [K]} \|\boldsymbol{a}_i\|_2 \leq L_1$, where $\|\cdot\|_2$ is the $\ell_2$-norm. In each time $t$, we pull one arm from $\mathcal{A} \subset \mathbb{R}^d$ with $|\mathcal{A}| = K$. The reward function is characterized by

$$X_t = \boldsymbol{a}_{A_t}^\top \boldsymbol{\theta} + \eta_t, \tag{1}$$

where $X_t$ is the observed reward, $\boldsymbol{a}_{A_t}$ is feature vector of arm $A_t$ sampled at time $t$, and $\eta_t$ represents a noise variable. We assume the noise distribution is conditionally $R$-sub-Gaussian, that is, for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[e^{\lambda \eta_t} | \boldsymbol{a}_{A_1}, \dots, \boldsymbol{a}_{A_{t-1}}, \eta_1, \dots, \eta_{t-1}\right] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right), \tag{2}$$

which requires the noise distribution to have zero expectation and $R$ or less variance.

## 2.2. All $\varepsilon$-Best Arms Identification Pure Exploration Tasks

~~We aim to identify all $\varepsilon$-best arms. The concept of all $\varepsilon$-best arms is particularly beneficial in scenarios where a wide exploration of nearly optimal options can lead to significant breakthroughs or efficiencies. For example, in drug discovery, drug development is costly and risky. By advancing multiple promising candidates, the strategy distributes risk and increases the chances that at least one candidate will successfully progress through clinical trials and regulatory approval. In addition, other compounds slightly less effective in preliminary tests might exhibit better properties later, such as fewer side effects, easier manufacturing, or lower resistance rates.~~

~~For some other examples, like e-commerce or streaming services, all $\varepsilon$-best (e.g., products, movies, or songs) can enhance user satisfaction. This approach allows for a diversity of recommendations that cater to slightly varying tastes or preferences, rather than recommending the same top items to everyone. In manufacturing, different processes might yield products of nearly identical quality. Identifying many process settings can help in optimizing production lines for cost, speed, and quality, providing flexibility in response to supply chain or demand changes. In investment portfolio management, identifying all $\varepsilon$-best investment opportunities (e.g., stocks, bonds, or other assets) can help in diversifying investments to manage risk and optimize returns. This strategy ensures that portfolio managers are not overly reliant on a small number of assets that appear optimal under current conditions but might be volatile or risky.~~ (CH: We may move these two paragraphs to Introduction. )(ZK L: Moved.)

DEFINITION 1. ($\varepsilon$-Best): For a given $\varepsilon > 0$, arm $i$ is *$\varepsilon$-best* if $\mu_i \geq \mu_1 - \varepsilon$.

The formal description of $\varepsilon$-best is given in Definition 1. We note that we define $\varepsilon$-best in an additive way. There is also a multiplicative $\varepsilon$-best, defined as $\mu_i \geq (1-\varepsilon)\mu_1$. Our paper focuses on the additive setting and the analysis for the multiplicative is similar. We define the set containing all $\varepsilon$-best arms when with means $\boldsymbol{\mu}$ as $G_\varepsilon(\boldsymbol{\mu}) \triangleq \{i : \mu_i \geq \mu_1 - \varepsilon\}$. Throughout the paper, we use $G_\varepsilon$ as a shorthand when referring to the true mean.

## 2.3. Probably Approximately Correct (PAC) with High Confidence

In this study, we aim to design an algorithm that identifies all $\varepsilon$-best arms with high confidence, while using a minimal sampling budget. As an overview, our algorithm comprises a sampling rule, a stopping rule, and a decision rule.

At each time step $t$, we determine whether to halt based on a stopping rule, denoted as $\tau_\delta$. If we stop at $t = \tau_\delta$, we provide an estimation $\widehat{\mathcal{I}}_{\tau_\delta}$ of the answer $\mathcal{I}(\boldsymbol{\mu})$ using a decision rule. In our context, $\mathcal{I}(\boldsymbol{\mu}) = G_\varepsilon(\boldsymbol{\mu})$. Alternatively, if we do not stop, we pull an arm according to a sampling rule and observe the random rewards. Specifically, we focus on such algorithms that are probably approximately correct with high confidence, referred to as *$\delta$-PAC*.

DEFINITION 2. An algorithm is said to be $\delta$-*PAC* if the probability of correct decision upon stopping time is with probability at least $1-\delta$ for any problem instance $\boldsymbol{\mu} \in M$, where $M$ is the set of models of interest.

$$\mathbb{P}_{\boldsymbol{\mu}}\left(\tau_\delta < \infty,\ \widehat{\mathcal{I}}_{\tau_\delta} = \mathcal{I}(\boldsymbol{\mu})\right) \geq 1-\delta, \qquad \forall \boldsymbol{\mu} \in M. \tag{3}$$

Upon stopping, $\delta$-*PAC* algorithms return all $\varepsilon$-best arms with high confidence. The primary goal is to design such an algorithm that stops as fast as possible. In the context of drug discovery, that would mean we find good drug candidates with the minimum amount of time and number of experiments. This could save both time and money in the initial stage of drug candidate screening.

## 2.4. Optimal Design

**2.4.1. Least Squares Estimators.** Let $A_1, A_2, \ldots, A_n$ be the sequence of arms pulled by the decision-maker and $X_1, X_2, \ldots, X_n$ be the corresponding noisy rewards. Suppose that the corresponding arm feature vectors $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n$ span space $\mathbb{R}^d$, then the classical ordinary least squares (OLS) estimator of parameter $\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\theta}}_t = \boldsymbol{V}_t^{-1} \sum_{t=1}^{n} \boldsymbol{a}_{A_t} X_t, \tag{4}$$

where the information matrix $\boldsymbol{V}_t = \sum_{t=1}^{n} \boldsymbol{a}_{A_t} \boldsymbol{a}_{A_t}^\top \in \mathbb{R}^{d\times d}$ is assumed to be invertible. By applying the properties of sub-Gaussian random variables, the confidence bound for the OLS estimator, which is in the form of $B_{t,\delta} = \|\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}\|_{\boldsymbol{V}_t}$, can be derived as in Proposition 1. This proposition is established with the fixed sampling sequence, where the rewards of samples are not observed and do not have any influence on the sampling policy.

PROPOSITION 1. *In the linear bandit, let the noise variable be bounded by* $[-c, c]$ *for* $c > 0$. *Then, for any fixed sampling policy, the statement*

$$\left|\boldsymbol{x}^\top \left(\hat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}\right)\right| \leq \|x\|_{\boldsymbol{V}_t^{-1}} B_{t,\delta} \tag{5}$$

*holds for given* $x \in \mathbb{R}^d$ *and all* $t > 0$ *with probability at least* $1-\delta$, *where the anytime confidence bound of empirical estimate of parameter* $\boldsymbol{\theta}$ *is defined as*

$$B_{t,\delta} = 2c\sqrt{2\log\left(\frac{6t^2K}{\delta\pi^2}\right)}. \tag{6}$$

Furthermore, by exploiting the martingale-based method, the adaptive confidence bound for the OLS estimator can be defined as in Proposition 2. This proposition is proposed with an adaptive sampling sequence, where the random rewards of samples are observed and thus bring randomness to the result and design of the sampling policy.

PROPOSITION 2. *(Abbasi-Yadkori et al. 2011) (Theorem 2). In the linear bandit with conditionally $R$-sub-Gaussian noise, if the $l_2$-norm of parameter $\boldsymbol{\theta}$ is less than constant $C_5$(ZK L: We got several other constant variables through this paper and we may make some final adjustments to the notation together.) and the sampling policy depends on past observations, then statement*

$$\left|\boldsymbol{x}^\top\left(\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}\right)\right| \le \|x\|_{\boldsymbol{V}_t^{-1}} B_{t,\delta} \tag{7}$$

*holds for given $x\in\mathbb{R}^d$ and all $t>0$ with probability at least $1-\delta$, where the anytime confidence bound of empirical estimate of parameter $\boldsymbol{\theta}$ is defined as*

$$B_{t,\delta} = R\sqrt{2\log\frac{\det\left(\boldsymbol{V}_t^{-1}\right)^{\frac{1}{2}}\det(\lambda\boldsymbol{I})^{-\frac{1}{2}}}{\delta}} + \lambda^{\frac{1}{2}}C_5, \tag{8}$$

*where $\lambda$ is the regularization coefficient.*

The confidence interval in Proposition 1 and Proposition 2 represents the connection between the arm allocation policy in linear bandits and experimental design theory (Pukelsheim 2006).

**2.4.2. Projection** For any linear bandit instance, if the corresponding arm vectors do not span $\mathbb{R}^d$, the agent can work with a set of dimensionality-reduced arm vectors.~~As LinFACTE does elimination from round to round, it is very likely that the remaining arms in $\mathcal{A}_t$ cannot span the whole space $\mathbb{R}^d$ and make design matrix $\boldsymbol{V}_\tau$ in some rounds not invertible. In fact, in our application on drug discovery, the original arm matrix has a lower rank than the whole space.~~

To solve this problem, we project all the arms into the subspace spanned by $\mathcal{A}$. To be specific, based on the work of Yang and Tan (2022), we define the dimension of subspace with ~~$d_\tau$ in round $t$~~ $d'$, then we can find an orthonormal basis of the subspace spanned by $\mathcal{A}$, denoted by $B\in\mathbb{R}^{d\times d'}$, where the projected arm is simply $\boldsymbol{a}'=B^\top\boldsymbol{a}$. This is simply because $BB^\top$ is a projection matrix, and $\langle\boldsymbol{\theta},\boldsymbol{a}\rangle=\langle\boldsymbol{\theta},BB^\top\boldsymbol{a}\rangle=\langle B^\top\boldsymbol{\theta},B^\top\boldsymbol{a}\rangle$. (ZK L: Moved from the section of algorithm design. Rewrite part of the content since we haven't introduced our algorithm here.)

**2.4.3. G-Optimal Design.** Different from the stochastic bandit, where the mean values can only be obtained by pulling the arms, in the linear bandit setting, the mean values can be obtained by making good estimates of the parameters $\boldsymbol{\theta}$. The G-optimal design minimizes the maximum variance of the predicted responses. In other words, this type of design aims to make the predictions as precise as possible, ensuring that the maximum prediction variance over all possible values of the explanatory variables is minimized. Formally, the $G$-optimal design problem aims at finding a probability distribution $\pi\in\mathcal{P}(\mathcal{A})$, defined as $\pi:\mathcal{A}\to[0,1]$ on $\mathcal{A}$ with $\sum_{\boldsymbol{a}\in\mathcal{A}}\pi(\boldsymbol{a})=1$, that minimizes

$$g(\pi)=\max_{\boldsymbol{a}\in\mathcal{A}}\|\boldsymbol{a}\|^2_{\boldsymbol{V}(\pi)^{-1}}, \tag{9}$$

where $\boldsymbol{V}(\pi) = \sum_{\boldsymbol{a}\in\mathcal{A}} \pi(\boldsymbol{a})\boldsymbol{a}\boldsymbol{a}_i^\top$ can be seen as the average version of the information matrix $\boldsymbol{V}_t$. Besides, the following Theorem 1 states the equivalence between $G$-optimal design and $D$-optimal design and the existence of a $G$-optimal design with a small core set. A design that maximizes function $f(\pi)$ is known as a $D$-optimal design. The D-optimal design aims to maximize the determinant of the information matrix, the Fisher Information Matrix, of the parameters in the model. Maximizing this determinant leads to minimizing the volume of the confidence ellipsoid around the parameter estimates, which effectively means achieving the smallest possible joint confidence region for the model parameters.

THEOREM 1. *(Kiefer and Wolfowitz 1960). If the arm vectors $\boldsymbol{a} \in \mathcal{A}$ span $\mathbb{R}^d$, for $\pi \in \mathcal{P}(\mathcal{A})$, the following statements are equivalent:*

1. *$\pi^*$ is a minimizer of function $g$.*
2. *$\pi^*$ is a maximizer of function $f(\pi) = \log\det \boldsymbol{V}(\pi)$.*
3. *$g(\pi^*) = d$.*

*Furthermore, there exists a minimizer $\pi^*$ of $g$ such that $|Supp(\pi^*)| \le d(d+1)/2$.*

(CH: Do we need this theorem? Or we can just use a sentence to summarize it? )(ZK L: It serves as the basis of algorithm design. It is used in Section EC.2.0.12 and the proof of Lemma EC.4.3.)

**2.4.4. $\mathcal{XY}$-Optimal Design.** Despite being classical and optimal in different applications, $G$-optimal design is minimizing a rather loose upper bound on the quantity used to test the gap between the arms (Soare et al. 2014). The optimization problem in equation (9) is equivalent to finding the solution that minimizes the maximum prediction error among mean values of all arms and providing the confidence interval of mean value estimation as tight as possible. However, in the context of the pure exploration problem, we are more interested in comparing the relative magnitudes of mean values among different arms, leading to a requirement for the precise estimation of gaps between different arms. Thus, we define an alternative optimal design that targets the gap quantity more directly by changing the term in the norm of the equation (9). Through the optimization problem described in equation (10), we can achieve a quick and accurate estimation of gaps directly, minimizing the maximum prediction error among all the gaps.

Take $\mathcal{S} \subset \mathcal{A}$ to be a subset of the active set. $\mathcal{Y}(\mathcal{S}) \triangleq \{\boldsymbol{a} - \boldsymbol{a}' : \forall \boldsymbol{a}, \boldsymbol{a}' \in \mathcal{S}, \boldsymbol{a} \neq \boldsymbol{a}'\}$ is defined as the directions obtained from the differences between each pair of arms. We define the $\mathcal{XY}$-optimal design as minimizing $g_{\mathcal{XY}}(\pi) = \max_{\boldsymbol{y}\in\mathcal{Y}(\mathcal{A})} \|\boldsymbol{y}\|^2_{\boldsymbol{V}(\pi)^{-1}}$, given by

$$\arg\min_{\pi\in\mathcal{P}(\mathcal{A})} g_{\mathcal{XY}}(\pi) \qquad \text{subject to } \sum_{\boldsymbol{a}\in\mathcal{A}} \pi(\boldsymbol{a}) = 1. \tag{10}$$

## 2.5. Further Notation

Before proceeding, we introduce several critical quantities that we will keep track of throughout the analysis. Define $\alpha_\varepsilon \triangleq \min_{i\in G_\varepsilon} \mu_i - (\mu_1 - \varepsilon)$ as the distance from the smallest additive $\varepsilon$-best arm, ~~denoted~~

~~$\mu_\alpha$,~~ to the threshold $\mu_1 - \varepsilon$. Additionally, if $G_\varepsilon^c$, the complement of $G_\varepsilon$, is not empty, we define $\beta_\varepsilon = \min_{i \in G_\varepsilon^c} (\mu_1 - \varepsilon) - \mu_i$ as the distance of the largest arm that is not additive $\varepsilon$-best, ~~denoted $\mu_\beta$,~~ to the threshold.

(ZK L: Move it to the part of $\mathcal{XY}$-Optimal Design.) (ZK L: To be supplemented.)

For two probability measures $P$ and $Q$ over a common measurable space, if $P$ is absolutely continuous with respect to $Q$, the Kullback-Leibler divergence between $P$ and $Q$ is

$$\mathrm{KL}(P, Q) = \begin{cases} \int \log\left(\frac{dP}{dQ}\right) dP, & \text{if} Q \ll P; \\ \infty, & \text{otherwise}, \end{cases} \tag{11}$$

where $dP/dQ$ is the Radon-Nikodym derivative of $P$ with respect to $Q$. (CH: This is what I copied from the Russo and Van Roy paper. Please refine it to match our notation. )(ZK L: Refined.)

## 3. Algorithm

### 3.1. General Algorithm Framework for the Pure Exploration

An arm $A_t$ is selected at each time step $t$, and a reward $X_t$ is drawn from the unknown distribution. Let $\mathcal{F}_t = \sigma(A_1, X_1, A_2, X_2, \ldots, A_t, X_t)$ be the $\sigma$-algebra generated by the observations up to time $t$, then any pure exploration algorithm can then be defined as a triple $\mathcal{H} = (A_t, \tau_\delta, \hat{a}_\tau)$ where:

- a sampling rule $A_t$, where $A_t$ is $\mathcal{F}_{t-1}$-measurable;
- a stopping rule $\tau_\delta$, which is a stopping time with respect to $\mathcal{F}_t$;
- a $\mathcal{F}_t$-measurable decision rule $\hat{a}_\tau$.

In such a setting, the sequence of events unfolds as follows. Based on the observed data, the decision-maker assesses whether to terminate the experiment according to a predefined stopping rule. If the experiment is concluded, then the decision-maker recommends the optimal estimate of the answer utilizing a decision rule. Conversely, if the experiment continues, a sampling rule determines the next alternative to be sampled.

### 3.2. Linear Fast Arm Classification with Threshold Estimation

In this section, we propose an algorithm called **LinFACTE** (Linear Fast Arm Classification with Threshold Estimation) algorithm for All $\varepsilon$-Best Arms Identification problem in the linear context.

The algorithm is shown in Algorithm 1. At a high level, LinFACTE is a phase-based classification algorithm, developing as round $r$ progresses and maintains sets $G_r$ and $B_r$ of arms classified to be good (empirically $\varepsilon$-best) or bad (not empirically $\varepsilon$-best) (**A**rm **C**lassification). The algorithm stops when all the classifications have been finished and returns a set of arms based on the decision rule.

**3.2.1. Sampling Rule.** To minimize the number of sample budgets, the decision-maker has to select arms that can provide as much information as possible about the mean value of the arms themselves or the gaps between arms to complete the classification quickly and with high confidence. Different from the stochastic bandit, where the mean values can only be obtained by pulling the specific arms, in the linear bandit setting, the mean values can be inferred by the estimated parameters $\boldsymbol{\theta}$. Thus, we resort to $G$-optimal and $\mathcal{XY}$-optimal design criteria that we discussed in Section 2.4.

In linear bandit (**Lin**ear), arms are pulled based on either $G$-optimal sampling policy (9) or $\mathcal{XY}$-optimal sampling policy (10), to achieve an increasingly accurate estimate of the unknown threshold $\boldsymbol{\mu}_1 - \varepsilon$ (**T**hreshold **E**stimate). This is critical because the estimated threshold is the standard for us to judge whether an arm is $\varepsilon$-best. To do so, LinFACTE refines an estimate of true parameter $\boldsymbol{\theta}$ and maintains an anytime confidence width $C_{\delta/K}(r)$ subsequently, such that for each arm's empirical mean value $\hat{\mu}_i$, we have $\mathbb{P}\left(\bigcap_{i\in\mathcal{A}_I(r)}\bigcap_{r\in\mathbb{N}}|\hat{\mu}_i(r)-\mu_i|\le C_{\delta/K}(r)\right)\ge 1-\delta$. We take $C_{\delta/K}(r)=\varepsilon_r$, which is halved with each iteration of the rounds and has a simple form.

The anytime confidence width $C_{\delta/K}(r)$ is maintained by the design of the sample budget in each round. In LinFACTE, the first optimal design-based budget allocation policy depends on the $G$-optimal design with a simple rounding-up operation, given by

$$\begin{cases} T_r(\boldsymbol{a}) = \left\lceil \frac{2d\pi_r(\boldsymbol{a})}{\varepsilon_r^2}\log\left(\frac{Kr(r+1)}{\delta}\right)\right\rceil. \\ T_r = \sum_{\boldsymbol{a}\in\mathcal{A}} T_r(\boldsymbol{a}) \end{cases} \tag{12}$$

where $\pi_r$ is the distribution of remaining arms as the active set $\mathcal{A}$ gradually shrinks along with round $r$.

The second budget allocation policy, which is also an optimal design, is designed based on the $\mathcal{XY}$-optimal design and a rounding operation with a $(1+\epsilon)$ approximation (Allen-Zhu et al. 2017). Let $r(\varepsilon)$ be the error due to the rounding procedure, we have

$$\begin{cases} T_r = \max\left\{\left\lceil \frac{2g_{\mathcal{XY}}(\mathcal{Y}(\mathcal{A}))(1+\epsilon)}{\varepsilon_r^2}\log\left(\frac{2K(K-1)r(r+1)}{\delta}\right)\right\rceil, r(\epsilon)\right\}. \\ T_r(\boldsymbol{a}) = \text{ROUND}(\pi_r, T_r) \end{cases} \tag{13}$$

**3.2.2. Stopping Rule and Decision Rule.** The LinFACTE stops once all the arms are assigned into one of the good or bad groups with enough confidence, i.e., $G_r \cup B_r = [K]$. In the stopping round, the set with good arms $G_r$ is returned as the final output of the algorithm.

**Algorithm 1** LinFACTE

1: **Input:** $\varepsilon, \delta$, Bandit Instance $v$, slackness $\gamma \ge 0$.
2: Let $G_0 = \emptyset$ be the set of ~~good~~ all $\varepsilon$-best arms and $B_0 = \emptyset$ the set of ~~bad~~ arms this are not $\varepsilon$-best.
3: Let ~~unchangeable action set~~ $\mathcal{A}$ be the active set (the set of uneliminated arms), ~~$\mathcal{A}_I = [K]$~~ $\mathcal{A}_I$ be the set of indices corresponding to the remaining arms ~~and the following indices $i \in \mathcal{A}_I$ is in serious correspondence with each arm $\boldsymbol{a}_i \in \mathcal{A}$~~. (CH: Please revise this.)(ZK L: Got it.)
4: ~~Let $C_{\delta/K}(r)$ be an anytime $\delta$-correct confidence width in round $r$.~~

**for** $r = 1, 2, \ldots$ **do**

Let $\varepsilon_r = 2^{-r}$ and initialize $G_r = G_{r-1}$, $B_r = B_{r-1}$.

**// Projection**

Project $\mathcal{A}_I$ to dimension $d_r$, which is the dimension of subspace that ~~$\mathcal{A}_I$~~ $\mathcal{A}$ spans.

**// Sampling**

(*G-optimal design*). (CH: Please change the following all to this style)(ZK L: Corrected.) Find the $G$-optimal design $\pi_r \in \mathcal{P}(\mathcal{A})$ with $\text{Supp}(\pi_r) \le d(d+1)/2$ according to equation (9). ~~$\log\det V(\pi_r)$ subject to $\sum_{\boldsymbol{a}\in\mathcal{A}} \pi_r(\boldsymbol{a}) = 1$~~

($\mathcal{XY}$-*optimal design*). Find the $\mathcal{XY}$-optimal design $\pi_r \in \mathcal{P}(\mathcal{A})$ according to equation (10). ~~$\arg\min_{\pi_r\in\mathcal{P}(\mathcal{A})} g_{\mathcal{XY}}(\pi)$ subject to $\sum_{\boldsymbol{a}\in\mathcal{A}} \pi_r(\boldsymbol{a}) = 1$~~

**for all** $i \in \mathcal{A}_I$ **do**

(*G-optimal design*). With a simple round-up operation, compute the sample budgets $T_r(\boldsymbol{a})$ in round $r$ based on equation (12) ~~$T_r(\boldsymbol{a}) = \left\lceil \frac{2d\pi_r(\boldsymbol{a})}{\varepsilon_r^2} \log\left(\frac{Kr(r+1)}{\delta}\right)\right\rceil$ and $T_r = \sum_{\boldsymbol{a}\in\mathcal{A}} T_r(\boldsymbol{a})$~~

($\mathcal{XY}$-*optimal design*). With a rounding procedure of $(1+\epsilon)$ approximation, compute the sample budgets $T_r(\boldsymbol{a})$ in round $r$ based on equation (13) ~~$T_r = \max\left\{\left\lceil \frac{2g_{\mathcal{XY}}(\mathcal{Y}(\mathcal{A}))(1+\epsilon)}{\varepsilon_r^2} \log\left(\frac{2K(K-1)r(r+1)}{\delta}\right)\right\rceil, r(\epsilon)\right\}$ and $T_r(\boldsymbol{a}) = \text{ROUND}(\pi_r, T_r)$~~

Sample each action $\boldsymbol{a} \in \mathcal{A}$ exactly $T_r(\boldsymbol{a})$ times and calculate the empirical estimate with standard Least Squares Estimation (LSE)~~:~~ by equation (4). ~~$\hat{\boldsymbol{\theta}}_r = \boldsymbol{V}_r^{-1} \sum_{\ell=1}^{T_r} \boldsymbol{a}_{A_\ell} X_\ell$ with $\boldsymbol{V}_r = \sum_{\boldsymbol{a}\in\mathcal{A}} T_r(\boldsymbol{a})\boldsymbol{a}\boldsymbol{a}^\top$~~

**// Arm Filter: find good arms in $G_\varepsilon$ and bad arms in $G_\varepsilon^c$**

Let $U_r = \max_{i\in\mathcal{A}_I} \hat{\mu}_i + C_{\delta/K}(r) - \varepsilon$.

Let $L_r = \max_{i\in\mathcal{A}_I} \hat{\mu}_i - C_{\delta/K}(r) - \varepsilon$.

**for all** $i \in \mathcal{A}_I$ **do**

**if** $\hat{\mu}_i - C_{\delta/K}(r) > U_r$ **then**

Add $i$ to $G_r$.

**if** $\hat{\mu}_i + C_{\delta/K}(r) < L_r$ **then**

Add $i$ to $B_r$ and remove $i$ from $\mathcal{A}_I$.

**if** $i \in G_r$ and $\hat{\mu}_i + C_{\delta/K}(r) \le \max_{j\in\mathcal{A}_I} \hat{\mu}_i - C_{\delta/K}(r)$ **then**

Remove $i$ from $\mathcal{A}_I$.

**if** $G_r \cup B_r = [K]$ **then**

**Output:** the set $G_r$. // Stopping condition for returning $G_\varepsilon$ exactly.

**if** $U_r - L_r \le \gamma/2$ **then**

**Output:** the set $\mathcal{A}_I \cup G_r$ // Stopping condition for $\gamma > 0$.

(CH: We may want to shorten the algorithm)(ZK L: Adjusted and shortened.)

## 4. Main Theoretical Results

Within the framework of All $\varepsilon$-Best Arms Identification in the linear bandits, our main theoretical results concern the lower bound representing the problem complexity and upper bounds that show the optimality of the proposed LinFACTE algorithm. In this section, we first introduce the current lower bound results for the All $\varepsilon$-Best Arms Identification problem in the stochastic setting. We extend the lower bound from the stochastic setting to the linear setting, which is the first result of this problem to the best of our knowledge. Then, we provide three upper bounds for the expected sample complexity to show the optimality of the LinFACTE algorithm. The upper bound based on $\mathcal{XY}$-optimal sampling policy is proved to be instance-optimal up to a logarithmic term.

Besides, to facilitate our theoretical analysis, we will provide a comprehensive model for the pure exploration problem and some graphical insights to assist the understanding of the problem complexity, the stopping rule, and the framework of the pure exploration problem fundamentally, which function as the basis of our theoretical guarantees derivation.

### 4.1. General Pure Exploration Model

The decision-maker aims to address a query (CH: please briefly explain what is a query and what is an answer here)(ZK L: Supplemented, but I'm not sure if I am doing right in answering this comment.) related to the mean parameters $\boldsymbol{\mu}$, achieved through adaptive allocation of the sampling budget to the set of arms. This query task involves finding a specific group of arms based on various criteria and we are trying to find the right answer with high probability. ~~We assume that the query question has a unique answer for each bandit instance under $\boldsymbol{\mu}$, denoted as $\mathcal{I}=\mathcal{I}(\boldsymbol{\mu})$.~~ Recall that $\mathcal{I}=\mathcal{I}(\boldsymbol{\mu})$ is the correct answer, the set with all $\varepsilon$-best arms. Let $\widetilde{M}$ denote the set of parameters associated with a unique answer. Let $\Xi$ be the set of all possible answers, and for each $\mathcal{I}' \in \Xi$, define $M_{\mathcal{I}'} \triangleq \left\{\boldsymbol{\vartheta} \in \widetilde{M} : \mathcal{I}(\boldsymbol{\vartheta}) = \mathcal{I}'\right\}$ as the set of parameters with $\mathcal{I}'$ as the correct answer. The parameter space $M \triangleq \cup_{\mathcal{I}' \in \Xi} M_{\mathcal{I}'}$ is the focus of the problem.

~~Based on Definition 2, we denote by~~ Recall that an algorithm is defined as a triple $\mathcal{H} = (A_t,\, \tau_\delta,\, \hat{a}_\tau)$ ~~the class of algorithms that are $\delta$-PAC for any $\delta > 0$~~. The algorithm's sample complexity is quantified by the number of samples, denoted as $\tau_\delta$, at the point of termination. The objective is to formulate algorithms that minimize the expected sample complexity $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]$ across the set $\mathcal{H}$. As stated in Garivier and Kaufmann (2016), the problem complexity of a problem instance $\boldsymbol{\mu}$ can be defined as

$$\kappa(\boldsymbol{\mu}) \triangleq \inf_{Algo \in \mathcal{H}} \limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)}. \tag{14}$$

This instance-dependent complexity indicates the smallest possible constant such that the expected sample complexity $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]$ scales in alignment with $\log(1/\delta)$, as $\delta$ decrease to 0. Another non-asymptotic version can be acquired from the inequality $\mathrm{d}(\delta, 1-\delta) \geq \log(1/2.4\delta)$ that holds for all $\delta \in (0,1)$, as given

in Kaufmann et al. (2016). Here $\mathrm{d}(\cdot,\cdot)$ is the Kullback-Leibler (KL) divergence of two distributions for a binary reward, which is defined as $\mathrm{d}(x,\,y) \triangleq x\log(x/y) + (1-x)\log((1-x)/(1-y))$.

The problem complexity $\kappa(\boldsymbol{\mu})$ is subject to an information-theoretic lower bound. This lower bound can be expressed as the optimal solution of an allocation problem, which we show in Proposition 4. To build this framework, we will first introduce two important concepts: culprits and $C_x$ function.

**4.1.1. Culprits and Alternative Sets.** A culprit denotes a decision that can lead to suboptimal outcomes. Mathematically, a culprit specifies the set of arms with the smallest cardinality that can cause the wrong answer in the easiest way. ~~Identifying culprits is crucial as it allows for the characterization of the problem complexity and the design of algorithms.~~ We define $\mathcal{X}(\boldsymbol{\mu})$ as the set of culprits when the true mean vector is $\boldsymbol{\mu}$. These culprits cause the deviation of the correct answer of $\boldsymbol{\mu}$ from $\mathcal{I}(\boldsymbol{\mu})$. The set of culprits differs for each exploration task, and identifying culprits is crucial as it allows for the characterization of the problem complexity and the design of algorithms.

As an example, in the ~~context of BAI~~ task of identifying the single best arm, ~~we define~~ $\mathcal{X}(\boldsymbol{\mu}) = [K]\backslash\{I^*(\boldsymbol{\mu})\}$, which is the set of all arms without the best arm. For each culprit $x \in \mathcal{X}(\boldsymbol{\mu})$, if arm $x$ is a better arm under another mean vector $\boldsymbol{\vartheta}$ than the best arm $I^*(\boldsymbol{\vartheta})$, then $\boldsymbol{\vartheta}$ is the parameter that can bring a wrong answer caused by $x$. Associated with each culprit $x \in \mathcal{X}(\boldsymbol{\mu})$ is an alternative set, which is the set of parameters that causes the identification of the wrong answer. We have $\mathrm{Alt}_x(\boldsymbol{\mu}) = \left\{\boldsymbol{\vartheta} \in M : \vartheta_x \geq \vartheta_{I^*(\boldsymbol{\vartheta})}\right\}$ for $x \in \mathcal{X}(\boldsymbol{\mu})$. (CH: Do we need to define the culprit and Alt in terms of $\theta$?)(ZK L: This paragraph is still an example of BAI. The culprit and Alt in terms of parameter $\boldsymbol{\theta}$ are the specific problems related to the linear bandit, which will be defined when we formally encounter it in the proof of Section EC.1. Besides, I made some further explanation in the proof of Section EC.1.) (CH: Please also use epsilon best as an example to define these terms. )(ZK L: Supplemented.)

Futhermore, $\varepsilon$-Best Arm Identification ($\varepsilon$-BAI) (also known as R&S with probability of good selection guarantee) is another pure exploration task that has been well studied (Jourdan et al. 2024) and is close to our setting. In this context, let $\mathcal{I}(\boldsymbol{\mu}) \triangleq \{i \in [K] : \mu_i \geq \mu_{I^*} - \varepsilon\}$ denote the set of good enough arms, whose suboptimality relative to the best arm (assumed to be unique) $I^* = \arg\max_i \mu_i$ is less than a predefined tolerance of $\varepsilon$. The $\varepsilon$-BAI problem aims to find any arm $i \in I(\mu)$ and stops whenever one arm is returned.

An algorithm that finds the correct answer must distinguish different problem instances in the parameter space $M$. Therefore, for any instance $\boldsymbol{\mu} \in M$, the instance-dependent problem complexity $\kappa(\boldsymbol{\mu})$ is related to the alternative set

$$\cup_{x\in\mathcal{X}(\boldsymbol{\mu})}\mathrm{Alt}_x(\boldsymbol{\mu}) = \mathrm{Alt}(\boldsymbol{\mu}) \triangleq \{\boldsymbol{\vartheta} \in M : \mathcal{I}(\boldsymbol{\vartheta}) \neq \mathcal{I}(\boldsymbol{\mu})\}, \tag{15}$$

which represents the set of parameters that return a solution that is different from the correct solution $\mathcal{I}(\boldsymbol{\mu})$.

**4.1.2. $C_x$ unction.** The process of finding the correct answer (CH: Please specify what process)(ZK L: Specified.) is solved by a sequential hypothesis test using the test of generalized log-likelihood ratio (GLR) (Kaufmann and Koolen 2021). ~~It gives the definition and meaning of $C_x$ function.~~ A GLR statistic is defined for testing a potentially composite null hypothesis $H_0 : (\boldsymbol{\mu} \in \Omega_0)$ versus a potentially composite alternative hypothesis $H_1 : (\boldsymbol{\mu} \in \Omega_1)$ by

$$\text{GLR}_t = \frac{\sup_{\boldsymbol{\lambda} \in \Omega_0 \cup \Omega_1} L(X_1, X_2, \ldots, X_t; \boldsymbol{\lambda})}{\sup_{\boldsymbol{\vartheta} \in \Omega_0} L(X_1, X_2, \ldots, X_t; \boldsymbol{\vartheta})}, \tag{16}$$

where $X_1, X_2, \ldots, X_t$ are observed rewards from arm pulling, and $L(\cdot)$ is the likelihood function with these observed data and some unknown parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\vartheta}$. $\Omega_0$ represents the restricted parameter space under the null hypothesis and $\Omega_0 \cup \Omega_1$ represents the parameter space under the alternative hypothesis, encompassing the full, unrestricted model space that allows for the greatest flexibility in fitting the data to the model. These parameter spaces are the alternative sets defined in the previous section. A large value of $\text{GLR}_t$ means we are more confident in rejecting the alternative hypothesis $H_1$.

~~In our setting, the observations are obtained from a designed sampling rule.~~

We limit our setting to a single-parameter exponential family parameterized by its mean, as in Garivier and Kaufmann (2016), which includes Bernoulli distribution, Poisson distribution, Gamma distributions with known shape parameter, or Gaussian distribution with known variance.

We need to test the following hypotheses: $H_{0,x} : \boldsymbol{\mu} \in \text{Alt}_x(\hat{\boldsymbol{\mu}})$ against $H_{1,x} : \boldsymbol{\mu} \notin \text{Alt}_x(\hat{\boldsymbol{\mu}})$ for each culprit $x \in \mathcal{X}(\hat{\boldsymbol{\mu}})$, where $\hat{\boldsymbol{\mu}}$ is the empirical mean based on observed data. If $\hat{\boldsymbol{\mu}}(t) \in \Omega_0 \cup \Omega_1$, equation (16) can be shown in the ~~following~~ form of self-normalized sum, giving us the formal expression of GLR statistic in Proposition 3 through the calculation of maximum likelihood estimation (MLE) and the rewriting of KL divergence. (CH: This is not clear, please explain briefly)(ZK L: Adjusted.)

~~$$\ln(\text{GLR}_t) = \inf_{\boldsymbol{\vartheta} \in \Omega_0} \sum_{i=1}^{K} N_i(t) \text{KL}(\hat{\mu}_i(t), \vartheta_i). \tag{17}$$~~

PROPOSITION 3. *The Generalized Likelihood Ratio statistic with respect to each culprit $x \in \mathcal{X}(\boldsymbol{\mu})$ and each time step $t$ for the pure exploration setting is defined as*

$$\begin{aligned} \hat{\Lambda}_{t,x} &= \ln(GLR_t) \\ &= \inf_{\boldsymbol{\vartheta} \in Alt_x(\hat{\boldsymbol{\mu}}(t))} \ln \frac{L(\hat{\boldsymbol{\mu}})}{L(X_1, X_2, \ldots, X_t; \boldsymbol{\vartheta})} \\ &= \inf_{\boldsymbol{\vartheta} \in Alt_x(\hat{\boldsymbol{\mu}}(t))} \sum_{i \in [K]} N_i(t) KL(\hat{\mu}_i(t), \vartheta_i), \end{aligned} \tag{18}$$

*where $KL(\cdot, \cdot)$ represents the KL divergence of the two distributions parameterized by their means, $L(\cdot)$ represents the likelihood depending on the observed data along with some unknown parameter, and $N_i(t) = t \cdot p_i$ for every $i \in [K]$ is the random observations allocated to each arm up to time step $t$. (CH: Change this to a proposition)(ZK L: Changed. All the references of this proposition in this paper have been adjusted as well.)*

This connects the GLR test with information-theoretic methods. To quantify the amount of information and confidence we have to assert that the unknown true mean value does not belong to $\text{Alt}_x$ for all $x \in \mathcal{X}$ (CH: which Alt?)(ZK L: Explained. $C_x$ corresponds $\text{Alt}_x$), we define the $C_x$ function as the population version of the GLR statistic, which is the same form as equation (18).

$$C_x(\boldsymbol{p}) = C_x(\boldsymbol{p}; \boldsymbol{\mu}) \triangleq \inf_{\boldsymbol{\vartheta} \in \text{Alt}_x} \sum_{i \in [K]} p_i d(\mu_i, \vartheta_i), \text{ for every } x \in \mathcal{X}(\boldsymbol{\mu}). \tag{19}$$

Then with the introduction of culprits and $C_x$ function, we define the following optimal allocation problem in Proposition 4.

PROPOSITION 4. *(Qin and You 2023) For any $\boldsymbol{\mu} \in M$, there exists set $\mathcal{X} = \mathcal{X}(\boldsymbol{\mu})$ and functions $\{C_x\}_{x \in \mathcal{X}}$ with $C_x : \mathcal{S}_K \times M \to \mathbb{R}_+$ such that*

$$\kappa(\boldsymbol{\mu}) \geq \left(\Gamma^*_{\boldsymbol{\mu}}\right)^{-1}, \tag{20}$$

*where*

$$\Gamma^*_{\boldsymbol{\mu}} = \max_{\boldsymbol{p} \in \mathcal{S}_K} \min_{x \in \mathcal{X}} C_x(\boldsymbol{p}; \boldsymbol{\mu}). \tag{21}$$

However, computing the lower bound can still be hard since it requires the solution of the minimax problem in equation (21). While the KL divergence in equation (19) is convex for Gaussians, it can be non-convex to minimize the $C_x$ function over the culprit set $\mathcal{X}(\boldsymbol{\mu})$. To solve this problem, depending on the following Proposition 5, we can write $\mathcal{X}(\boldsymbol{\mu})$ as a union of several convex sets ~~as follows.~~ . The following three equivalent expressions represent different ways of describing the lower bound, making the minimax problem in equation (21) tractable with the introduction of culprit set $\mathcal{X}$ and the corresponding alterantive set $\text{Alt}_x$ for every $x \in \mathcal{X}$.

$$\Gamma^*_{\boldsymbol{\mu}} = \max_{\boldsymbol{p} \in \mathcal{S}_K} \inf_{\boldsymbol{\vartheta} \in \text{Alt}(\boldsymbol{\mu})} \sum_{i \in [K]} p_i d(\mu_i, \vartheta_i) = \max_{\boldsymbol{p} \in \mathcal{S}_K} \min_{x \in \mathcal{X}} \inf_{\boldsymbol{\vartheta} \in \text{Alt}_x} \sum_{i \in [K]} p_i d(\mu_i, \vartheta_i) \tag{22}$$

$$= \max_{\boldsymbol{p} \in \mathcal{S}_K} \min_{x \in \mathcal{X}} \sum_{i \in [K]} p_i d(\mu_i, \vartheta_i^x) \tag{23}$$

$$= \max_{\boldsymbol{p} \in \mathcal{S}_K} \min_{x \in \mathcal{X}} C_x(\boldsymbol{p}), \tag{24}$$

where we used the existence of finite union set and unique minimizer $\boldsymbol{\vartheta}^x$ in Proposition 5. (CH: Please make it more clear. )(ZK L: Adjusted.)

PROPOSITION 5. *Assuming that the arms distribution follows a canonical single-parameter exponential family parameterized by its mean value. Then, for each culprit $x \in \mathcal{X}(\boldsymbol{\mu})$,*

1. *(Wang et al. 2021) (Assumption 1). For each problem instance $\boldsymbol{\mu} \in M$, the alternative set $Alt(\boldsymbol{\mu})$ is a finite union of convex sets. Namely, there exists a finite collection of convex sets $\{Alt_x(\boldsymbol{\mu}) : x \in \mathcal{X}(\boldsymbol{\mu})\}$ such that $Alt(\boldsymbol{\mu}) = \cup_{x \in \mathcal{X}(\boldsymbol{\mu})} Alt_x(\boldsymbol{\mu})$.*
2. *Given a specific simplex distribution $\boldsymbol{p}$, there exists a unique $\boldsymbol{\vartheta}^x \in Alt_x(\boldsymbol{\mu})$ that achieves the infimum in (19).*

**4.1.3. Stopping Rule** ~~In this section, we will introduce the deviation inequalities to substantiate the correctness of a stopping rule for the general sequential adaptive hypothesis testing problem~~ In this section, we introduce the stopping rule, which suggests when to stop the algorithm and returns an answer that gives all $\epsilon$-best arms with probability of at least $1-\delta$. This stopping rule is based on the deviation inequalities that are linked to the generalized likelihood ratio test (Kaufmann and Koolen 2021).

For each $x_i \in \mathcal{X}(\boldsymbol{\mu})$ and $i \in [|\mathcal{X}|]$, let $M_i(\boldsymbol{\mu}) = \text{Alt}_{x_i}(\boldsymbol{\mu})$ be a partition of the whole realizable parameter space $M$ considered in Section 4.1.1, with each partition $M_i(\boldsymbol{\mu})$ being characterized by a culprit in $\mathcal{X}(\boldsymbol{\mu})$. This means for each $\boldsymbol{\mu} \in M$, we can construct a unique partition of the parameter space $M$ to form our hypothesis test. $M_0(\boldsymbol{\mu})$ denotes the set of parameters where $\boldsymbol{\mu}$ truly resides. If $\boldsymbol{\mu} \in M$, define $i^*(\boldsymbol{\mu})$ as the index of the unique element in the partition where the true mean value $\boldsymbol{\mu}$ belongs and here $i^*(\boldsymbol{\mu}) = 0$. In other words, we have $\boldsymbol{\mu} \in M_0$ and $\text{Alt}(\boldsymbol{\mu}) = M \backslash M_0$. Since we do not care about the order among suboptimal arms, $M_i(\boldsymbol{\mu})$ for $i \in \{0, 1, 2, \ldots, |\mathcal{X}|\}$ can form a partition of $M$ for each $\boldsymbol{\mu}$. The alternative set can thus be further defined as

$$\text{Alt}(\boldsymbol{\mu}) = \cup_{i:\boldsymbol{\mu} \notin M_i(\boldsymbol{\mu})} M_i(\boldsymbol{\mu}) = M \backslash M_{i^*(\boldsymbol{\mu})} = M \backslash M_0. \tag{25}$$

Given a bandit instance $\boldsymbol{\mu}$, we consider $|\mathcal{X}(\boldsymbol{\mu})| + 1$ hypotheses, given by

$$H_0 = (\boldsymbol{\mu} \in M_0(\boldsymbol{\mu})), H_1 = (\boldsymbol{\mu} \in M_1(\boldsymbol{\mu})), \ldots, H_{|\mathcal{X}|} = (\boldsymbol{\mu} \in M_{|\mathcal{X}|}(\boldsymbol{\mu})). \tag{26}$$

Substitute true mean value $\boldsymbol{\mu}$ with the empirical value $\hat{\boldsymbol{\mu}}(t)$, the GLR test? depends on the empirical value $\hat{\boldsymbol{\mu}}(t)$ in each time $t$. The hypotheses tested at time $t$ are data-dependent. If $\hat{\boldsymbol{\mu}}(t) \in M$, define $\hat{i}(t) = i^*(\hat{\boldsymbol{\mu}}(t))$ as the index of the partition to which $\hat{\boldsymbol{\mu}}(t)$ belong; in other words, $\hat{\boldsymbol{\mu}}(t) \in M_{\hat{i}(t)}$; otherwise if $\hat{\boldsymbol{\mu}}(t) \notin M$, $\hat{\Lambda}_{t,x} = 0$ for each pair of hypotheses in this time step and the process does not stop. In practice, if $\hat{\boldsymbol{\mu}}(t) \notin M$, any designed algorithm can be replaced by the uniform exploration. With this forced exploration, as true mean value ~~$\boldsymbol{\mu}(t) \in M$~~ $\boldsymbol{\mu} \in M$ (which is an open set by assumption), the law of large numbers ensures that at some point the empirical mean value will fall back into our parameter space again, i.e., $\hat{\boldsymbol{\mu}}(t) \in M$.

~~Our objective is to sequentially sample the arms until a decision is reached, confirming the correctness of one of the hypotheses.~~ We run $|\mathcal{X}|$ time-varying GLR tests in parallel, each of which tests $H_0$ against $H_i$ for $i \in [|\mathcal{X}(\hat{\boldsymbol{\mu}}(t))|]$. We stop when one of the tests rejects $H_0$. This means that we have empirically found the alternative set that is most easily rejected and we have identified the accepted hypothesis for $\hat{\boldsymbol{\mu}}(t) \in M$ with the highest probability of being right. Given a sequence of exploration rate $(\hat{\beta}_t(\delta))_{t\in\mathbb{N}}$, the GLR stopping rule for the pure exploration setting ~~with this time-dependent threshold~~ is defined as

$$\tau_\delta \triangleq \inf\left\{t \in \mathbb{N} : \min_{x \in \mathcal{X}(\hat{\boldsymbol{\mu}}(t))} \hat{\Lambda}_{t,x} > \hat{\beta}_t(\delta)\right\} = \inf\left\{t \in \mathbb{N} : t \cdot \min_{x \in \mathcal{X}(\hat{\boldsymbol{\mu}}(t))} C_x(\boldsymbol{p}_t; \hat{\boldsymbol{\mu}}(t)) > \hat{\beta}_t(\delta)\right\}, \tag{27}$$

where The GLR statistic $\hat{\Lambda}_{t,x}$ is defined in equation (18). The testing process is quite similar to the classical one except that the hypotheses in each round are data-dependent, changing with time.

We also give some insight from the perspective of the confidence region, which is quite common in the bandit literature, enabling us to understand many other algorithms intuitively and equivalently. It can be noted that ~~$\left\{\hat{\Lambda}_t > \hat{\beta}_t(\delta)\right\} = \left\{\mathcal{C}_t(\delta) \subseteq M_{\hat{i}(t)}\right\}$~~ $\left\{\min_{x\in\mathcal{X}(\hat{\boldsymbol{\mu}}(t))} \hat{\Lambda}_{t,x} > \hat{\beta}_t(\delta)\right\} = \left\{\mathcal{C}_t(\delta) \subseteq M_{\hat{i}(t)}\right\}$, where $\mathcal{C}_t(\delta)$ is the confidence region of mean value, given by

$$\mathcal{C}_t(\delta) \triangleq \left\{\boldsymbol{\lambda} : \sum_{i=1}^{K} N_i(t) d(\hat{\mu}_i(t), \lambda_i) \leq \hat{\beta}_t(\delta)\right\}. \tag{28}$$

The stopping rule can thus be interpreted as follows: the algorithm stops when the confidence region of mean value parameters shrinks into one part of the partition. Based on this stopping framework, different research has been conducted to explore the design of the exploration rate $\hat{\beta}_t(\delta)$ and concentration inequalities, achieving the $\delta$-*PAC* performance for any sampling rule (Kaufmann and Koolen 2021). The anytime and decomposition characteristics of the stopping rule make it possible for our algorithm to concentrate on the sampling rule and be used in the fixed-budget setting directly.

## 4.2. A Lower Bound on the Sample Complexity of All $\varepsilon$-Best Arms Identification Problem with Linear Context

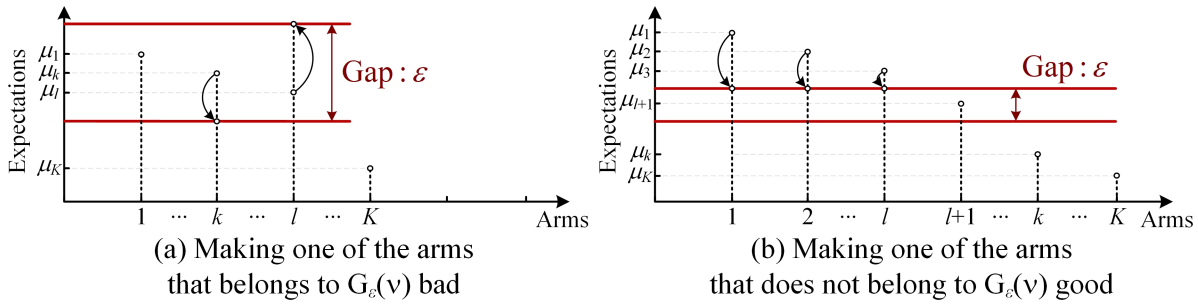
For the All $\varepsilon$-Best Arms Identification problem, we initially introduce the current lower bound results in the stochastic setting. Then, we extend the lower bound from the stochastic setting to the linear setting, which serves as the analysis basis of our algorithm and subsequent analyses. Furthermore, this lower bound is also the first result of this problem to the best of our knowledge.

**4.2.1. Lower Bound for the All $\varepsilon$-Best Arms Identification in Stochastic Bandit.** As stated in the last section, identifying the lower bound involves constructing the largest possible alternative set and pinpointing the most challenging one to identify, (CH: We might want to explain again in one sentence what is the relationship of the alternative set to the culprit. )(ZK L: Content adjusted, it is unnecessary to mention the culprit here.) ~~Before finding this culprit,~~ which means it is necessary to construct an alternative set as comprehensive as possible to get the tightest lower bound.

For the All $\varepsilon$-Best Arms Identification in stochastic bandit, Marjani et al. (2022) considered all possible situations, providing a complete characterization of the alternative bandit instances that the optimal sampling strategy needs to rule out, making their the lower bound the tightest up to now. Their characterization is based on two cases shown in Figure 1, clearly outlining the methods for constructing the alternative set. The following theorem displays the lower bound results, the detailed proof of which can be checked in Section 3 of Marjani et al. (2022).

We recall the following major components and we need to make the forms of these components clear to give the current lower bound result of All $\varepsilon$-Best Arms Identification in stochastic bandit.

1. The correct answer $G_\varepsilon(\nu)$ corresponding to the preset exploration task.




(a) Making one of the arms that belongs to $G_\varepsilon(\nu)$ bad

(b) Making one of the arms that does not belong to $G_\varepsilon(\nu)$ good

**Figure 1** Marjani et al. (2022) The left Figure (a) makes one of the arms that belong to $G_\varepsilon(\nu)$ bad. This is achieved by decreasing the expectation of some $\varepsilon$-best arm $k$ while increasing the expectation of some other arm $\ell$ to the point where $k$ is no more $\varepsilon$-best. The right Figure (b) makes one of the arms that does not belong to $G_\varepsilon(\nu)$ good. This is achieved by increasing the expectation of some arm $k$ that does not belong to $G_\varepsilon(\nu)$ while decreasing the expectations of the arms with the largest means to the point where $k$ becomes $\varepsilon$-best.

2. The culprit set $\mathcal{X}(\boldsymbol{\mu})$.
3. The alternative set $\text{Alt}(\boldsymbol{\mu})$ based on the culprit set $\mathcal{X}(\boldsymbol{\mu})$, which satisfies the existence of optimal solutions and the union structure in Proposition 5.
4. The $C_x(\boldsymbol{p})$ function that quantifies the amount of information we have to reject the wrong hypothesis.

Following the description in Figure 1 by (Marjani et al. 2022). The culprit $x$ and the alternative set $\text{Alt}_x$ can be defined as

$$\mathcal{X}(\boldsymbol{\mu}) = \{(i,j,m,\ell) : i \in G_\varepsilon(\nu), j \neq i, m \notin G_\varepsilon(\nu), \ell \in [|1, m-1|]\}, \tag{29}$$

$$\text{Alt}_x(\boldsymbol{\mu}) = \text{Alt}_{i,j}(\boldsymbol{\mu}) \cup \text{Alt}_{m,\ell}(\boldsymbol{\mu}) \text{ for all } x \in \mathcal{X}(\boldsymbol{\mu}), \tag{30}$$

$$\text{Alt}_{i,j}(\boldsymbol{\mu}) = \{\boldsymbol{\mu} : \mu_i - \mu_j < -\varepsilon\} \text{ for all } x \in \mathcal{X}(\boldsymbol{\mu}), \tag{31}$$

$$\text{Alt}_{m,\ell}(\boldsymbol{\mu}) = \left\{\boldsymbol{\mu} : \mu_\ell \geq \mu_\varepsilon^{m,\ell} p + \varepsilon \geq \mu_{\ell+1}\}\right\} \text{ for all } x \in \mathcal{X}(\boldsymbol{\mu}), \tag{32}$$

where

$$\mu_\varepsilon^{m,\ell}(p) \triangleq \frac{p_m \mu_m + \sum_{i=1}^{\ell} p_i \left(\mu_i - \varepsilon\right)}{p_m + \sum_{i=1}^{\ell} p_i}. \tag{33}$$

Then we can formally introduce the following Theorem 2 to present the latest lower bound for All $\varepsilon$-Best Arms Identification in stochastic bandit.

THEOREM 2. *Marjani et al. (2022) (Section 3: Best response oracle). Fix $\delta, \varepsilon > 0$ and consider $K$ arms. Let the $i^{th}$ be distributed according to $\mathcal{N}(\mu_i, 1)$ and any $\delta$-PAC algorithm for the additive setting satisfies*

$$\begin{aligned}
\frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/2.4\delta)} &\geq \min_{p \in \mathcal{S}_K} \max \left\{ \frac{1}{\min_{x \in \mathcal{X}} C_{i,j}(p)}, \frac{1}{\min_{x \in \mathcal{X}} C_{m,\ell}(p)} \right\} \\
&= \min_{p \in \mathcal{S}_K} \max_{x \in \mathcal{X}} \max \left\{ \frac{1}{C_{i,j}(p)}, \frac{1}{C_{m,\ell}(p)} \right\} \\
&= 2 \min_{p \in \mathcal{S}_K} \max_{x \in \mathcal{X}} \max \left\{ \frac{1/p_i + 1/p_j}{(\mu_i - \mu_j + \varepsilon)^2}, \right.
\end{aligned}$$

$$\left.\frac{1}{\inf_{\boldsymbol{\mu}'\in Alt_{m,\ell}(\boldsymbol{\mu})}\left[\sum_{i=1}^{\ell}(\mu_i-\mu_\varepsilon^{m,\ell}(p)-\varepsilon)^2p_i+(\mu_m-\mu_\varepsilon^{m,\ell}(p))^2p_m\right]}\right\}$$
$$=\left(\Gamma_{\boldsymbol{\mu}}^*\right)^{-1}, \tag{34}$$

*where the indices $i$, $j$, $m$, and $l$, which specify the culprit set, are introduced in equation (29). The alternative set $Alt_{m,\ell}(\boldsymbol{\mu})$ is defined in equation (32)*

In addition, by choosing special culprits $x\in\mathcal{X}$ in Theorem 2, we can recover the former lower bound result for All $\varepsilon$-Best Arms Identification in stochastic bandit proposed by Mason et al. (2020). Specifically, we have

1. Let $j=1$, $j\neq i$, $i\in G_\varepsilon(\nu)$, $\ell=1$ and $\mu_\varepsilon^{m,\ell}(p)=\mu_1-\varepsilon$, the first term and second term in Theorem 2 recover the first term in Theorem 2.1 of Mason et al. (2020).
2. Let $i=k$ and thus $j\neq k$, the first term in Theorem 2 recover the second term in Theorem 2.1 of Mason et al. (2020).

**4.2.2. Extension to Linear Bandit** We recall the following major components:

1. The correct answer $G_\varepsilon(\nu)$ corresponding to the preset exploration task.
2. The culprit set $\mathcal{X}(\boldsymbol{\mu})$.
3. The alternative set $\mathrm{Alt}(\boldsymbol{\mu})$ based on the culprit set $\mathcal{X}(\boldsymbol{\mu})$, which satisfies the existence of solutions and the union structure in Proposition 5.
4. The $C_x(\boldsymbol{p})$ function that quantifies the amount of information we have to reject the matching hypothesis.

(ZK L: The content related to the linear bandit has been merged into the section of Problem Formulation and thus the code is commented out.)

In the context of stochastic bandits, the alternative set comprises instances featuring distinct mean value vectors. However, in the linear bandit, the feature vector for each arm $i\in[K]$ is known, and the alternative set is characterized by variations in the parameter $\boldsymbol{\theta}$. This leads to the following critical extension results:

1. The correct answer $\mathcal{I}=i^*\triangleq\arg\max_{i\in[K]}\mu_i$, where $\mu_i=\langle\boldsymbol{\theta},\boldsymbol{a}_i\rangle$.
2. The culprit set $\mathcal{X}(\boldsymbol{\mu})=[K]\backslash\{I^*\}$.
3. The alternative set $\mathrm{Alt}(\boldsymbol{\mu})=\cup_{x\in\mathcal{X}(\boldsymbol{\mu})}\mathrm{Alt}_x(\boldsymbol{\mu})=\cup_{j\in\mathcal{X}(\boldsymbol{\mu})}\{\boldsymbol{\theta}:\langle\boldsymbol{\theta},\boldsymbol{a}_j-\boldsymbol{a}_{I^*}\rangle>0\}$.
4. The infimum $C_j(p)=\sum_{i=1}^K p_i d(\mu_i,\langle\boldsymbol{\vartheta}_j,\boldsymbol{a}_i\rangle)$ for all $j\in\mathcal{X}$ (CH: $j$ appears twice here), where $j$ belongs to the set $\boldsymbol{\vartheta}=\{\boldsymbol{\vartheta}_j:j\in\mathcal{X}\}$, given by

$$\boldsymbol{\vartheta}_j=\boldsymbol{\theta}-\left(\frac{\langle\boldsymbol{\theta},\boldsymbol{a}_{I^*}-\boldsymbol{a}_j\rangle}{\|\boldsymbol{a}_{I^*}-\boldsymbol{a}_j\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}}\boldsymbol{V}_{\boldsymbol{p}}^{-1}\right)(\boldsymbol{a}_{I^*}-\boldsymbol{a}_j), \tag{35}$$

where $\boldsymbol{V_p} = \sum_{i=1}^{K} p_i \boldsymbol{a}_i \boldsymbol{a}_i^\top$ is closely related to the Fisher information matrix (Chaloner and Verdinelli 1995) Without loss of generality, we assume that $\mu_i \sim \mathcal{N}(\boldsymbol{a}_i\boldsymbol{\theta}, 1)$, and the problem complexity derived in Proposition 4 can be written as

$$\Gamma^* = \max_{p\in\mathcal{S}_K} \min_{j\in\mathcal{X}} \frac{\langle \boldsymbol{\theta}, \boldsymbol{a}_{I^*} - \boldsymbol{a}_j\rangle^2}{2\|\boldsymbol{a}_{I^*} - \boldsymbol{a}_j\|^2_{\boldsymbol{V_p}^{-1}}}, \tag{36}$$

which means that

$$C_j(p) = \frac{\langle \boldsymbol{\theta}, \boldsymbol{a}_{I^*} - \boldsymbol{a}_j\rangle^2}{2\|\boldsymbol{a}_{I^*} - \boldsymbol{a}_j\|^2_{\boldsymbol{V_p}^{-1}}}. \tag{37}$$

(CH: Please explain in more detail how you obtain the previous two equations.)

**4.2.3. Lower Bound for the All $\varepsilon$-Best Arms Identification in Linear Bandit.** In Mason et al. (2020), Marjani et al. (2022), the All $\varepsilon$-Best Arms Identification problem in linear bandit was introduced and its complexity was analyzed in this section. However, the All $\varepsilon$-Best Arms Identification problem within the linear setting remains unsolved. The following Theorem 3 is the lower bound of All $\varepsilon$-Best Arms Identification problem in the linear setting, which is, to the best of our knowledge, the first result in the literature. Furthermore, we also provide a graphical explanation for the lower bound. From Figure 2, we develop a new method of deriving the lower bound, providing the result that is exactly the same form as equation (38) in Theorem 3 and revealing the nature of the problem at the same time.

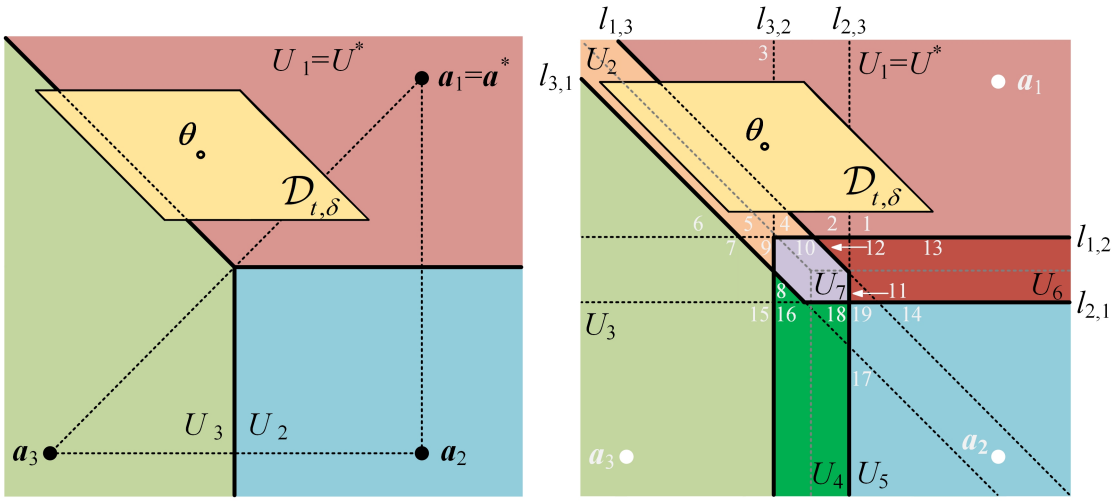
Our lower bound is an extension from the result in Theorem 2. However, the lower bound in this theorem is extremely complex and is quite hard to utilize or reference by new research. Thus, we fine-tune the formula of the alternative set depicted in Figure 1, simplifying its form while ensuring the completeness of the alternative set and the tightness of the lower bound. Compared to equation (34) in Theorem 2, the result in Theorem 3 assumes that the mean value of arm 1 remains the same, partly sacrificing the completeness of the alternative set in equation (38) regarding its second term but is presented in an explicit and symmetric form, which is tight and proves to be of great help to our theoretical analysis. We provide the proof and detailed notation in EC.1.

THEOREM 3. *Fix $\delta, \varepsilon > 0$. Consider arms, such that the $i^{th}$ is distributed according to $\mathcal{N}(\mu_i, 1)$. Any $\delta$-PAC algorithm for the additive setting with linear context satisfies*

$$\frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/2.4\delta)} \geq (\Gamma^*)^{-1} = \min_{\boldsymbol{p}\in S_K} \max_{(i,j,m)\in\mathcal{X}} \max\left\{ \frac{2\|\boldsymbol{a}_i - \boldsymbol{a}_j\|^2_{\boldsymbol{V_p}^{-1}}}{\left(\boldsymbol{a}_i^\top\boldsymbol{\theta} - \boldsymbol{a}_j^\top\boldsymbol{\theta} + \varepsilon\right)^2}, \frac{2\|\boldsymbol{a}_1 - \boldsymbol{a}_m\|^2_{\boldsymbol{V_p}^{-1}}}{\left(\boldsymbol{a}_1^\top\boldsymbol{\theta} - \boldsymbol{a}_m^\top\boldsymbol{\theta} - \varepsilon\right)^2} \right\}, \tag{38}$$

*where $\mathcal{X}(\boldsymbol{\mu}) = \{(i,j,m) : i \in G_\varepsilon(\nu), j \neq i, m \notin G_\varepsilon(\nu)\}$.*

Besides, note that the stochastic bandit is a special case of the linear bandit. Let $\boldsymbol{e}_i$ ($i \in [d]$) be the unit vector and $\mathcal{A} = \{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_d\}$, then the linear bandit problem reduces to the stochastic setting. Considering this relationship, we can recover the lower bound result for All $\varepsilon$-best Identification in the stochastic bandit in Theorem 2 by taking the set of arms as $\mathcal{A} = \{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_d\}$ and letting parameter $\boldsymbol{\theta}$ equal $\boldsymbol{\mu}$ in




(a) Best Arm Identification (BAI) in Linear Bandit

(b) All $\varepsilon$-Best Arms Identification in Linear Bandit

**Figure 2** (a) The dots correspond to three arms with ~~$\boldsymbol{\theta} \in U(\boldsymbol{a}_1)$~~ $\boldsymbol{\theta} \in U_1$ and the best arm $\boldsymbol{a}^* = \boldsymbol{a}_1$. The confidence set $\mathcal{D}_{t,\delta}$ is aligned with directions $\boldsymbol{a}_1 - \boldsymbol{a}_2$ and $\boldsymbol{a}_1 - \boldsymbol{a}_3$. Three solid boundaries form the decision regions. $\mathcal{D}_{t,\delta}$ spans two different decision regions ($U_1$ and $U_2$), thus both $\boldsymbol{a}_1$ and $\boldsymbol{a}_3$ may be the optimal arm. (b) The decision regions correspond to seven areas formed by 19 partitions with three arms (dots). Each partition is formed by ~~xxx~~ the intersection of dotted lines on both sides of the BAI decision boundaries, namely $l_{i,j}$. The true parameter $\boldsymbol{\theta}$ lies in $U_1 = U^*$ and we have one arm that is $\varepsilon$-best, i.e., $G_\varepsilon(\nu) = \{1\}$. Given the uncertainty in the region $\mathcal{D}_{t,\delta}$, it spans two different decision regions ($U_1$ and $U_2$) and the set of all $\varepsilon$-best arms can be either $G_\varepsilon(\nu) = \{1\}$ or $G_\varepsilon(\nu) = \{1,3\}$. In fact, the extension relationship between BAI and All-$\varepsilon$-Best Arms Identification can be clearly seen from the figure as the situation of BAI is the limiting case of All $\varepsilon$-Best Arms Identification when the gap $\varepsilon$ approaches 0.

Theorem 3. (ZK L: Here I want to mention two points. The first is that we can't recover the exact form in Theorem 3 since we made the simplification. The second point is that I did not provide detailed process for this recovery.)

Furthermore, from (28), we use a figure to illustrate the essence of the stopping rule and the problem complexity for All $\varepsilon$-Best Arms in linear bandit. Assume that the oracle knows exactly which arms are $\varepsilon$-best. A $d$-dimensional Euclidean space $\mathbb{R}^d$ can be divided by lines $l_{i,j}$ for any $i, j \in [K]$ $(j \neq i)$, which are perpendicular to the direction of $\boldsymbol{a}_i - \boldsymbol{a}_j$, forming the following half-spaces.

$$H_{i,j}^{+} = \left\{ \boldsymbol{\vartheta} \in \mathbb{R}^d \mid (\boldsymbol{a}_i - \boldsymbol{a}_j)^\top \boldsymbol{\vartheta} > \varepsilon \right\}, \tag{39}$$

$$H_{i,j}^{-} = \left\{ \boldsymbol{\vartheta} \in \mathbb{R}^d \mid (\boldsymbol{a}_i - \boldsymbol{a}_j)^\top \boldsymbol{\vartheta} \leq \varepsilon \right\}. \tag{40}$$

Graphically, it can be understood from an intuitive angle that the lines on both sides of the BAI decision boundaries, namely $l_{i,j}$, come closer to the boundaries as the gap $\varepsilon$ approaches 0 and the decision regions $U_i$ $(i = 2, 4, 6, 7)$ shrink until they disappear.

The relationship between parameter $\boldsymbol{\theta}$ and half-spaces determines which arms belong to set $G_\varepsilon(\nu)$. If $\boldsymbol{\theta} \in H_{i,j}^+$, then arm $j$ does not belong to the set of $\varepsilon$-best arms $G_\varepsilon(\nu)$; conversely, if $\boldsymbol{\theta} \in H_{i,j}^-$ for every $j \neq i$, arm $j$ belongs to the set of $\varepsilon$-best arms $G_\varepsilon(\nu)$. As illustrated in Figure 2, the intersection of these half-spaces, with the form of $\left\{\cap_{i,j\in[K]\ (j\neq i)} H_{i,j} \mid H_{i,j} \in \{H_{i,j}^+, H_{i,j}^-\}\right\}$, forms a partition of the space. Each partition returns an answer set $G_\varepsilon(\nu)$, representing all $\varepsilon$-best arms of a bandit instance if the true parameter $\boldsymbol{\theta}$ lies in this area. The decision regions, denoted by $U_i \in \mathcal{U}$, can thus be defined as the union of partitions with the same returned answer set.

Take the case of 3 arms for instance, the space is divided into 19 partitions as shown in Figure 2. These partitions can further form seven ($2^3 - 1 = 7$) decision regions from $U_1 = U^*$ to $U_7$, which is summarized in Table 1. In addition, from Figure 2 we find that the situation of BAI is the limiting case of All $\varepsilon$-Best Arms Identification as gap $\varepsilon$ approaches 0. We assume that for any sampling policy, it is possible to con-

**Table 1** 19 Partitions and Seven Decision Regions in the Case of Three Arms

| Decision Region $U$ | Partition Index | Set Expression of Decision Region | All $\varepsilon$-Best Arms $G_\varepsilon(\nu)$ |
|---|---|---|---|
| $U_1$ | 1 | $H_{1,2}^+ \cap H_{1,3}^+ \cap H_{3,1}^- \cap H_{2,1}^-$ | $\{1\}$ |
| | 2 | $H_{1,2}^+ \cap H_{1,3}^+ \cap H_{3,1}^- \cap H_{2,1}^-$ | $\{1\}$ |
| | 3 | $H_{1,2}^+ \cap H_{1,3}^+ \cap H_{3,1}^- \cap H_{2,1}^-$ | $\{1\}$ |
| $U_2$ | 4 | $H_{1,2}^+ \cap H_{3,1}^- \cap H_{2,1}^- \cap H_{1,3}^- \cap H_{2,3}^-$ | $\{1, 3\}$ |
| | 5 | $H_{1,2}^+ \cap H_{3,1}^- \cap H_{2,1}^- \cap H_{1,3}^- \cap H_{2,3}^-$ | $\{1, 3\}$ |
| | 9 | $H_{1,2}^+ \cap H_{3,1}^+ \cap H_{1,3}^- \cap H_{2,3}^-$ | $\{1, 3\}$ |
| $U_3$ | 6 | $H_{1,2}^+ \cap H_{3,1}^+ \cap H_{1,3}^- \cap H_{2,3}^-$ | $\{3\}$ |
| | 7 | $H_{3,2}^+ \cap H_{3,1}^+ \cap H_{1,3}^- \cap H_{2,3}^-$ | $\{3\}$ |
| | 15 | $H_{3,2}^+ \cap H_{2,1}^+ \cap H_{2,3}^- \cap H_{1,3}^-$ | $\{3\}$ |
| $U_4$ | 8 | $H_{3,2}^- \cap H_{1,2}^- \cap H_{3,1}^+ \cap H_{1,3}^- \cap H_{2,3}^-$ | $\{2, 3\}$ |
| | 16 | $H_{2,1}^+ \cap H_{2,3}^- \cap H_{1,3}^- \cap H_{1,2}^- \cap H_{3,2}^-$ | $\{2, 3\}$ |
| | 18 | $H_{2,1}^+ \cap H_{1,2}^- \cap H_{3,2}^- \cap H_{2,3}^- \cap H_{1,3}^-$ | $\{2, 3\}$ |
| $U_5$ | 14 | $H_{1,3}^+ \cap H_{2,1}^+ \cap H_{1,2}^- \cap H_{3,2}^-$ | $\{2\}$ |
| | 17 | $H_{2,1}^+ \cap H_{2,3}^+ \cap H_{1,2}^- \cap H_{3,2}^-$ | $\{2\}$ |
| | 19 | $H_{2,1}^+ \cap H_{2,3}^+ \cap H_{1,2}^- \cap H_{3,2}^-$ | $\{2\}$ |
| $U_6$ | 11 | $H_{2,3}^+ \cap H_{3,1}^- \cap H_{2,1}^- \cap H_{1,2}^- \cap H_{3,2}^-$ | $\{1, 2\}$ |
| | 12 | $H_{1,3}^+ \cap H_{3,1}^- \cap H_{2,1}^- \cap H_{1,2}^- \cap H_{3,2}^-$ | $\{1, 2\}$ |
| | 13 | $H_{2,3}^+ \cap H_{3,1}^- \cap H_{2,1}^- \cap H_{1,2}^- \cap H_{3,2}^-$ | $\{1, 2\}$ |
| $U_7$ | 10 | $H_{1,3}^- \cap H_{2,3}^- \cap H_{3,1}^- \cap H_{2,1}^- \cap H_{1,2}^- \cap H_{3,2}^-$ | $\{1, 2, 3\}$ |

struct a confidence region $\mathcal{D}_{t,\delta} \subset \mathbb{R}^d$ that $\boldsymbol{\theta} \in \mathcal{D}_{t,\delta}$ and the empirical estimate $\hat{\boldsymbol{\theta}}_t$ belongs to $\mathcal{D}_{t,\delta}$ with high probability, i.e., $\mathbb{P}\left(\hat{\boldsymbol{\theta}}_t \in \mathcal{D}_{t,\delta}\right) \geq 1 - \delta$. Thus, as stated in equation (28), the oracle stopping criterion simply checks whether the confidence region $\mathcal{D}_{t,\delta}$ is contained in some decision region $U \in \mathcal{U}$ or not. We assume

that the oracle knows the decision region $U^*$ containing all the parameters for which all $\varepsilon$-best arms are correctly identified. In fact, from Figure 2 we can find that for a sampling policy whenever the region $\mathcal{D}_{t,\delta}$ overlaps different decision regions, there exists ambiguity about finding all the arms belonging to $G_\varepsilon(\nu)$. On the other hand when all possible values of $\boldsymbol{\theta}_t$ are included with high probability in the right decision region, then the set of all $\varepsilon$-best arms is correctly returned.

From Figure 2 it can be seen that the more $\mathcal{D}_{t,\delta}$ is aligned with the boundaries of the decision region, the easier and faster it is to shrink into $U^*$. To put it formally, the condition $\mathcal{D}_{t,\delta} \subset U^*$ can be represented as

$$\forall i \in G_\varepsilon(\nu), j \neq i, m \notin G_\varepsilon(\nu), \forall \boldsymbol{\vartheta} \in \mathcal{D}_{t,\delta}, \boldsymbol{\vartheta} \in H_{j,i}^- \text{ and } \boldsymbol{\vartheta} \in H_{1,m}^+. \tag{41}$$

where $\boldsymbol{\vartheta} \in H_{j,i}^-$ and $\boldsymbol{\vartheta} \in H_{1,m}^+$ are equivalent to the following inequalities.

$$\begin{cases} (\boldsymbol{a}_j - \boldsymbol{a}_i)^\top (\boldsymbol{\theta} - \boldsymbol{\vartheta}) \geq (\boldsymbol{a}_j - \boldsymbol{a}_i)^\top \boldsymbol{\theta} - \varepsilon \\ (\boldsymbol{a}_1 - \boldsymbol{a}_m)^\top (\boldsymbol{\theta} - \boldsymbol{\vartheta}) < (\boldsymbol{a}_1 - \boldsymbol{a}_m)^\top \boldsymbol{\theta} - \varepsilon \end{cases}. \tag{42}$$

Considering that $(\boldsymbol{a}_j - \boldsymbol{a}_i)^\top \boldsymbol{\theta} - \varepsilon < 0$ and $(\boldsymbol{a}_1 - \boldsymbol{a}_m)^\top \boldsymbol{\theta} - \varepsilon > 0$, we have

$$\begin{cases} \left|(\boldsymbol{a}_j - \boldsymbol{a}_i)^\top (\boldsymbol{\theta} - \boldsymbol{\vartheta})\right| \leq \left|(\boldsymbol{a}_j - \boldsymbol{a}_i)^\top \boldsymbol{\theta} - \varepsilon\right| \\ \left|(\boldsymbol{a}_1 - \boldsymbol{a}_m)^\top (\boldsymbol{\theta} - \boldsymbol{\vartheta})\right| < \left|(\boldsymbol{a}_1 - \boldsymbol{a}_m)^\top \boldsymbol{\theta} - \varepsilon\right| \end{cases}. \tag{43}$$

Thus the confidence region can be easily constructed as

$$\mathcal{D}_{t,\delta} = \left\{ \boldsymbol{\vartheta} \in \mathbb{R}^d, \forall i \in G_\varepsilon(\nu), j \neq i, m \notin G_\varepsilon(\nu), \begin{cases} \left|(\boldsymbol{a}_j - \boldsymbol{a}_i)^\top (\boldsymbol{\theta} - \boldsymbol{\vartheta})\right| \leq \|\boldsymbol{a}_j - \boldsymbol{a}_i\|_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}} B_{t,\delta} \\ \left|(\boldsymbol{a}_1 - \boldsymbol{a}_m)^\top (\boldsymbol{\theta} - \boldsymbol{\vartheta})\right| \leq \|\boldsymbol{a}_1 - \boldsymbol{a}_m\|_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}} B_{t,\delta} \end{cases} \right\}, \tag{44}$$

where the confidence bound for parameter $\boldsymbol{\theta}$, $B_{t,\delta}$, can either be a fixed confidence bound by Proposition 1 or a looser adaptive confidence bound by Proposition 2. Thus the stopping condition $\mathcal{D}_{t,\delta} \subset U^*$ can be reformulated. For $\forall i \in G_\varepsilon(\nu), j \neq i, m \notin G_\varepsilon(\nu)$, we have

$$\begin{cases} \|\boldsymbol{a}_j - \boldsymbol{a}_i\|_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}} B_{t,\delta} \leq \left|(\boldsymbol{a}_j - \boldsymbol{a}_i)^\top \boldsymbol{\theta} - \varepsilon\right| \\ \|\boldsymbol{a}_1 - \boldsymbol{a}_m\|_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}} B_{t,\delta} \leq \left|(\boldsymbol{a}_1 - \boldsymbol{a}_m)^\top \boldsymbol{\theta} - \varepsilon\right| \end{cases}. \tag{45}$$

From this condition, the oracle allocation strategy and lower bound can be derived as

$$\boldsymbol{p}^* = \arg\min_{\boldsymbol{p} \in S_K} \max_{i \in G_\varepsilon(\nu), j \neq i, m \notin G_\varepsilon(\nu)} \max \left\{ \frac{2\|\boldsymbol{a}_i - \boldsymbol{a}_j\|_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}^2}{\left(\boldsymbol{a}_i^\top \boldsymbol{\theta} - \boldsymbol{a}_j^\top \boldsymbol{\theta} + \varepsilon\right)^2}, \frac{2\|\boldsymbol{a}_1 - \boldsymbol{a}_m\|_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}^2}{\left(\boldsymbol{a}_1^\top \boldsymbol{\theta} - \boldsymbol{a}_m^\top \boldsymbol{\theta} - \varepsilon\right)^2} \right\}, \tag{46}$$

which is exactly the same form as equation (38) in Theorem 3.

### 4.3. Upper Bounds of the LinFACTE Algorithm

The following Theorem 4 and Theorem 5 demonstrate the upper bound of the proposed **LinFACTE** algorithm. Theorem 4 is the LinFACTE algorithm with a $G$-optimal sampling policy, the upper bound which does not match the lower bound in any form. Theorem 5 is the LinFACTE algorithm with a $\mathcal{XY}$-optimal sampling policy, proved to be instance-optimal up to some logarithmic term.

THEOREM 4. *Fix $\varepsilon > 0$, let $T_G$ denote the number of samples taken based on the G-optimal design, then for universal constant $C_1$, $C_2$, and $C_3$, there exists an event $\mathcal{E}$ such that $\mathbb{P}(\mathcal{E}) \geq 1-\delta$ and on $\mathcal{E}$, LinFACTE algorithm with the G-optimal sampling policy terminates and returns $G_\varepsilon$ with an expected sample complexity upper bound, given by*

$$\mathbb{E}\left[T_G \mid \mathcal{E}\right] \leq C_1 d \frac{64}{\alpha_\varepsilon^2} \log\left(\frac{K}{\delta}\log_2\left(\frac{8}{\alpha_\varepsilon}\right)\right) + \frac{d(d+1)}{2}\log_2\frac{4}{\alpha_\varepsilon} + \sum_{i\in G_\varepsilon^c}\left(\frac{d(d+1)}{2}\log_2\left(\frac{4}{\Delta_i-\epsilon}\right) + C_2\frac{64d}{(\Delta_i-\varepsilon)^2}\log\left(\frac{K}{\delta}\log_2\frac{8}{\Delta_i-\varepsilon}\right)\right). \tag{47}$$

*Furthermore, we have a refined version that removes the summation from the upper bound, given by*

$$\mathbb{E}\left[T_G \mid \mathcal{E}\right] \leq C_3 \max\left\{\frac{64d}{\alpha_\varepsilon^2}\log\left(\frac{K}{\delta}\log_2\frac{8}{\alpha_\varepsilon}\right), \frac{64d}{\beta_\varepsilon^2}\log\left(\frac{K}{\delta}\log_2\frac{8}{\beta_\varepsilon}\right)\right\}. \tag{48}$$

However, from the proof we can see that it is impossible for the algorithm based on $G$-optimal design to have a matching upper bound in any form. Therefore, we switch our sampling policy from $G$-optimal design to $\mathcal{XY}$-optimal design to improve our algorithms.

THEOREM 5. *Fix $\varepsilon > 0$, and an instance $\nu$ of arms that $\min_{i\in G_\varepsilon\setminus\{1\}}\|\boldsymbol{a}_1-\boldsymbol{a}_i\|^2 \geq L_2$. Let $T_{\mathcal{XY}}$ denote the number of samples taken based on the $\mathcal{XY}$-optimal design and we have the following theorem and consider an $\epsilon$-efficient rounding procedure. There exists an event $\mathcal{E}$ such that $\mathbb{P}(\mathcal{E}) \geq 1-\delta$ and on $\mathcal{E}$, the LinFACTE algorithm with the $\mathcal{XY}$-optimal sampling policy terminates and returns $G_\varepsilon$. The expected sample complexity is instance-optimal up to logarithmic factors, given by*

$$\mathbb{E}\left[T_{\mathcal{XY}} \mid \mathcal{E}\right] \leq C_4\left[dR_{upper}\log\left(\frac{2K(R_{upper}+1)}{\delta}\right)\right](\Gamma^*)^{-1} + r(\epsilon)R_{upper}, \tag{49}$$

*where $C_4$ is a universal constant, $R_{upper} = \max\left\{\left\lceil\log_2\frac{4}{\alpha_\varepsilon}\right\rceil, \left\lceil\log_2\frac{4}{\beta_\varepsilon}\right\rceil\right\}$, and $(\Gamma^*)^{-1}$ is the critical term in the lower bound of linear bandit with All $\varepsilon$-Best pure exploration task, defined in Theorem 3.*

The detailed proof can be seen in Section EC.2 and in the following we give a sketch of proof.

**4.3.1. Proof Sketch of Theorem 4** The proof of Theorem 4 can be divided into two parts. As the first result in Theorem 4, equation (47) is a non-matching upper bound with summation in the result. In the second part, we provided another non-matching result in equation (48), which removes the summation from the upper bound as an improvement.

The key idea in analyzing the sample complexity is to learn and estimate the round $R_{\max}$ in which the final arm from $G_\varepsilon$ is added to $G_r$. With the estimation of $R_{\max}$, we split the total number of samples drawn as the number taken through round $R_{\max}$ and the number taken from $R_{\max}+1$ until termination if the algorithm does not terminate in round $R_{\max}$.

Steps 0 and 1 prove the correctness of LinFACTE with a G-optimal sampling procedure under a high probability event $\mathcal{E}_1$ defined as equation (EC.2.1). ~~The initial point of our analysis is the following core~~

~~decomposition in Section EC.2.0.3.~~ First, we ~~decompose xxx as xxx.~~ decompose the total number of budgets as the budgets taken through round $R_{\max}$ and the budgets taken from $R_{\max}+1$ until termination if the algorithm does not terminate in round $R_{\max}$.

$$\begin{aligned} T &\le \sum_{r=1}^{\infty} \mathbb{1}\left[G_{r-1} \cup B_{r-1} \neq [K]\right] \sum_{\boldsymbol{a}\in\mathcal{A}} T_r(\boldsymbol{a}) \\ &= \sum_{r=1}^{\infty} \mathbb{1}\left[G_{r-1} \neq G_\varepsilon\right] \mathbb{1}\left[G_{r-1} \cup B_{r-1} \neq [K]\right] \sum_{\boldsymbol{a}\in\mathcal{A}} T_r(\boldsymbol{a}) \qquad (50) \\ &+ \sum_{r=1}^{\infty} \mathbb{1}\left[G_{r-1} = G_\varepsilon\right] \mathbb{1}\left[G_{r-1} \cup B_{r-1} \neq [K]\right] \sum_{\boldsymbol{a}\in\mathcal{A}} T_r(\boldsymbol{a}). \qquad (51) \end{aligned}$$

Then we bound these two terms separately. The first term (50) is the samples before round $R_{\max}$. It can be bounded by the Lemma 1 with a clear characterization of $R_{\max}$. The second sample complexity after round $R_{\max}$, equation (51), is bounded by the Lemma 2. This lemma is proved in Section EC.2.0.7.

LEMMA 1.

$$\begin{aligned} &\sum_{r=1}^{\infty} \mathbb{1}\left[G_{r-1} \neq G_\varepsilon\right] \mathbb{1}\left[G_{r-1} \cup B_{r-1} \neq [K]\right] \sum_{\boldsymbol{a}\in\mathcal{A}} T_r(\boldsymbol{a}) \\ &\le C_1 2^{2R_{\max}+1} d \log\left(\frac{KR_{\max}(R_{\max}+1)}{\delta}\right) + \frac{d(d+1)}{2} R_{\max}. \qquad (52) \end{aligned}$$

LEMMA 2.

$$\begin{aligned} &\sum_{r=1}^{\infty} \mathbb{1}\left[G_{r-1} = G_\varepsilon\right] \mathbb{1}\left[G_{r-1} \cup B_{r-1} \neq [K]\right] \sum_{\boldsymbol{a}\in\mathcal{A}} T_r(\boldsymbol{a}) \\ &\le \sum_{i\in G_\varepsilon^c} \sum_{r=1}^{\infty} \mathbb{1}\left[i \notin B_{r-1}\right] \left(\frac{2d}{\varepsilon_r^2} \log\left(\frac{Kr(r+1)}{\delta}\right) + \frac{d(d+1)}{2}\right). \qquad (53) \end{aligned}$$

As for the second part that removes the summation, it utilizes the main results in the proof of the first part, making some refinements to the analysis to remove the summation over set $G_\varepsilon$ from the upper bound in Section EC.2.0.11. Specifically, instead of focusing on the round $R_{\max} = \min\{r : G_r = G_\varepsilon\}$ where all the arms from $G_\varepsilon$ is added to $G_r$, we concentrate on a newly defined unknown round $R_{\text{upper}} = \max\left\{\left\lceil \log_2 \frac{4}{\alpha_\varepsilon} \right\rceil, \left\lceil \log_2 \frac{4}{\beta_\varepsilon} \right\rceil\right\}$ where all the classifications have been finished and the algorithm is terminated. Following this way, the refined version without summation in Theorem 4 is acquired. (CH: This is not very clear.)(ZK L: Supplemented.)

Furthermore, we consider the relationship between the lower bound and the upper bound from another angle in Section EC.2.0.12, providing some interesting insights into the algorithm optimality. This insight is exactly the motivation for us to designate the final algorithm and theorem based on $\mathcal{XY}$-optimal design in Theorem 5.

**4.3.2. Proof Sketch of Theorem 5** In proving Theorem 5, The high probability event $\mathcal{E}_3$ is defined as equation (EC.3.1) to provide the confidence widths for the gaps of arms. Considering an $\epsilon$-efficient rounding procedure, the key idea in this proof is twofold. The first is the definition of $R_{\text{upper}}$, (CH: What is the difference to R max?)(ZK L: This question is answered and can be checked in the last comment on this page.) which represents the round where all the classifications of arms into $G_\varepsilon$ or $G_\varepsilon^c$. The second is to introduce the lower bound $(\Gamma^*)^{-1}$ from Theorem 3 to help with the derivation of our upper bound. Then, the total sample complexity in each round is bounded in Lemma 3.

LEMMA 3.

$$T_r \leq \max\left\{\left\lceil \frac{2g_{\mathcal{XY}}\left(\mathcal{Y}(G_r' \cup B_r')\right)(1+\epsilon)}{\varepsilon_r^2}\log\left(\frac{2K(K-1)r(r+1)}{\delta}\right)\right\rceil, r\left(\epsilon\right)\right\}. \tag{54}$$

Together with this lemma, the design of critical round number $R_{\text{upper}}$ and inequality (EC.2.57) make it possible to establish the matching relationship between the lower bound and our upper bound, which is in the form of being instance-optimal up to logarithmic factors.

## 5. Simulation Experiments

In this ~~session~~ section, we compare our ~~methods~~ newly proposed LinFACTE with Knowledge Gradient and BayesGap, in the context of drug development with Free-Wilson model (Free and Wilson 1964) and data given by Katz et al. (1977).

### 5.1. Experiment Setup

Free-Wilson model proposed that the overall value of the compound is a linear sum of value contributed by each substituent on the base molecule. As for each site on the base molecule, there are many candidate substituents that can be chosen from, Negoescu et al. (2011) gives a modeling method where each compound corresponds to an arm represented as a vector of indicator variables. To be specific, assuming that we have $N$ sites, each has $l_n$ ($n \in [N]$) candidate substituents. Then each arm $\boldsymbol{a}$ is a vector ~~in~~ belongs to $\mathbb{R}^{1+\sum_{n\in[N]} l_n}$, where 1 ~~is the dimension for~~ represents the intercept for the linear model. For each ~~$l_n$ number~~ period of indices with a length of $l_n$ in the arm vector, there is exactly one 1 that indicates which substituent is chosen for this site and all other $l_n - 1$ ~~numbers~~ dimensions are 0. Then, there are totally $\prod_{n\in[N]} l_n$ possible compounds.(ZK L: Maybe we can find a better way to show the structure of our drug model. It may be better to have a figure to assist our introduction.) TM: I think this will make our paper exceed page limit.(ZK L: Got it.)

We use ~~the 1-70 data~~ the data of compounds 1-70 from Katz et al. (1977), with only non-zero values remained. Then, we have 5 sites, each has 4, 5, 4, 3, 4 candidate substituents, namely there are ~~totally~~ total 960 compounds ~~and~~ with each arm $\boldsymbol{a} \in \mathbb{R}^{21}$. Our experiment is implemented on AMD Ryzen 9 5950x 16-core processor and DDR4 memory with a speed of 2133MT/s.

We compare our algorithm LinFACTE with KGCB (Negoescu et al. 2011) and BayesGap (Hoffman et al. 2014) on the number of true positive and false positive. With respect to the correct set of 20 ~~$\epsilon$~~ $\varepsilon$ arms, the proposed set of arms by algorithms can be seen as a result of classification. True positive means that the algorithm correctly propose an arm that is in the true set, and false positive means that the algorithm wrongly propose an arm that is not in the true set.

While our algorithm is ~~a~~ based on the fixed confidence with ~~non-constant~~ unfixed total budget, we fix the budget of KGCB and BayesGap to be the same as the final budget of LinFACTE in each run. All three algorithms are run for 50 times for LinFACTE failure probability $\delta$ ranging from 0.1 to 0.9 with stepsize 0.1.

### 5.2. Experiment Results

TM: Should we put the following paragraph here or above? Unfortunately, KGCB and BayesGap are too time-consuming compared to LinFACTE. As shown in Table 2, even for the largest failure probability $\delta = 0.9$, LinFACTE requires an average budget $T = 5415.25$, ~~in~~ with which BayesGap needs to run for 179.56 seconds and many hours for KGCB. Therefore, we set a running time threshold $T/100$, namely ~~both~~ all algorithms will stop ~~early~~ immediately when the running ~~for more than~~ time exceeds $T/100$ seconds and $T$ is the ~~average~~ time budget of LinFACTE under ~~each~~ different failure probability $\delta$. Our results are shown

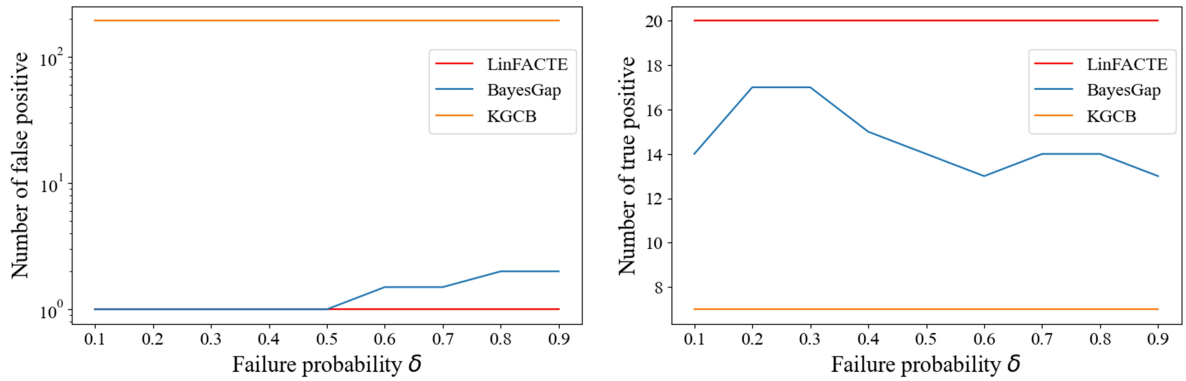
**Table 2** LinFACTE time budget and run time of algorithms changing with failure probability $\delta$

| Failure probability $\delta$ | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|---|---|---|
| LinFACTE time budget $T$ | 7249.73 | 6681.30 | 6342.81 | 6117.41 | 5909.06 | 5755.59 | 5628.74 | 5491.35 | 5415.25 |
| Run time threshold | 72.50 | 66.81 | 63.42 | 61.17 | 59.09 | 57.56 | 56.29 | 54.91 | 54.15 |
| LinFACTE run time (s) | 7.68 | 7.62 | 7.86 | 7.81 | 7.87 | 7.79 | 7.88 | 7.71 | 7.61 |
| BayesGap run time (s) | 246.13 | 222.50 | 215.94 | 207.36 | 198.01 | 190.45 | 190.70 | 188.96 | 179.56 |
| KGCB run time | More than two hours | | | | | | | | |

in Figure 3(a) and Figure 3(b). From Figure 3(a), we can see that all arms can be proposed by LinFACTE, while BayesGap and KGCB can only propose a subset. From Figure 3(b), we can see that while KGCB performs really badly, LinFACTE and BayesGap only propose ~~a~~ one wrong arm. Although these results might be due to the run time threshold on BayesGap and KGCB, our algorithm still outperforms others ~~in~~ to a certain extent.

## 6. Conclusion

(ZK L: To be finished.)




(a) Number of true positive among 20 ε-best arms (b) Number of false positive among 20 ε-best arms

**Figure 3** Number of misclassifications changing with LinFACTE failure probability $\delta$.

## References

Abbasi-Yadkori Y, Pál D, Szepesvári C (2011) Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems* 24.

Abe N, Long PM (1999) Associative reinforcement learning using linear probabilistic concepts. *ICML*, 3–11 (Citeseer).

Abernethy JD, Amin K, Zhu R (2016) Threshold bandits, with and without censored feedback. *Advances In Neural Information Processing Systems* 29.

Al Marjani A, Kocak T, Garivier A (2022) On the complexity of all $\varepsilon$-best arms identification. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 317–332 (Springer).

Allen-Zhu Z, Li Y, Singh A, Wang Y (2017) Near-optimal discrete optimization for experimental design: A regret minimization approach.

Allen-Zhu Z, Li Y, Singh A, Wang Y (2021) Near-optimal discrete optimization for experimental design: A regret minimization approach. *Mathematical Programming* 186:439–478.

Anishchenko I, Pellock SJ, Chidyausiku TM, Ramelot TA, Ovchinnikov S, Hao J, Bafna K, Norn C, Kang A, Bera AK, et al. (2021) De novo protein design by deep network hallucination. *Nature* 600(7889):547–552.

Audibert JY, Bubeck S, Munos R (2010) Best arm identification in multi-armed bandits. *COLT*, 41–53.

Auer P (2002) Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research* 3(Nov):397–422.

Azizi MJ, Kveton B, Ghavamzadeh M (2021) Fixed-budget best-arm identification in structured bandits. *arXiv preprint arXiv:2106.04763* .

Azuma K (1967) Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series* 19(3):357–367.

Bechhofer RE (1954) A single-sample multiple decision procedure for ranking means of normal populations with known variances. *The Annals of Mathematical Statistics* 16–39.

Bechhofer RE, Kiefer J, Sobel M (1968) Sequential identification and ranking procedures: with special reference to koopman-darmois populations. *(No Title)* .

Bertsimas D, Zhuo Y (2020) Novel target discovery of existing therapies: Path to personalized cancer therapy. *INFORMS Journal on Optimization* 2:ijoo.2019.0019, URL http://dx.doi.org/10.1287/ijoo.2019.0019.

Borkowski J, Pukelsheim F (1994) Optimal design of experiments. *Technometrics* 36:214, URL http://dx.doi.org/10.2307/1270234.

Bubeck S, Cesa-Bianchi N, et al. (2012) Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning* 5(1):1–122.

Bubeck S, Munos R, Stoltz G (2009) Pure exploration in multi-armed bandits problems. *Algorithmic Learning Theory: 20th International Conference, ALT 2009, Porto, Portugal, October 3-5, 2009. Proceedings 20*, 23–37 (Springer).

Bubeck S, Munos R, Stoltz G (2011) Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science* 412(19):1832–1852.

Bubeck S, Wang T, Viswanathan N (2013) Multiple identifications in multi-armed bandits. *International Conference on Machine Learning*, 258–265 (PMLR).

Cai C, Wang S, Xu Y, Zhang W, Tang K, Ouyang Q, Lai L, Pei J (2020) Transfer learning for drug discovery. *Journal of Medicinal Chemistry* 63(16):8683–8694.

Carpentier A, Locatelli A (2016) Tight (lower) bounds for the fixed budget best arm identification bandit problem. *Conference on Learning Theory*, 590–604 (PMLR).

Chaloner K, Verdinelli I (1995) Bayesian experimental design: A review. *Statistical science* 273–304.

Chen CH, Lin J, Yücesan E, Chick SE (2000) Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Discrete Event Dynamic Systems* 10:251–270.

Chen S, Lin T, King I, Lyu MR, Chen W (2014) Combinatorial pure exploration of multi-armed bandits. *Advances in neural information processing systems* 27.

Chernoff H (1959) Sequential design of experiments. *Annals of Mathematical Statistics* 30:345–360, URL https://api.semanticscholar.org/CorpusID:85510705.

Christmann-Franck S, van Westen GJ, Papadatos G, Beltran Escudie F, Roberts A, Overington JP, Domine D (2016) Unprecedently large-scale kinase inhibitor set enabling the accurate prediction of compound–kinase activities: a way toward selective promiscuity by design? *Journal of chemical information and modeling* 56(9):1654–1675.

Das P, Sercu T, Wadhawan K, Padhi I, Gehrmann S, Cipcigan F, Chenthamarakshan V, Strobelt H, Dos Santos C, Chen PY, et al. (2021a) Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nature Biomedical Engineering* 5(6):613–623.

Das P, Sercu T, Wadhawan K, Padhi I, Gehrmann S, Cipcigan F, Chenthamarakshan V, Strobelt H, Dos Santos C, Chen PY, et al. (2021b) Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nature Biomedical Engineering* 5(6):613–623.

Degenne R, Koolen WM, Ménard P (2019) Non-asymptotic pure exploration by solving games. *Advances in Neural Information Processing Systems* 32.

Doob, Joseph L (1990) *Stochastic processes* (New York: Wiley).

Even-Dar E, Mannor S, Mansour Y, Mahadevan S (2006) Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research* 7(6).

Fan W, Hong LJ, Nelson BL (2016) Indifference-zone-free selection of the best. *Operations Research* 64(6):1499–1514.

Fiez T, Jain L, Jamieson KG, Ratliff L (2019) Sequential experimental design for transductive linear bandits. *Advances in neural information processing systems* 32.

Foster DJ, Gentile C, Mohri M, Zimmert J (2020) Adapting to misspecification in contextual bandits. *Advances in Neural Information Processing Systems* 33:11478–11489.

Frazier PI (2018) A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811* .

Frazier PI, Powell WB, Dayanik S (2008) A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization* 47(5):2410–2439.

Frazier PI, Wang J (2016) Bayesian optimization for materials design. *Information science for materials discovery and design*, 45–75 (Springer).

Free SM, Wilson JW (1964) A mathematical contribution to structure-activity studies. *Journal of medicinal chemistry* 7(4):395–399.

Gabillon V, Ghavamzadeh M, Lazaric A (2012) Best arm identification: A unified approach to fixed budget and fixed confidence. *Advances in Neural Information Processing Systems* 25.

Gabillon V, Ghavamzadeh M, Lazaric A, Bubeck S (2011) Multi-bandit best arm identification. *Advances in Neural Information Processing Systems* 24.

Garivier A, Kaufmann E (2016) Optimal best arm identification with fixed confidence. *Conference on Learning Theory*, 998–1027 (PMLR).

Garivier A, Kaufmann E (2021) Nonasymptotic sequential tests for overlapping hypotheses applied to near-optimal arm identification in bandit models. *Sequential Analysis* 40(1):61–96.

Ghosh A, Chowdhury SR, Gopalan A (2017) Misspecified linear bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Hoffman M, Shahriari B, Freitas N (2014) On correlation and budget constraints in model-based bandit optimization with application to automatic machine learning. *Artificial Intelligence and Statistics*, 365–374 (PMLR).

Hong LJ, Fan W, Luo J (2021) Review on ranking and selection: A new perspective. *Frontiers of Engineering Management* 8(3):321–343.

Jourdan M, Degenne R, Kaufmann E (2024) An $\varepsilon$-best-arm identification algorithm for fixed-confidence and beyond. *Advances in Neural Information Processing Systems* 36.

Juneja S, Krishnasamy S (2019) Sample complexity of partition identification using multi-armed bandits. *Conference on Learning Theory*, 1824–1852 (PMLR).

Kalyanakrishnan S, Stone P (2010) Efficient selection of multiple bandit arms: Theory and practice. *ICML*, volume 10, 511–518.

Kalyanakrishnan S, Tewari A, Auer P, Stone P (2012) Pac subset selection in stochastic multi-armed bandits. *ICML*, volume 12, 655–662.

Katz R, Osborne SF, Ionescu F (1977) Application of the free-wilson technique to structurally related series of homologs. quantitative structure-activity relationship studies of narcotic analgetics. *Journal of Medicinal Chemistry* 20(11):1413–1419.

Katz-Samuels J, Jain L, Jamieson KG, et al. (2020) An empirical process approach to the union bound: Practical algorithms for combinatorial and linear bandits. *Advances in Neural Information Processing Systems* 33:10371–10382.

Kaufmann E, Cappé O, Garivier A (2016) On the complexity of best arm identification in multi-armed bandit models. *Journal of Machine Learning Research* 17:1–42.

Kaufmann E, Kalyanakrishnan S (2013) Information complexity in bandit subset selection. *Conference on Learning Theory*, 228–251 (PMLR).

Kaufmann E, Koolen WM (2021) Mixture martingales revisited with applications to sequential tests and confidence intervals. *The Journal of Machine Learning Research* 22(1):11140–11183.

Kiefer J, Wolfowitz J (1960) The equivalence of two extremum problems. *Canadian Journal of Mathematics* 12:363–366.

Kim SH, Nelson BL (2001) A fully sequential procedure for indifference-zone selection in simulation. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 11(3):251–273.

Koenig LW, Law AM (1985) A procedure for selecting a subset of size m containing the l best of k independent normal populations, with applications to simulation. *Communications in Statistics-Simulation and Computation* 14(3):719–734.

Lattimore T, Szepesvári C (2020) *Bandit algorithms* (Cambridge University Press).

Li Y, Reyes KG, Vazquez-Anderson J, Wang Y, Contreras LM, Powell WB (2018) A knowledge gradient policy for sequencing experiments to identify the structure of rna molecules using a sparse additive belief model. *INFORMS Journal on Computing* 30(4):750–767.

Locatelli A, Gutzeit M, Carpentier A (2016) An optimal algorithm for the thresholding bandit problem. *International Conference on Machine Learning*, 1690–1698 (PMLR).

Lou B, Wu L (2021) Ai on drugs: can artificial intelligence accelerate drug development? evidence from a large-scale examination of bio-pharma firms. *MIS Quarterly* 45(3).

Mannor S, Tsitsiklis JN (2004) The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research* 5(Jun):623–648.

Marjani AA, Kocák T, Garivier A (2022) On the complexity of all $\varepsilon$-best arms identification.

Mason B, Jain L, Tripathy A, Nowak R (2020) Finding all $\epsilon$-good arms in stochastic bandits. *Advances in Neural Information Processing Systems* 33:20707–20718.

Michel T, Hajiabolhassan H, Ortner R (2023) Regret bounds for satisficing in multi-armed bandit problems. *Transactions on Machine Learning Research* .

Millan MJ (2006) Multi-target strategies for the improved treatment of depressive states: conceptual foundations and neuronal substrates, drug discovery and therapeutic application. *Pharmacology & therapeutics* 110(2):135–370.

Mouchlis VD, Afantitis A, Serra A, Fratello M, Papadiamantis AG, Aidinis V, Lynch I, Greco D, Melagraki G (2021) Advances in de novo drug design: From conventional to machine learning methods. *International journal of molecular sciences* 22(4):1676.

Nasrollahzadeh AA, Khademi A (2022) Dynamic programming for response-adaptive dose-finding clinical trials. *INFORMS Journal on Computing* 34(2):1176–1190.

Negoescu DM, Frazier PI, Powell WB (2011) The knowledge-gradient algorithm for sequencing experiments in drug discovery. *INFORMS Journal on Computing* 23(3):346–363.

Pacchiano A, Phan M, Abbasi Yadkori Y, Rao A, Zimmert J, Lattimore T, Szepesvari C (2020) Model selection in contextual stochastic bandit problems. *Advances in Neural Information Processing Systems* 33:10328–10337.

Paulson E (1964) A sequential procedure for selecting the population with the largest mean from k normal populations. *The Annals of Mathematical Statistics* 174–180.

Powell WB, Ryzhov IO (2012) *Optimal learning*, volume 841 (John Wiley & Sons).

Pukelsheim F (2006) *Optimal design of experiments* (SIAM).

Qin C (2022) Open problem: Optimal best arm identification with fixed-budget. *Conference on Learning Theory*, 5650–5654 (PMLR).

Qin C, You W (2023) Dual-directed algorithm design for efficient pure exploration. *arXiv preprint arXiv:2310.19319* .

Ravina E (2011) *The evolution of drug discovery: from traditional medicines to modern drugs* (John Wiley & Sons).

Réda C, Kaufmann E, Delahaye-Duriez A (2020) Machine learning applications in drug development. *Computational and structural biotechnology journal* 18:241–252.

Réda C, Tirinzoni A, Degenne R (2021) Dealing with misspecification in fixed-confidence linear top-m identification. *Advances in Neural Information Processing Systems* 34:25489–25501.

Reverdy P, Srivastava V, Leonard NE (2016) Satisficing in multi-armed bandit problems. *IEEE Transactions on Automatic Control* 62(8):3788–3803.

Ringel MS, Scannell JW, Baedeker M, Schulze U (2020) Breaking eroom's law. *Nature Reviews Drug Discovery* 19(12):833–835.

Robbins H (1952) Some aspects of the sequential design of experiments .

Robbins H (1970) Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics* 41(5):1397–1409, ISSN 00034851, URL `http://www.jstor.org/stable/2239848`.

Siegmund D, Siegmund D (1985) The sequential probability ratio test. *Sequential Analysis: Tests and Confidence Intervals* 8–33.

Silvey S, Sibson B (1972) Discussion of dr. wynn's and of dr. laycock's papers. *Journal of Royal Statistical Society (B)* 34(174-175):270.

Simchowitz M, Jamieson K, Recht B (2017) The simulator: Understanding adaptive sampling in the moderate-confidence regime. *Conference on Learning Theory*, 1794–1834 (PMLR).

Simon HA (1955) A behavioral model of rational choice. *The quarterly journal of economics* 99–118.

Soare M, Lazaric A, Munos R (2014) Best-arm identification in linear bandits. *Advances in Neural Information Processing Systems* 27.

Takemura K, Ito S, Hatano D, Sumita H, Fukunaga T, Kakimura N, Kawarabayashi Ki (2021) A parameter-free algorithm for misspecified linear contextual bandits. *International Conference on Artificial Intelligence and Statistics*, 3367–3375 (PMLR).

Tallorin L, Wang J, Kim WE, Sahu S, Kosa NM, Yang P, Thompson M, Gilson MK, Frazier PI, Burkart MD, et al. (2018) Discovering de novo peptide substrates for enzymes using machine learning. *Nature communications* 9(1):1–10.

Tao C, Blanco S, Zhou Y (2018) Best arm identification in linear bandits with linear dimension dependency. *International Conference on Machine Learning*, 4877–4886 (PMLR).

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

Todd MJ (2016) *Minimum-volume ellipsoids: Theory and algorithms* (SIAM).

Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, Li B, Madabhushi A, Shah P, Spitzer M, et al. (2019) Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery* 18(6):463–477.

Wald A (1992) Sequential tests of statistical hypotheses. *Breakthroughs in statistics: Foundations and basic theory*, 256–298 (Springer).

Wang PA, Tzeng RC, Proutiere A (2021) Fast pure exploration via frank-wolfe. *Advances in Neural Information Processing Systems* 34:5810–5821.

Xu L, Honda J, Sugiyama M (2018) A fully adaptive algorithm for pure exploration in linear bandits. *International Conference on Artificial Intelligence and Statistics*, 843–851 (PMLR).

Yang J, Tan V (2022) Minimax optimal fixed-budget best arm identification in linear bandits. Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K, Oh A, eds., *Advances in Neural Information Processing Systems*, volume 35, 12253–12266 (Curran Associates, Inc.), URL https://proceedings.neurips.cc/paper_files/paper/2022/file/4f9342b74c3bb63f6e030d8263082ab6-Paper-Conference.pdf.

Yoshida M, Hinkley T, Tsuda S, Abul-Haija YM, McBurney RT, Kulikov V, Mathieson JS, Galiñanes Reyes S, Castro MD, Cronin L (2018) Using evolutionary algorithms and machine learning to explore sequence space for the discovery of antimicrobial peptides. *Chem* 4(3):533–543, ISSN 2451-9294, URL http://dx.doi.org/https://doi.org/10.1016/j.chempr.2018.01.005.

Yu K, Bi J, Tresp V (2006) Active learning via transductive experimental design. *Proceedings of the 23rd international conference on Machine learning*, 1081–1088.

Zaki M, Mohan A, Gopalan A (2020) Explicit best arm identification in linear bandits using no-regret learners. *arXiv preprint arXiv:2006.07562* .

Zhang W, He J, Fan Z, Gu Q (2023) On the interplay between misspecification and sub-optimality gap in linear contextual bandits. *arXiv preprint arXiv:2303.09390* .

# E-Companion —XX

## EC.1. Proof of Problem Complexity and Lower Bound in Theorem 3

Proof. For All $\varepsilon$-Best Arms Identification in the Linear Bandit, the definition of the correct answer $\mathcal{I}$, the culprit set $\mathcal{X}(\boldsymbol{\mu})$, and the alternative set $Alt(\boldsymbol{\mu})$ are established as follows.

1. The correct answer $G_\varepsilon(\nu) \triangleq \{i : \langle \boldsymbol{\theta}, \boldsymbol{a}_i \rangle \geq \max_i \langle \boldsymbol{\theta}, \boldsymbol{a}_i \rangle - \varepsilon\}$ for all $i \in [K]$.
2. The culprit set $\mathcal{X}(\boldsymbol{\mu}) = \{(i, j, m, \ell) : i \in G_\varepsilon(\nu), j \neq i, m \notin G_\varepsilon(\nu), \ell \in [|1, m-1|]\}$.
3. The alternative set $Alt(\boldsymbol{\mu}) = \cup_{x \in \mathcal{X}(\boldsymbol{\mu})} Alt_x(\boldsymbol{\mu})$ and the form of $\text{Alt}_x(\boldsymbol{\mu})$ is given by

$$\text{Alt}_x(\boldsymbol{\mu}) = \text{Alt}_{i,j}(\boldsymbol{\mu}) \cup \text{Alt}_{m,\ell}(\boldsymbol{\mu}) \text{ for all } x \in \mathcal{X}(\boldsymbol{\mu}). \tag{EC.1.1}$$

As stated in Theorem 2, two distinct parts of the alternative sets can be individually represented as

$$\text{Alt}_{i,j}(\boldsymbol{\mu}) = \text{Alt}_{i,j}(\boldsymbol{\theta}) = \{\boldsymbol{\theta} : \langle \boldsymbol{\theta}, \boldsymbol{a}_i - \boldsymbol{a}_j \rangle < -\varepsilon\} \text{ for all } x \in \mathcal{X}(\boldsymbol{\mu}) \tag{EC.1.2}$$

and

$$\text{Alt}_{m,\ell}(\boldsymbol{\mu}) = \left\{\boldsymbol{\mu} : \mu_\ell \geq \mu_\varepsilon^{m,\ell} p + \varepsilon \geq \mu_{\ell+1}\right\} \text{ for all } x \in \mathcal{X}(\boldsymbol{\mu}), \tag{EC.1.3}$$

$$\text{Alt}_{m,\ell}(\boldsymbol{\mu}) = \text{Alt}_{m,\ell}(\boldsymbol{\theta}) = \left\{\boldsymbol{\theta} : \boldsymbol{\theta}^\top \boldsymbol{a}_\ell \geq \mu_\varepsilon^{m,\ell} p + \varepsilon \geq \boldsymbol{\theta}^\top \boldsymbol{a}_{\ell+1}\right\} \text{ for all } x \in \mathcal{X}(\boldsymbol{\mu}), \tag{EC.1.4}$$

where

$$\mu_\varepsilon^{m,\ell}(p) \triangleq \frac{p_m \mu_m + \sum_{i=1}^{\ell} p_i (\mu_i - \varepsilon)}{p_m + \sum_{i=1}^{\ell} p_i}. \tag{EC.1.5}$$

Considering that there are two ways of constructing the alternative set, the alternative set $\boldsymbol{\vartheta}_{(i,j,m,\ell)}$ is composed of two distinct parts, i.e. $\boldsymbol{\vartheta}_{(i,j)}$ and $\boldsymbol{\vartheta}_{(m,\ell)}$.

The first part is

$$\boldsymbol{\vartheta}_{(i,j)}(\varepsilon, p, \alpha) = \boldsymbol{\theta} - \frac{\boldsymbol{y}_{i,j}^\top \boldsymbol{\theta} + \varepsilon + \alpha}{\|\boldsymbol{y}_{i,j}\|_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}^2} \boldsymbol{V}_{\boldsymbol{p}}^{-1} \boldsymbol{y}_{i,j}, \tag{EC.1.6}$$

where $\boldsymbol{V}_{\boldsymbol{p}} = \sum_{i=1}^{K} p_i \boldsymbol{a}_i \boldsymbol{a}_i^\top$ is closely related to the Fisher information matrix (Chaloner and Verdinelli 1995) and ~~$\|\boldsymbol{a}_{I^*} - \boldsymbol{a}_j\|_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}^2$~~ $\|\boldsymbol{y}_{i,j}\|_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}^2$ is a Mahalanobis norm(ZK L: It is indeed in the form of Mahalanobis distance but still a little different.) defined in terms of the matrix $\boldsymbol{V}_{\boldsymbol{p}}^{-1}$. (CH: Please briefly explain what does $\boldsymbol{\vartheta}_j$ mean here. )(ZK L: It has been explained in the following equation (EC.1.11) and equation (EC.1.12).) $\boldsymbol{y}_{i,j} = \boldsymbol{a}_i - \boldsymbol{a}_j$ and $\alpha > 0$. Then we have

$$\boldsymbol{y}_{i,j}^\top \boldsymbol{\vartheta}_{(i,j)}(\varepsilon, p, \alpha) = -\varepsilon - \alpha < -\varepsilon. \tag{EC.1.7}$$

Without loss of generality, we still assume the Gaussian distribution, i.e. $\mu_i \sim \mathcal{N}(\boldsymbol{a}_i^\top \boldsymbol{\theta}, 1)$. Then the KL divergence between the mean value based on the true parameter $\boldsymbol{\theta}$ against the mean value based on the alternative parameter $\boldsymbol{\vartheta}_{(i,j)}$ is

$$\begin{aligned}\mathrm{KL}(\boldsymbol{a}_i^\top \boldsymbol{\theta}, \boldsymbol{a}_i^\top \boldsymbol{\vartheta}_{(i,j)}) &= \frac{\left(\boldsymbol{a}_i^\top \left(\boldsymbol{\theta} - \boldsymbol{\vartheta}_{(i,j)}(\varepsilon, p, \alpha)\right)\right)^2}{2(1)^2} \\ &= \boldsymbol{y}_{i,j}^\top \boldsymbol{V}_{\boldsymbol{p}}^{-1} \frac{\left(\boldsymbol{y}_{i,j}^\top \boldsymbol{\theta} + \varepsilon + \alpha\right)^2 \boldsymbol{a}_i \boldsymbol{a}_i^\top}{2\left(\|\boldsymbol{y}_{i,j}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}\right)^2} \boldsymbol{V}_{\boldsymbol{p}}^{-1} \boldsymbol{y}_{i,j}. \end{aligned} \tag{EC.1.8}$$

Then by Proposition 4, the lower bound can be calculated as

$$\begin{aligned}\frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/2.4\delta)} &\geq \min_{\boldsymbol{p} \in S_K} \max_{\boldsymbol{\vartheta} \in Alt(\boldsymbol{\mu})} \frac{1}{\sum_{n=1}^K p_n \mathrm{KL}(\boldsymbol{a}_n^\top \boldsymbol{\theta}, \boldsymbol{a}_n^\top \boldsymbol{\vartheta})} \\ &\geq \min_{\boldsymbol{p} \in S_K} \max_{(i,j,m) \in \mathcal{X}} \sup_{\alpha > 0} \frac{1}{\sum_{n=1}^K p_n \mathrm{KL}(\boldsymbol{a}_n^\top \boldsymbol{\theta}, \boldsymbol{a}_n^\top \boldsymbol{\vartheta}_{(i,j)}(\varepsilon, p, \alpha))} \\ &= \min_{\boldsymbol{p} \in S_K} \max_{(i,j,m) \in \mathcal{X}} \frac{2\|\boldsymbol{y}_{i,j}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}}{\left(\boldsymbol{y}_{i,j}^\top \boldsymbol{\theta} + \varepsilon\right)^2}. \end{aligned} \tag{EC.1.9}$$

From this lower bound we can also define the $C_x$ function for All $\varepsilon$-Best Arms Identification in Linear Bandit, which is

$$C_{i,j}(p) = \frac{\left(\boldsymbol{y}_{i,j}^\top \boldsymbol{\theta} + \varepsilon\right)^2}{2\|\boldsymbol{y}_{i,j}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}}. \tag{EC.1.10}$$

Notice that we let $\alpha \to 0$ establish the result by realizing $\sup_{\alpha > 0}$. Furthermore, it needs to be mentioned that the form of the alternative set in the first part actually comes from the solution of an optimization problem, given by

$$\begin{aligned} \underset{\boldsymbol{\vartheta} \in \mathbb{R}^d}{\arg\min} \quad & \|\boldsymbol{\vartheta} - \boldsymbol{\theta}\|^2_{V_p} && \text{(EC.1.11)} \\ \text{s.t.} \quad & \boldsymbol{y}_{i,j}^\top \boldsymbol{\vartheta} = -\varepsilon - \alpha. && \text{(EC.1.12)} \end{aligned}$$

Then for the second part, if we are considering a simpler case, following the same way of derivation, we have the second part of the alternative set as

$$\boldsymbol{\vartheta}_{(m)}(\varepsilon, p, \alpha) = \boldsymbol{\theta} - \frac{\boldsymbol{y}_m^\top \boldsymbol{\theta} + \varepsilon + \alpha}{\|\boldsymbol{y}_m\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}} \boldsymbol{V}_{\boldsymbol{p}}^{-1} \boldsymbol{y}_m, \tag{EC.1.13}$$

where $\boldsymbol{y}_m = \boldsymbol{a}_1 - \boldsymbol{a}_m$ and $\alpha > 0$. Then we have

$$\boldsymbol{y}_m^\top \boldsymbol{\vartheta}_{(m)}(\varepsilon, p, \alpha) = \varepsilon - \alpha < \varepsilon. \tag{EC.1.14}$$

Hence,

$$C_m(p) = \frac{\left(\boldsymbol{y}_m^\top \boldsymbol{\theta} - \varepsilon\right)^2}{2\|\boldsymbol{y}_m\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}}. \tag{EC.1.15}$$

Then similarly by the Proposition 4, combining equation (EC.1.10) and equation (EC.1.15), the final lower bound can be calculated as

$$\frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/2.4\delta)} \geq \min_{\boldsymbol{p}\in S_K} \max_{(i,j,m)\in\mathcal{X}} \max\left\{ \frac{2\|\boldsymbol{y}_{i,j}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}}{\left(\boldsymbol{y}_{i,j}^\top\boldsymbol{\theta}+\varepsilon\right)^2}, \frac{2\|\boldsymbol{y}_m\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}}{\left(\boldsymbol{y}_m^\top\boldsymbol{\theta}-\varepsilon\right)^2} \right\} \tag{EC.1.16}$$

$$= \min_{\boldsymbol{p}\in S_K} \max_{(i,j,m)\in\mathcal{X}} \max\left\{ \frac{2\|\boldsymbol{a}_i-\boldsymbol{a}_j\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}}{\left(\boldsymbol{a}_i^\top\boldsymbol{\theta}-\boldsymbol{a}_j^\top\boldsymbol{\theta}+\varepsilon\right)^2}, \frac{2\|\boldsymbol{a}_1-\boldsymbol{a}_m\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}}{\left(\boldsymbol{a}_1^\top\boldsymbol{\theta}-\boldsymbol{a}_m^\top\boldsymbol{\theta}-\varepsilon\right)^2} \right\}. \tag{EC.1.17}$$

□

## EC.2. Proof of Theorem 4

The complete proof of Theorem 4 can be divided into two parts. In the first part, we provide the first proof based on $G$-optimal design, providing the result of the equation (47), which is a non-matching upper bound with summation in the result. In the second part, we find another angle to prove our upper bounds. Though still non-matching, the result in equation (48) removes the summation from the upper bound.

The overall idea and the structure of deriving the first part result are concluded below. The unknown round $R_{\max}$ is the round in which the final arm from $G_\varepsilon$ is added to $G_r$. We may split the total number of samples drawn as the number taken through round $R_{\max}$ and the number taken from $R_{\max}+1$ until termination if the algorithm does not terminate in round $R_{\max}$. The proof is split into eight steps detailed as follows.

#### EC.2.0.1. Preliminary: Clean Event $\mathcal{E}_1$

Proof. We give the proof process for the $G$-optimal design. The initial point of the algorithm design and the proof is the definition of the events $\mathcal{E}_1$ below.

$$\mathcal{E}_1 = \left\{ \bigcap_{i\in\mathcal{A}_I(r)} \bigcap_{r\in\mathbb{N}} |\hat{\mu}_i(r)-\mu_i| \leq C_{\delta/K}(r) \right\}. \tag{EC.2.1}$$

Considering that the arm is sampled based on the preset allocation, i.e. the fixed design, here we introduce the following lemma to give a simple confidence region of estimated parameter $\boldsymbol{\theta}$.

LEMMA EC.2.1. *Let $\delta\in(0,1)$, it holds that for each arm $\boldsymbol{a}\in\mathcal{A}$, we have*

$$\mathbb{P}\left\{ |\langle\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}^*,\boldsymbol{a}\rangle| \geq \sqrt{2\|\boldsymbol{a}\|^2_{\boldsymbol{V}^{-1}}\log\left(\frac{1}{\delta}\right)} \right\} \leq \delta., \tag{EC.2.2}$$

If we follow the $G$-optimal sampling rule as stated in line 13 and line 15 in the pseudocode, for each round we have

$$\boldsymbol{V} = \sum_{\boldsymbol{a}\in\text{Supp}(\pi_r)} T_r(\boldsymbol{a})\boldsymbol{a}\boldsymbol{a}^\top \succeq \frac{2d}{\varepsilon_r^2}\log\left(\frac{Kr(r+1)}{\delta}\right)\boldsymbol{V}(\pi). \tag{EC.2.3}$$

Thus with the inequality in lemma EC.2.1 and the result in Lemma EC.4.2, for any arm $\boldsymbol{a} \in \mathcal{A}$, with probability at least $1-\delta/Kr(r+1)$, we have

$$\begin{aligned}\left|\langle \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*, \boldsymbol{a}\rangle\right| &\leq \sqrt{2\|\boldsymbol{a}\|^2_{\boldsymbol{V}^{-1}} \log\left(\frac{Kr(r+1)}{\delta}\right)} \\ &= \sqrt{2\boldsymbol{a}^\top \boldsymbol{V}^{-1}\boldsymbol{a} \log\left(\frac{Kr(r+1)}{\delta}\right)} \\ &\leq \sqrt{2\boldsymbol{a}^\top \left(\frac{\varepsilon_r^2}{2d}\frac{1}{\log\left(\frac{Kr(r+1)}{\delta}\right)}\boldsymbol{V}(\pi)^{-1}\right)\boldsymbol{a} \log\left(\frac{Kr(r+1)}{\delta}\right)} \\ &\leq \varepsilon_r. \end{aligned} \tag{EC.2.4}$$

Thus with the standard result of the $G$-optimal design, we have

$$C_{\delta/K}(r) \triangleq \varepsilon_r. \tag{EC.2.5}$$

To give the probability guarantee, we have

$$\begin{aligned}\mathbb{P}(\mathcal{E}_1^c) &= \mathbb{P}\left\{\bigcup_{i\in\mathcal{A}_I(r)}\bigcup_{r\in\mathbb{N}} |\hat{\mu}_i(r) - \mu_i| > C_{\delta/K}(r)\right\} \\ &\leq \sum_{r=1}^{\infty}\mathbb{P}\left\{\bigcup_{i\in\mathcal{A}_I(r)} |\hat{\mu}_i(r) - \mu_i| > C_{\delta/K}(r)\right\} \\ &\leq \sum_{r=1}^{\infty}\sum_{i=1}^{K}\frac{\delta}{Kr(r+1)} \\ &= \delta. \end{aligned} \tag{EC.2.6}$$

Therefore, considering the union bounds over the rounds $r \in \mathbb{N}$, we have

$$P(\mathcal{E}_1) \geq 1-\delta. \tag{EC.2.7}$$

Considering another event, given by

$$\mathcal{E}_2 = \bigcap_{i\in G_\varepsilon}\bigcap_{j\in\mathcal{A}_I}\bigcap_{r\in\mathbb{N}} |(\hat{\mu}_j(r) - \hat{\mu}_i(r)) - (\mu_j - \mu_i)| \leq 2\varepsilon_r. \tag{EC.2.8}$$

By equation (EC.2.4), for $i, j \in \mathcal{A}_I(r)$, we have

$$\begin{aligned}\mathbb{P}\left\{|(\hat{\mu}_j - \hat{\mu}_i) - (\mu_j - \mu_i)| > 2\varepsilon_r \mid \mathcal{E}_1\right\} &\leq \mathbb{P}\left\{|\hat{\mu}_j - \mu_j| + |\hat{\mu}_i - \mu_i| > 2\varepsilon_r \mid \mathcal{E}_1\right\} \\ &\leq \mathbb{P}\left\{|\hat{\mu}_j - \mu_j| > \varepsilon_r \mid \mathcal{E}_1\right\} + \mathbb{P}\left\{|\hat{\mu}_i - \mu_i| > \varepsilon_r \mid \mathcal{E}_1\right\} \\ &= 0, \end{aligned} \tag{EC.2.9}$$

which means

$$\mathbb{P}(\mathcal{E}_2 \mid \mathcal{E}_1) = 1. \tag{EC.2.10}$$

□

#### EC.2.0.2. Step 0: Correctness

On $\mathcal{E}_1$, we first prove that if there exist a random round $r$, at which $G_r \cup B_r = [K]$, then $G_r = G_\varepsilon$. Besides, we also prove that on $\mathcal{E}_1$, if $\mathcal{A}_I \subset G_r$, then $G_r = G_\varepsilon$. As we can see, for either of these two stopping conditions of LinFACTE in line 27, the LinFACTE algorithm will return the set $G_\varepsilon$ correctly on the clean event $\mathcal{E}_1$.

To prove the correctness of the first condition, we need the following **Claim 0** and **Claim 1**.

**Claim 0: On $\mathcal{E}_1$, for all $r \in \mathbb{N}$, $G_r \subset G_\varepsilon$**

Proof. Firstly we show $1 \in \mathcal{A}_I$ for all $r \in \mathbb{N}$, that is, the best arm is never removed from $\mathcal{A}$. Note for any arm $i$

$$\hat{\mu}_1(r) + C_{\delta/K}(r) \geq \mu_1 \geq \mu_i \geq \hat{\mu}_i(r) - C_{\delta/K}(r) > \hat{\mu}_i(r) - C_{\delta/K}(r) - \varepsilon, \tag{EC.2.11}$$

which particularly shows that $\hat{\mu}_1(r) + C_{\delta/K}(r) > \max_{i\in\mathcal{A}_I} \hat{\mu}_i - C_{\delta/K}(r) - \varepsilon = L_r$ and $\hat{\mu}_1(r) + C_{\delta/K}(r) \geq \max_{i\in\mathcal{A}} \hat{\mu}_i(r) - C_{\delta/K}$ showing that 1 will never exit $\mathcal{A}_I$ in line 22 or line 24.

Secondly, we show that at all rounds $r$, $\mu_1 - \varepsilon \in [L_r, U_r]$. Since $\mu_1$ never exists $\mathcal{A}_I$,

$$U_r = \max_{i\in\mathcal{A}_I} \hat{\mu}_i + C_{\delta/K}(r) - \varepsilon \geq \hat{\mu}_1(r) + C_{\delta/K}(r) - \varepsilon \geq \mu_1 - \varepsilon, \tag{EC.2.12}$$

and for any $i$,

$$\mu_1 - \varepsilon \geq \mu_i - \varepsilon \geq \hat{\mu}_i - C_{\delta/K}(r) - \varepsilon. \tag{EC.2.13}$$

Hence,

$$\mu_1 - \varepsilon \geq \max_i \hat{\mu}_i - C_{\delta/K}(r) - \varepsilon = L_r. \tag{EC.2.14}$$

Next, we show that $G_r \subset G_\varepsilon$ for all $r \geq 1$. Suppose not, then it means that $\exists r \in \mathbb{N}$ and $\exists i \in G_\varepsilon^c \cap G_r$ such that,

$$\mu_i \geq \hat{\mu}_i - C_{\delta/K}(r) \geq U_r \geq \mu_1 - \varepsilon > \mu_i, \tag{EC.2.15}$$

at which the last inequality gives the contradiction. □

**Claim 1: On $\mathcal{E}_1$, for all $r \in \mathbb{N}$, $B_r \subset G_\varepsilon^c$**

Proof. Next, we will show that $B_r \subset G_\varepsilon^c$. Suppose not, then a good arm was added to the bad set by the arm filter. Consider this case that the arm filter added an arm in $G_\varepsilon$ to $B_r$ for some $r$. By definition, $B_0 = \emptyset$ and $B_{r-1} \subset B_r$ for all $r$. Then there must exist some $r \in \mathbb{N}$ and an $i \in G_\varepsilon$ such that $i \in B_r$ and $i \notin B_{r-1}$. Following the line 22 of the algorithm, this occurs if and only if

$$\max_{j\in\mathcal{A}_I} \hat{\mu}_j - \hat{\mu}_i > 2\varepsilon_r + \varepsilon. \tag{EC.2.16}$$

On the clean event $\mathcal{E}_1$, the above implies $\exists j \in \mathcal{A}_I$ such that

$$\mu_j - \mu_i + 2\varepsilon_r \geq \hat{\mu}_j - \hat{\mu}_i \geq 2\varepsilon_r + \varepsilon, \tag{EC.2.17}$$

which shows that $\mu_j - \mu_i \geq \varepsilon$, contradicting the assertion that $i \in G_\varepsilon$. □

The above **Claim 0** and **Claim 1** shows that under $\mathcal{E}_1$, $G_r \cup B_r = [K]$ can lead to the result $G_r = G_\varepsilon$. Since $\mathbb{P}\{\mathcal{E}_1\} \geq 1-\delta$, if LinFACTE terminates, it can correctly provide the correct decision rule with a probability as least $1-\delta$, being a $\delta$-PAC algorithm.

**Claim 2: On $\mathcal{E}_1$, if $U_r - L_r \leq \gamma/2$, then $\mathcal{A}_I \cup G_r \subset G_{\varepsilon+\gamma}$**

Proof. Assume $U_r - L_r \leq \gamma/2$, this implies that

$$U_r - L_r = \left(\max_{i\in\mathcal{A}_I} \hat{\mu}_i + C_{\delta/K}(r) - \varepsilon\right) - \left(\max_{i\in\mathcal{A}_I} \hat{\mu}_i - C_{\delta/K}(r) - \varepsilon\right) = 2C_{\delta/K}(r) \leq \gamma/2. \tag{EC.2.18}$$

Suppose for contradiction that there exists $i \in G^c_{\varepsilon+\gamma}$ such that $i \in \mathcal{A}_I \cup G_r$. Since $G_\varepsilon \cap G^c_{\varepsilon+\gamma} = \emptyset$ and we have previously shown that $G_r \subset G_\varepsilon$ for all $r$, we have that $i \in \mathcal{A}_I \setminus G_r$.

Therefore, by the condition in line 22, $\hat{\mu}_i + C_{\delta/K}(r) > L_r$. Hence

$$\mu_i + 2C_{\delta/K}(r) \geq \hat{\mu}_i + C_{\delta/K}(r) > L_r. \tag{EC.2.19}$$

By assumption, we have that $U_r - \gamma/2 \leq L_r$. The event $\mathcal{E}_1$ implies that $U_r \geq \mu_1 - \varepsilon$. Therefore, $\mu_i + 2C_{\delta/K}(r) > U_r - \gamma/2 \geq \mu_1 - \varepsilon - \gamma/2$. Combining this with the inequality $2C_{\delta/K}(r) \leq \gamma/2$, we have that

$$\gamma \geq 2C_{\delta/K}(r) + \gamma/2 \geq \mu_1 - \varepsilon - \mu_i > \gamma, \tag{EC.2.20}$$

which is a contradiction. □

Up to now, we have proved the correctness of the stopping rule of the algorithm. Then we will bound the sample complexity in the following parts.

**EC.2.0.3. Step 1: An expression for the total number of samples drawn and introducing several helper random variables**

We write an expression for the total number of samples drawn by LinFACTE. Specifically, we introduce several sums that we will use in the remainder of the proof controlling the sample budget.

For $i \in G_\varepsilon$, let $R_i$ denote the random variable of the number of rounds arm $i$ is sampled by the arm filter when it is added to $G_r$ in line 20. For $i \in G^c_\varepsilon$, let $R_i$ denote the random variable of the number of rounds arm $i$ is sampled by the arm filter when it is removed from $\mathcal{A}_I$ and added to $B_r$ in line 22. ~~For any arm $i$, let $R'_i$ denote the random variable of the number of rounds $i$ is sampled by the arm filter when $\hat{\mu}_i(t) + C_{\delta/K}(r) \leq \max_{j\in\mathcal{A}_I} \hat{\mu}_j(t) - C_{\delta/K}(r)$. Lastly,~~ Let $R_\gamma$ denote the random variable of the number of rounds any arm is sampled by the arm filter when $U_r - L_r < \gamma/2$.

We may write the total number of samples drawn by the algorithm as

$$T = \sum_{r=1}^{\infty} \mathbb{1}\left[G_{r-1} \cup B_{r-1} \neq [K] \text{ and } U_{r-1} - L_{r-1} \geq \gamma/2\right] \sum_{\boldsymbol{a}\in\mathcal{A}} T_r(\boldsymbol{a}). \tag{EC.2.21}$$

$$
\begin{aligned}
T &\le \sum_{r=1}^{\infty} \mathbb{1}\left[G_{r-1} \cup B_{r-1} \neq [K]\right] \sum_{\boldsymbol{a}\in\mathcal{A}} T_r(\boldsymbol{a}) \\
&= \sum_{r=1}^{\infty} \mathbb{1}\left[G_{r-1} \neq G_\varepsilon\right] \mathbb{1}\left[G_{r-1} \cup B_{r-1} \neq [K]\right] \sum_{\boldsymbol{a}\in\mathcal{A}} T_r(\boldsymbol{a}) && \text{(EC.2.22)} \\
&+ \sum_{r=1}^{\infty} \mathbb{1}\left[G_{r-1} = G_\varepsilon\right] \mathbb{1}\left[G_{r-1} \cup B_{r-1} \neq [K]\right] \sum_{\boldsymbol{a}\in\mathcal{A}} T_r(\boldsymbol{a}). && \text{(EC.2.23)}
\end{aligned}
$$

Then by definition, $R_i$ ~~and $R'_i$~~ can be defined as

$$
R_i = \min\left\{ r : \begin{cases} i \in G_r(r) & \text{if } i \in G_\varepsilon \\ i \notin \mathcal{A}_I(r) & \text{if } i \in G_\varepsilon^c \end{cases} \right\} = \min\left\{ r : \begin{cases} \hat{\mu}_i - C_{\delta/K}(r) \ge U_r & \text{if } i \in G_\varepsilon \\ \hat{\mu}_i + C_{\delta/K}(r) \le L_r & \text{if } i \in G_\varepsilon^c \end{cases} \right\}, \tag{EC.2.24}
$$

~~$R'_i = \min\left\{ r : \hat{\mu}_i + C_{\delta/K}(r) \le \max_{j\in\mathcal{A}_I}\left(\hat{\mu}_j - C_{\delta/K}(r)\right)\right\}$. (EC.2.25)~~

Define $R_i = \infty$ ~~and $R'_i = \infty$~~ if ~~they~~ it never occurs. Note that this may happen if LinFACTE terminates due to the stopping condition in line 29 that $U_r - L_r < \gamma/2$. Finally, we define the time $R_\gamma$ such that $U_r - L_r < \gamma/2$, which is designed to ensure that no arm is sampled more than $R_\gamma$ times by the arm filter, controlling the case that $R_i$ ~~or $R'_i$~~ ~~are~~ is infinite.

$$
R_\gamma = \min\left\{ r : U_r - L_r < \gamma/2 \right\}. \tag{EC.2.26}
$$

#### EC.2.0.4. Step 2: Bounding $R_i$ ~~and $R'_i$~~ for $i \in G_\varepsilon$

**Claim 0: For $i \in G_\varepsilon$, we have that $R_i \le h\left(0.25\left(\varepsilon - \Delta_i\right)\right)$**

Proof. Helper function $h(\cdot,\cdot)$ is defined in Lemma EC.4.1 to assist our derivation. Note that $4C_{\delta/K}(r) \le \mu_i - (\mu_1 - \varepsilon)$ is true when $\boldsymbol{r \ge h\left(0.25\left(\varepsilon - \Delta_i\right)\right)}$, implies that for all $j \in \mathcal{A}_I$,

$$
\begin{aligned}
\hat{\mu}_i - C_{\delta/K}(r) &\overset{\mathcal{E}_1}{\ge} \mu_i - 2C_{\delta/K}(r) \\
&\ge \mu_1 + 2C_{\delta/K}(r) - \varepsilon \\
&\ge \mu_j + 2C_{\delta/K}(r) - \varepsilon \\
&\overset{\mathcal{E}_1}{\ge} \hat{\mu}_j + C_{\delta/K}(r) - \varepsilon.
\end{aligned} \tag{EC.2.27}
$$

Thus, in particular, $\hat{\mu}_i - C_{\delta/K}(r) \ge \max_{j\in\mathcal{A}_I} \hat{\mu}_j(t) + C_{\delta/K}(r) - \varepsilon = U_r$. □

Additionally, we define a round $R_{\max}$ when all good arms have been added into $G_r$.

**Claim 1: Defining $R_{\max} \triangleq \min\left\{r : G_r = G_\varepsilon\right\} = \max_{i\in G_\varepsilon} R_i$, we have that $R_{\max} \le h\left(0.25\alpha_\varepsilon\right)$ (in other words, if $r \ge h\left(0.25\alpha_\varepsilon\right)$, then we have that $G_r = G_\varepsilon$ )**

Proof. Recall that $\alpha_\varepsilon = \min_{i\in G_\varepsilon} \mu_i - \mu_1 + \varepsilon = \min_{i\in G_\varepsilon} \varepsilon - \Delta_i$. By **Claim 0**, $R_i \le h\left(0.25\left(\varepsilon - \Delta_i\right)\right)$. Furthermore, $h(\cdot)$ is monotonically decreasing in its argument. Then for any $\delta > 0$, $R_{\max} = \max_{i\in G_\varepsilon} R_i \le \max_{i\in G_\varepsilon} h\left(0.25\left(\varepsilon - \Delta_i\right)\right) = h\left(0.25\alpha_\varepsilon\right)$. □

**EC.2.0.5. Step 3: Bounding the total samples given to the arm filter at round $r = R_{\max}$**

Note that for a round $r = R$, the total samples given to the arm filter is $\sum_{s=1}^{R} \sum_{\boldsymbol{a} \in \mathcal{A}_s} T_s(\boldsymbol{a})$. Therefore, the total number of samples up to round $R_{\max}$ is $\sum_{s=1}^{R_{\max}} \sum_{\boldsymbol{a} \in \mathcal{A}_s} T_s(\boldsymbol{a})$. By line 13 of the algorithm, we have

$$T_r \leq \frac{2d}{\varepsilon_r^2} \log\left(\frac{Kr(r+1)}{\delta}\right) + \frac{d(d+1)}{2}. \tag{EC.2.28}$$

Hence,

$$\begin{aligned}
\sum_{s=1}^{R_{\max}} T_s &= \sum_{s=1}^{R_{\max}} \sum_{\boldsymbol{a} \in \mathcal{A}_s} T_s(\boldsymbol{a}) \\
&\leq \sum_{s=1}^{R_{\max}} \left(\frac{2d}{\varepsilon_s^2} \log\left(\frac{Ks(s+1)}{\delta}\right) + \frac{d(d+1)}{2}\right) \\
&\leq \sum_{s=1}^{R_{\max}} 2^{2s+1} d \log\left(\frac{KR_{\max}(R_{\max}+1)}{\delta}\right) + \frac{d(d+1)}{2} R_{\max} \\
&\leq C_1 2^{2R_{\max}+1} d \log\left(\frac{KR_{\max}(R_{\max}+1)}{\delta}\right) + \frac{d(d+1)}{2} R_{\max},
\end{aligned} \tag{EC.2.29}$$

where $C_1$ is a universal constant and recall that $R_{\max} \leq h\left(0.25\alpha_\varepsilon\right)$.

**EC.2.0.6. Step 4: Proof of Lemma 1 - Bounding Equation (EC.2.22)**

Recall that $R_{\max} \leq h(0.25\alpha_\varepsilon)$ is the round where $G_{R_{\max}} = G_\varepsilon$. Using the result of the previous step, we may bound the total number of samples taken through this round, controlling Equation (EC.2.22).

**Claim 0:**

$$\begin{aligned}
&\sum_{r=1}^{\infty} \mathbb{1}\left[G_{r-1} \neq G_\varepsilon\right] \mathbb{1}\left[G_{r-1} \cup B_{r-1} \neq [K]\right] \sum_{\boldsymbol{a} \in \mathcal{A}} T_r(\boldsymbol{a}) \\
&\leq C_1 2^{2R_{\max}+1} d \log\left(\frac{KR_{\max}(R_{\max}+1)}{\delta}\right) + \frac{d(d+1)}{2} R_{\max}.
\end{aligned} \tag{EC.2.30}$$

Proof. By definition of $R_{\max}$ and Equation (EC.2.29) in Step 3,

$$\begin{aligned}
&\sum_{r=1}^{\infty} \mathbb{1}\left[G_{r-1} \neq G_\varepsilon\right] \mathbb{1}\left[G_{r-1} \cup B_{r-1} \neq [K]\right] \sum_{\boldsymbol{a} \in \mathcal{A}} T_r(\boldsymbol{a}) \\
&= \sum_{r=1}^{R_{\max}} \mathbb{1}\left[G_{r-1} \cup B_{r-1} \neq [K]\right] \sum_{\boldsymbol{a} \in \mathcal{A}} T_r(\boldsymbol{a}) \\
&\leq C_1 2^{2R_{\max}+1} d \log\left(\frac{KR_{\max}(R_{\max}+1)}{\delta}\right) + \frac{d(d+1)}{2} R_{\max}.
\end{aligned} \tag{EC.2.31}$$

□

**EC.2.0.7. Step 5: Proof of Lemma 2 - Bounding Equation (EC.2.23)**

Proof.

$$\begin{aligned}
&\sum_{r=1}^{\infty} \mathbb{1}\left[G_{r-1} = G_\varepsilon\right] \mathbb{1}\left[G_{r-1} \cup B_{r-1} \neq [K]\right] \sum_{\boldsymbol{a}\in\mathcal{A}} T_r(\boldsymbol{a}) \\
&= \sum_{r=R_{\max}+1}^{\infty} \mathbb{1}\left[G_{r-1} \cup B_{r-1} \neq [K]\right] \sum_{\boldsymbol{a}\in\mathcal{A}} T_r(\boldsymbol{a}) \\
&= \sum_{r=R_{\max}+1}^{\infty} \mathbb{1}\left[B_{r-1} \neq G_\varepsilon^c\right] \sum_{\boldsymbol{a}\in\mathcal{A}} T_r(\boldsymbol{a}) \\
&\leq \sum_{r=R_{\max}+1}^{\infty} |G_\varepsilon^c \backslash B_{r-1}| \sum_{\boldsymbol{a}\in\mathcal{A}} T_r(\boldsymbol{a}) \\
&= \sum_{r=R_{\max}+1}^{\infty} \sum_{i\in G_\varepsilon^c} \mathbb{1}\left[i \notin B_{r-1}\right] \sum_{\boldsymbol{a}\in\mathcal{A}} T_r(\boldsymbol{a}) \\
&= \sum_{i\in G_\varepsilon^c} \sum_{r=R_{\max}+1}^{\infty} \mathbb{1}\left[i \notin B_{r-1}\right] \sum_{\boldsymbol{a}\in\mathcal{A}} T_r(\boldsymbol{a}) \\
&\leq \sum_{i\in G_\varepsilon^c} \sum_{r=1}^{\infty} \mathbb{1}\left[i \notin B_{r-1}\right] \sum_{\boldsymbol{a}\in\mathcal{A}} T_r(\boldsymbol{a}) \\
&\leq \sum_{i\in G_\varepsilon^c} \sum_{r=1}^{\infty} \mathbb{1}\left[i \notin B_{r-1}\right] \left(\frac{2d}{\varepsilon_r^2} \log\left(\frac{Kr(r+1)}{\delta}\right) + \frac{d(d+1)}{2}\right). \qquad \text{(EC.2.32)}
\end{aligned}$$

□

#### EC.2.0.8. Step 6: Bounding the expected total number of samples drawn by LinFACTE

Now we take expectations over the number of samples drawn. These expectations are conditioned on the high probability event $\mathcal{E}_1$.

$$\begin{aligned}
\mathbb{E}\left[T \mid \mathcal{E}_1\right] &\leq \sum_{r=1}^{\infty} \mathbb{E}\left[\mathbb{1}\left[G_r \cup B_r \neq [K]\right] \mid \mathcal{E}_1\right] \sum_{\boldsymbol{a}\in\mathcal{A}} T_r(\boldsymbol{a}) \\
&= \sum_{r=1}^{\infty} \mathbb{E}\left[\mathbb{1}\left[G_{r-1} \neq G_\epsilon\right] \mathbb{1}\left[G_{r-1} \cup B_{r-1} \neq [K]\right] \mid \mathcal{E}_1\right] \sum_{\boldsymbol{a}\in\mathcal{A}} T_r(\boldsymbol{a}) \\
&+ \sum_{r=1}^{\infty} \mathbb{E}\left[\mathbb{1}\left[G_{r-1} = G_\epsilon\right] \mathbb{1}\left[G_{r-1} \cup B_{r-1} \neq [K]\right] \mid \mathcal{E}_1\right] \sum_{\boldsymbol{a}\in\mathcal{A}} T_r(\boldsymbol{a}) \\
&\overset{\text{Step 4}}{\leq} C_1 2^{2R_{\max}+1} d \log\left(\frac{KR_{\max}(R_{\max}+1)}{\delta}\right) + \frac{d(d+1)}{2} R_{\max} \\
&+ \sum_{r=1}^{\infty} \mathbb{E}\left[\mathbb{1}\left[G_{r-1} = G_\epsilon\right] \mathbb{1}\left[G_{r-1} \cup B_{r-1} \neq [K]\right] \mid \mathcal{E}_1\right] \sum_{\boldsymbol{a}\in\mathcal{A}} T_r(\boldsymbol{a}) \\
&\overset{\text{Step 5}}{\leq} C_1 2^{2R_{\max}+1} d \log\left(\frac{KR_{\max}(R_{\max}+1)}{\delta}\right) + \frac{d(d+1)}{2} R_{\max} \\
&+ \sum_{i\in G_\varepsilon^c} \sum_{r=1}^{\infty} \mathbb{E}\left[\mathbb{1}\left[i \notin B_{r-1} \mid \mathcal{E}_1\right]\right] \left(\frac{2d}{\varepsilon_r^2} \log\left(\frac{Kr(r+1)}{\delta}\right) + \frac{d(d+1)}{2}\right). \qquad \text{(EC.2.33)}
\end{aligned}$$

**EC.2.0.9. Step 7: Bounding $\sum_{r=1}^{\infty} \mathbb{E}_\nu \left[\mathbb{1}\left[i \notin B_{r-1}\right] \mid \mathcal{E}_1\right] \left(\frac{2d}{\varepsilon_r^2} \log\left(\frac{Kr(r+1)}{\delta}\right) + \frac{d(d+1)}{2}\right)$**

Next, we bound the expectation remaining in Step 6. First, we bound the probability that for a given $i \in G_\epsilon^c$ and a given $r$, $i \notin B_r$. Note that by Borel-Cantelli, this implies that the probability that $i$ is never added to any $B_r$ is 0.

**Claim 0:**

For $i \in G_\epsilon^c$ and $r \geq \left\lceil \log_2\left(\frac{4}{\Delta_i - \epsilon}\right)\right\rceil \Longrightarrow \mathbb{E}_\nu \left[\mathbb{1}\left[i \notin B_r\right] \mid \mathcal{E}_1\right] = 0$

Proof. First, we have that for any $i \in G_\epsilon^c$

$$\mathbb{E}_\nu \left[\mathbb{1}\left[i \notin B_r\right] \mid \mathcal{E}_1\right] = \mathbb{E}_\nu \left[\mathbb{1}\left[\max_{j \in \mathcal{A}_I} \hat{\mu}_j - \hat{\mu}_i \leq 2\varepsilon_r + \varepsilon\right] \mid \mathcal{E}_1\right]. \tag{EC.2.34}$$

If $i \in B_{r-1}$, then $i \in B_r$ by definition. Otherwise, if $i \notin B_{r-1}$, then under event $\mathcal{E}_1$, we have

$$|(\hat{\mu}_j - \hat{\mu}_i) - (\mu_j - \mu_i)| \leq 2\varepsilon_r \tag{EC.2.35}$$

and

$$\max_{j \in \mathcal{A}_I} \hat{\mu}_j - \hat{\mu}_i \geq \mu_1 - \mu_i - 2\varepsilon_r = \Delta_i - 2\varepsilon_r = \Delta_i - 2^{-r+1}. \tag{EC.2.36}$$

Then for $i \in G_\varepsilon^c$ and $r \geq \left\lceil \log_2\left(\frac{4}{\Delta_i - \epsilon}\right)\right\rceil$, we have

$$\max_{j \in \mathcal{A}_I} \hat{\mu}_j - \hat{\mu}_i \geq \Delta_i - 2^{-r+1} \geq \varepsilon + 2\varepsilon_r, \tag{EC.2.37}$$

which implies that $i \in B_r$ by line 22 of the algorithm. In other words, we have proved the correctness of the algorithm if line 22 happens in step 0, and here we introduce when exactly line 22 will happen. In particular, under event $\mathcal{E}_1$, if $i \notin B_{r-1}$, for all $i \in G_\varepsilon^c$ and $r \geq \left\lceil \log_2\left(\frac{4}{\Delta_i - \epsilon}\right)\right\rceil$, we have

$$\mathbb{E}_\nu \left[\mathbb{1}\left[\max_{j \in \mathcal{A}_I} \hat{\mu}_j - \hat{\mu}_i > 2\varepsilon_r + \varepsilon\right] \mid i \notin B_{r-1}\right] = 1. \tag{EC.2.38}$$

Deterministically, $\mathbb{1}\left[i \notin B_r\right] \mathbb{1}\left[i \in B_{r-1}\right] = 0$. Therefore,

$$\begin{aligned}
&\mathbb{E}_\nu \left[\mathbb{1}\left[\max_{j \in \mathcal{A}_I} \hat{\mu}_j - \hat{\mu}_i \leq 2\varepsilon_r + \varepsilon\right] \mid \mathcal{E}_1\right] \\
&= \mathbb{E}_\nu \left[\mathbb{1}\left[\max_{j \in \mathcal{A}_I} \hat{\mu}_j - \hat{\mu}_i \leq 2\varepsilon_r + \varepsilon\right] \mathbb{1}\left[i \notin B_{r-1}\right] \mid \mathcal{E}_1\right] \\
&+ \mathbb{E}_\nu \left[\mathbb{1}\left[\max_{j \in \mathcal{A}_I} \hat{\mu}_j - \hat{\mu}_i \leq 2\varepsilon_r + \varepsilon\right] \mathbb{1}\left[i \in B_{r-1}\right] \mid \mathcal{E}_1\right] \\
&= \mathbb{E}_\nu \left[\mathbb{1}\left[\max_{j \in \mathcal{A}_I} \hat{\mu}_j - \hat{\mu}_i \leq 2\varepsilon_r + \varepsilon\right] \mathbb{1}\left[i \notin B_{r-1}\right] \mid \mathcal{E}_1\right] \\
&= \mathbb{E}_\nu \left[\mathbb{1}\left[\max_{j \in \mathcal{A}_I} \hat{\mu}_j - \hat{\mu}_i \leq 2\varepsilon_r + \varepsilon\right] \mathbb{1}\left[i \notin B_{r-1}\right] \mid i \notin B_{r-1}, \mathcal{E}_1\right] \mathbb{P}\left(i \notin B_{r-1} \mid \mathcal{E}_1\right) \\
&+ \mathbb{E}_\nu \left[\mathbb{1}\left[\max_{j \in \mathcal{A}_I} \hat{\mu}_j - \hat{\mu}_i \leq 2\varepsilon_r + \varepsilon\right] \mathbb{1}\left[i \notin B_{r-1}\right] \mid i \in B_{r-1}, \mathcal{E}_1\right] \mathbb{P}\left(i \in B_{r-1} \mid \mathcal{E}_1\right)
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_\nu \left[ \mathbb{1}\left[ \max_{j\in\mathcal{A}_I} \hat{\mu}_j - \hat{\mu}_i \le 2\varepsilon_r + \varepsilon \right] \mathbb{1}\left[ i \notin B_{r-1} \right] \mid i \notin B_{r-1}, \mathcal{E}_1 \right] \mathbb{P}\left( i \notin B_{r-1} \mid \mathcal{E}_1 \right) \\
&= \mathbb{E}_\nu \left[ \mathbb{1}\left[ \max_{j\in\mathcal{A}_I} \hat{\mu}_j - \hat{\mu}_i \le 2\varepsilon_r + \varepsilon \right] \mid i \notin B_{r-1}, \mathcal{E}_1 \right] \mathbb{E}_\nu \left[ \mathbb{1}\left[ i \notin B_{r-1} \right] \mid \mathcal{E}_1 \right] \\
&= 0. \qquad \text{(EC.2.39)}
\end{aligned}
$$

□

**Claim 1:** For $i \in G_\epsilon^c$,

$$
\begin{aligned}
&\sum_{r=1}^{\infty} \mathbb{E}_\nu \left[ \mathbb{1}\left[ i \notin B_{r-1} \right] \mid \mathcal{E}_1 \right] \left( \frac{2d}{\varepsilon_r^2} \log\left( \frac{Kr(r+1)}{\delta} \right) + \frac{d(d+1)}{2} \right) \\
&\le \frac{d(d+1)}{2} \log_2\left( \frac{4}{\Delta_i - \epsilon} \right) + C_2 \frac{64d}{(\Delta_i - \varepsilon)^2} \log\left( \frac{K}{\delta} \log_2 \frac{8}{\Delta_i - \varepsilon} \right). \qquad \text{(EC.2.40)}
\end{aligned}
$$

Proof.

$$
\begin{aligned}
&\sum_{r=1}^{\infty} \mathbb{E}_\nu \left[ \mathbb{1}\left[ i \notin B_{r-1} \right] \mid \mathcal{E}_1 \right] \left( \frac{2d}{\varepsilon_r^2} \log\left( \frac{Kr(r+1)}{\delta} \right) + \frac{d(d+1)}{2} \right) \\
&= \sum_{r=1}^{\left\lfloor \log_2\left(\frac{4}{\Delta_i-\epsilon}\right) \right\rfloor} \mathbb{E}_\nu \left[ \mathbb{1}\left[ i \notin B_{r-1} \right] \mid \mathcal{E}_1 \right] \left( \frac{2d}{\varepsilon_r^2} \log\left( \frac{Kr(r+1)}{\delta} \right) + \frac{d(d+1)}{2} \right) \\
&+ \sum_{r=\left\lceil \log_2\left(\frac{4}{\Delta_i-\epsilon}\right) \right\rceil}^{\infty} \mathbb{E}_\nu \left[ \mathbb{1}\left[ i \notin B_{r-1} \right] \mid \mathcal{E}_1 \right] \left( \frac{2d}{\varepsilon_r^2} \log\left( \frac{Kr(r+1)}{\delta} \right) + \frac{d(d+1)}{2} \right) \\
&= \sum_{r=1}^{\left\lfloor \log_2\left(\frac{4}{\Delta_i-\epsilon}\right) \right\rfloor} \mathbb{E}_\nu \left[ \mathbb{1}\left[ i \notin B_{r-1} \right] \mid \mathcal{E}_1 \right] \left( \frac{2d}{\varepsilon_r^2} \log\left( \frac{Kr(r+1)}{\delta} \right) + \frac{d(d+1)}{2} \right) + 0 \\
&\le \sum_{r=1}^{\left\lfloor \log_2\left(\frac{4}{\Delta_i-\epsilon}\right) \right\rfloor} \left( d2^{2r+1} \log\left( \frac{Kr(r+1)}{\delta} \right) + \frac{d(d+1)}{2} \right) \\
&\le \frac{d(d+1)}{2} \log_2\left( \frac{4}{\Delta_i - \epsilon} \right) + 2d \log\left( \frac{K}{\delta} \right) \sum_{r=1}^{\left\lfloor \log_2\left(\frac{4}{\Delta_i-\epsilon}\right) \right\rfloor} 2^{2r} + 4d \sum_{r=1}^{\left\lfloor \log_2\left(\frac{4}{\Delta_i-\epsilon}\right) \right\rfloor} 2^{2r} \log(r+1) \\
&\le \frac{d(d+1)}{2} \log_2\left( \frac{4}{\Delta_i - \epsilon} \right) + C_2 \frac{64d}{(\Delta_i - \varepsilon)^2} \log\left( \frac{K}{\delta} \log_2 \frac{8}{\Delta_i - \varepsilon} \right). \qquad \text{(EC.2.41)}
\end{aligned}
$$

where $C_2$ is a sufficiently large universal constant.

□

### EC.2.0.10. Step 8: Applying the result of Step 7 to the result of Step 6

Summarizing the aforementioned results, we have

$$
\begin{aligned}
\mathbb{E}\left[ T \mid \mathcal{E}_1 \right] \le\ & C_1 2^{2R_{\max}+1} d \log\left( \frac{KR_{\max}(R_{\max}+1)}{\delta} \right) + \frac{d(d+1)}{2} R_{\max} \\
&+ \sum_{i \in G_\varepsilon^c} \sum_{r=1}^{\infty} \mathbb{E}\left[ \mathbb{1}\left[ i \notin B_{r-1} \mid \mathcal{E}_1 \right] \right] \left( \frac{2d}{\varepsilon_r^2} \log\left( \frac{Kr(r+1)}{\delta} \right) + \frac{d(d+1)}{2} \right), \qquad \text{(EC.2.42)}
\end{aligned}
$$

$$R_{\max} \leq h\left(0.25\alpha_\varepsilon\right) = \log_2 \frac{4}{\alpha_\varepsilon}, \tag{EC.2.43}$$

and

$$\begin{aligned}
&\sum_{r=1}^{\infty} \mathbb{E}_\nu \left[\mathbb{1}\left[i \notin B_{r-1}\right] \mid \mathcal{E}_1\right] \left(\frac{2d}{\varepsilon_r^2} \log\left(\frac{Kr(r+1)}{\delta}\right) + \frac{d(d+1)}{2}\right) \\
&\leq \frac{d(d+1)}{2} \log_2\left(\frac{4}{\Delta_i - \epsilon}\right) + C_2 \frac{64d}{(\Delta_i - \varepsilon)^2} \log\left(\frac{K}{\delta} \log_2 \frac{8}{\Delta_i - \varepsilon}\right).
\end{aligned} \tag{EC.2.44}$$

We may repeat the result of Step 7 for every $G_\varepsilon^c$ and plug this inequality into the result of Step 6. Then we can return the final result as follows.

$$\begin{aligned}
\mathbb{E}\left[T \mid \mathcal{E}_1\right] &\leq C_1 2^{2R_{\max}+1} d \log\left(\frac{KR_{\max}(R_{\max}+1)}{\delta}\right) + \frac{d(d+1)}{2} R_{\max} \\
&+ \sum_{i \in G_\varepsilon^c} \sum_{r=1}^{\infty} \mathbb{E}\left[\mathbb{1}\left[i \notin B_{r-1} \mid \mathcal{E}_1\right]\right] \left(\frac{2d}{\varepsilon_r^2} \log\left(\frac{Kr(r+1)}{\delta}\right) + \frac{d(d+1)}{2}\right) \\
&\leq C_1 d \frac{64}{\alpha_\varepsilon^2} \log\left(\frac{K}{\delta} \log_2\left(\frac{8}{\alpha_\varepsilon}\right)\right) + \frac{d(d+1)}{2} \log_2 \frac{4}{\alpha_\varepsilon} \\
&+ \sum_{i \in G_\varepsilon^c} \left(\frac{d(d+1)}{2} \log_2\left(\frac{4}{\Delta_i - \epsilon}\right) + C_2 \frac{64d}{(\Delta_i - \varepsilon)^2} \log\left(\frac{K}{\delta} \log_2 \frac{8}{\Delta_i - \varepsilon}\right)\right).
\end{aligned} \tag{EC.2.45}$$

**EC.2.0.11. A refined version of our upper bound**

However, the final result derived from the former steps contains the form of summation over set $G_\varepsilon$, which can be improved by removing the summation.

**Claim 0:** For the expected sample complexity With high probability event $\mathcal{E}_1$, we have

$$\mathbb{E}\left[T \mid \mathcal{E}_1\right] \leq C_3 \max\left\{\frac{64d}{\alpha_\varepsilon^2} \log\left(\frac{K}{\delta} \log_2 \frac{8}{\alpha_\varepsilon}\right), \frac{64d}{\beta_\varepsilon^2} \log\left(\frac{K}{\delta} \log_2 \frac{8}{\beta_\varepsilon}\right)\right\}, \tag{EC.2.46}$$

where $C_3$ is a universal constant and $R_{\text{upper}} = \max\left\{\left\lceil \log_2 \frac{4}{\alpha_\varepsilon} \right\rceil, \left\lceil \log_2 \frac{4}{\beta_\varepsilon} \right\rceil\right\}$ is the round where all the classifications have been finished and the answer is returned.

Proof. Based on the analysis of this step, we can have another decomposition of $T$ in equation (EC.2.33). These expectations are conditioned on the high probability event $\mathcal{E}_1$, given by

$$\begin{aligned}
\mathbb{E}\left[T \mid \mathcal{E}_1\right] &\leq \sum_{r=1}^{\infty} \mathbb{E}\left[\mathbb{1}\left[G_r \cup B_r \neq [K]\right] \mid \mathcal{E}_1\right] \sum_{\boldsymbol{a} \in \mathcal{A}} T_r(\boldsymbol{a}) \\
&\leq \sum_{r=1}^{R_{\text{upper}}} \left(d 2^{2r+1} \log\left(\frac{Kr(r+1)}{\delta}\right) + \frac{d(d+1)}{2}\right) \\
&\leq \frac{d(d+1)}{2} R_{\text{upper}} + 2d \log\left(\frac{K}{\delta}\right) \sum_{r=1}^{R_{\text{upper}}} 2^{2r} + 4d \sum_{r=1}^{R_{\text{upper}}} 2^{2r} \log(r+1) \\
&\leq 4d \log\left[\frac{K}{\delta}\left(R_{\text{upper}}+1\right)\right] 2^{2R_{\text{upper}}} \\
&\leq C_3 \max\left\{\frac{64d}{\alpha_\varepsilon^2} \log\left(\frac{K}{\delta} \log_2 \frac{8}{\alpha_\varepsilon}\right), \frac{64d}{\beta_\varepsilon^2} \log\left(\frac{K}{\delta} \log_2 \frac{8}{\beta_\varepsilon}\right)\right\},
\end{aligned} \tag{EC.2.47}$$

where $C_3$ is a universal constant.

### EC.2.0.12. Additional insights into the algorithm optimality

From another perspective, we consider the relationship between the lower bound and the upper bound in the following section and give some additional insights into the algorithm optimality.

For $\forall i \in (\mathcal{A}_I \cap G_\varepsilon)$ and $\forall j \in \mathcal{A}_I$, in round $r$, we have

$$\boldsymbol{y}_{j,i}^\top(\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}) \leq 2\varepsilon_r \tag{EC.2.48}$$

and

$$\boldsymbol{y}_{j,i}^\top\hat{\boldsymbol{\theta}}_r - \varepsilon \leq \boldsymbol{y}_{j,i}^\top\boldsymbol{\theta} + 2\varepsilon_r - \varepsilon. \tag{EC.2.49}$$

**Claim 0:** Define $G_r'$ as $G_r' \triangleq \{\exists j \in \mathcal{A}_I, i : \boldsymbol{y}_{j,i}\boldsymbol{\theta} - \varepsilon > -4\varepsilon_r\}$. We can prove that for $i \in G_\varepsilon$, we always have $(\mathcal{A}_I \cap G_\varepsilon \cap G_r^c) \subset G_r'$.

Proof. with the idea of contradiction, if $i \in (\mathcal{A}_I \cap G_\varepsilon \cap G_r^c) \cap (G_r')^c$, then for $\forall j \in \mathcal{A}_I$, we have

$$\boldsymbol{y}_{j,i}^\top\boldsymbol{\theta} \leq -4\varepsilon_r + \varepsilon, \tag{EC.2.50}$$

Hence, with equation (EC.2.49), for $\forall j \in \mathcal{A}_I$, we have

$$\boldsymbol{y}_{j,i}^\top\hat{\boldsymbol{\theta}}_r - \varepsilon \leq -2\varepsilon_r \tag{EC.2.51}$$

which is exactly the condition for the arm filter to add arm $i$ into $G_r$ by line 20 of the algorithm. This contradiction leads to the result. Besides, note that when $r \geq \left\lceil \log_2 \frac{4}{\alpha_\varepsilon} \right\rceil$, we have $G_r' = \emptyset$. Furthermore, considering that $i \in G_\varepsilon$, we have $\boldsymbol{y}_{i,j}^\top\boldsymbol{\theta} + \varepsilon > 0$. □

On the other hand, for $i \in (\mathcal{A}_I \cap G_\varepsilon^c)$, in round $r$, we have

$$\boldsymbol{y}_{1,i}^\top(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_r) \leq 2\varepsilon_r \tag{EC.2.52}$$

and

$$\boldsymbol{y}_{1,i}^\top\hat{\boldsymbol{\theta}}_r - \varepsilon \geq \boldsymbol{y}_{1,i}^\top\boldsymbol{\theta} - 2\varepsilon_r - \varepsilon. \tag{EC.2.53}$$

**Claim 1:** Define $B_r'$ as $B_r' \triangleq \left\{i : \boldsymbol{y}_{1,i}^\top\boldsymbol{\theta} - \varepsilon < 4\varepsilon_r\right\}$. Then considering that 1 always belongs to $\mathcal{A}_I(r)$, we can prove that for $i \in G_\varepsilon^c$, we always have $(\mathcal{A}_I \cap G_\varepsilon^c) \subset B_r'$.

Proof. To prove this, with the same idea of recursion, if $i \in (\mathcal{A}_I \cap G_\varepsilon^c) \cap (B_r')^c$, then we have

$$\boldsymbol{y}_{1,i}^\top\boldsymbol{\theta} \geq 4\varepsilon_r + \varepsilon, \tag{EC.2.54}$$

Hence, with equation (EC.2.53), we have

$$\boldsymbol{y}_{1,i}^\top\hat{\boldsymbol{\theta}}_r - \varepsilon \geq 2\varepsilon_r, \tag{EC.2.55}$$

which is also exactly the condition for the arm filter to add arm $i$ into $B_r$ by line 22 of the algorithm. This contradiction leads to the result. Besides, note that when $r \geq \left\lceil \log_2 \frac{4}{\beta_\varepsilon} \right\rceil$, we have $B_r' = \emptyset$, when $r = \left\lfloor \log_2 \frac{4}{\beta_\varepsilon} \right\rfloor$, we have $G_r' = \beta$. Furthermore, for $i \in G_\varepsilon^c$, we can also conclude that $\boldsymbol{y}_{1,i}^\top\boldsymbol{\theta} - \varepsilon > 0$. □

**Claim 2:** For the stochastic linear bandit, considering the value of lower bound $(\Gamma^*)^{-1}$, we have

$$(\Gamma^*)^{-1} \geq \frac{1}{4R_{\text{upper}}} \sum_{r=1}^{R_{\text{upper}}} 2^{2r-3} \min_{\boldsymbol{p}\in S_K} \max_{i,j\in\mathcal{A}_I(r)} \|\boldsymbol{y}_{i,j}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}. \tag{EC.2.56}$$

where $R_{\text{upper}} = \max\left\{\left\lceil \log_2 \frac{4}{\alpha_\varepsilon}\right\rceil, \left\lceil \log_2 \frac{4}{\beta_\varepsilon}\right\rceil\right\}$ is the round where all the classifications have been finished and the answer is returned.

Proof. Let $R_{\text{upper}} = \max\left\{\left\lceil \log_2 \frac{4}{\alpha_\varepsilon}\right\rceil, \left\lceil \log_2 \frac{4}{\beta_\varepsilon}\right\rceil\right\}$. When round $r$ is larger than $R_{\text{upper}}$, newly defined sets $G'_r$ and $B'_r$ are both , meaning $G_\varepsilon \cup G_\varepsilon^c = [K]$ and the algorithm is terminated. From Theorem 3, we have

$$\begin{aligned}
(\Gamma^*)^{-1} &= \min_{\boldsymbol{p}\in S_K} \max_{(i,j,m)\in\mathcal{X}} \max\left\{\frac{2\|\boldsymbol{y}_{i,j}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}}{\left(\boldsymbol{y}_{i,j}^\top\boldsymbol{\theta}+\varepsilon\right)^2}, \frac{2\|\boldsymbol{y}_{1,m}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}}{\left(\boldsymbol{y}_{1,m}^\top\boldsymbol{\theta}-\varepsilon\right)^2}\right\} \\
&= \min_{\boldsymbol{p}\in S_K} \max_{r\leq R_{\text{upper}}} \max_{i\in G'_r} \max_{j\in\mathcal{A}_I} \max_{m\in B'_r} \max\left\{\frac{2\|\boldsymbol{y}_{i,j}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}}{\left(\boldsymbol{y}_{i,j}^\top\boldsymbol{\theta}+\varepsilon\right)^2}, \frac{2\|\boldsymbol{y}_{1,m}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}}{\left(\boldsymbol{y}_{1,m}^\top\boldsymbol{\theta}-\varepsilon\right)^2}\right\} \\
&\geq \min_{\boldsymbol{p}\in S_K} \max_{r\leq R_{\text{upper}}} \max_{i\in G'_r} \max_{j\in\mathcal{A}_I} \max_{m\in B'_r} \max\left\{\frac{2\|\boldsymbol{y}_{i,j}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}}{(4\varepsilon_r)^2}, \frac{2\|\boldsymbol{y}_{1,m}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}}{(4\varepsilon_r)^2}\right\} \\
&\overset{\text{(i)}}{\geq} \frac{1}{R_{\text{upper}}} \min_{\boldsymbol{p}\in S_K} \sum_{r=1}^{R_{\text{upper}}} \max_{i\in G'_r} \max_{j\in\mathcal{A}_I} \max_{m\in B'_r} \max\left\{\frac{2\|\boldsymbol{y}_{i,j}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}}{(4\varepsilon_r)^2}, \frac{2\|\boldsymbol{y}_{1,m}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}}{(4\varepsilon_r)^2}\right\} \\
&\overset{\text{(ii)}}{\geq} \frac{1}{R_{\text{upper}}} \sum_{r=1}^{R_{\text{upper}}} 2^{2r-3} \min_{\boldsymbol{p}\in S_K} \max_{i\in G'_r} \max_{j\in\mathcal{A}_I} \max_{m\in B'_r} \max\left\{\|\boldsymbol{y}_{i,j}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}, \|\boldsymbol{y}_{1,m}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}\right\} \\
&\overset{\text{(iii)}}{\geq} \frac{1}{R_{\text{upper}}} \sum_{r=1}^{R_{\text{upper}}} 2^{2r-3} \min_{\boldsymbol{p}\in S_K} \max_{i\in\mathcal{A}_I\cap G_\varepsilon\cap G_r^c} \max_{j\in\mathcal{A}_I} \max_{m\in\mathcal{A}_I\cap G_\varepsilon^c} \max\left\{\|\boldsymbol{y}_{i,j}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}, \|\boldsymbol{y}_{1,m}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}\right\} \\
&\overset{\text{(iv)}}{\geq} \frac{{\mathcal{G}_{\mathcal{Y}}}^2 L_2}{dL_1} \frac{1}{R_{\text{upper}}} \sum_{r=1}^{R_{\text{upper}}} 2^{2r-3} \min_{\boldsymbol{p}\in S_K} \max_{i\in\mathcal{A}_I\cap G_\varepsilon} \max_{m\in\mathcal{A}_I\cap G_\varepsilon^c} \max\left\{\|\boldsymbol{y}_{1,i}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}, \|\boldsymbol{y}_{1,m}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}\right\} \\
&\overset{\text{(v)}}{\geq} \frac{{\mathcal{G}_{\mathcal{Y}}}^2 L_2}{4R_{\text{upper}} dL_1} \sum_{r=1}^{R_{\text{upper}}} 2^{2r-3} \min_{\boldsymbol{p}\in S_K} \max_{i,j\in\mathcal{A}_I} \|\boldsymbol{y}_{i,j}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}},
\end{aligned} \tag{EC.2.57}$$

where $(i)$ follows from the fact that the maximum of positive numbers is always less than the average, and $(ii)$ by the fact that the minimum of a sum is greater than the sum of minimums. $(iii)$ comes from the inclusion relationship between sets in **Claim 0** and **Claim 1**. Take $j = 1$ and $(iv)$ is a conclusion of Lemma EC.4.3. To see $(v)$, note that for $i, j \in \mathcal{A}_I(r)$, we have $\max_{i,j\in\mathcal{A}_I} \|\boldsymbol{y}_{i,j}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}} \leq 4\max_{i\in\mathcal{A}_I} \|\boldsymbol{y}_{1,i}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}$. □

What's more, to give some additional insights, we can derive another form of the upper bound, given by

$$\begin{aligned}
\mathbb{E}\left[T \mid \mathcal{E}_1\right] &\leq \sum_{r=1}^{\infty} \mathbb{E}\left[\mathbb{1}\left[G_r \cup B_r \neq [K]\right] \mid \mathcal{E}_1\right] \sum_{\boldsymbol{a}\in\mathcal{A}} T_r(\boldsymbol{a}) \\
&\leq \sum_{r=1}^{R_{\text{upper}}} \left(d2^{2r+1} \log\left(\frac{Kr(r+1)}{\delta}\right) + \frac{d(d+1)}{2}\right)
\end{aligned}$$

$$\leq \frac{d(d+1)}{2} R_{\text{upper}} + 2d \log\left(\frac{K}{\delta}\right) \sum_{r=1}^{R_{\text{upper}}} 2^{2r} + 4d \sum_{r=1}^{R_{\text{upper}}} 2^{2r} \log(r+1)$$

$$\leq 4 \log\left[\frac{K}{\delta}\left(R_{\text{upper}}+1\right)\right] \sum_{r=1}^{R_{\text{upper}}} d 2^{2r}. \tag{EC.2.58}$$

From inequalities (EC.2.57) and (EC.2.58), it can be seen that to match the upper bound and the lower bound, we have to prove that $d \leq C_6 \min_{\boldsymbol{p}\in S_K} \max_{i,j\in\mathcal{A}_I(r)} \|\boldsymbol{y}_{i,j}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}$ with some constant $C_6$. However, with the Kiefer–Wolfowitz Theorem introduced in Theorem 1, we can only have the following inequality.

$$\min_{\boldsymbol{p}\in S_K} \max_{i,j\in\mathcal{A}_I(r)} \|\boldsymbol{y}_{i,j}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}} \leq 2 \min_{\boldsymbol{p}\in S_K} \max_{i\in\mathcal{A}_I(r)} \|\boldsymbol{a}_i\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}} = 2d, \tag{EC.2.59}$$

where the direction of the inequality is reversed, can not help us in bridging the gap between the lower bound and our upper bound. This is exactly the motivation for us to adjust the sampling policy of our algorithm. □

## EC.3. Proof of Theorem 5

From the former derivation and results in Theorem 4, it can be seen that the upper bound of the proposed algorithm can't match the lower bound in any form. Thus, to give a matching upper bound of the expected sample complexity, we give the following proof. The proof proceeds as follows. We begin by showing that the good event holds with probability at least $1-\delta_r$ in round $r$ given that the good event is held in round $r-1$. We then show that the probability of this good event holding in every round is at least $1-\delta$. As a result, we can simply sum over the bound on the sample complexity in each round given in the good event to obtain the stated bound on the sample complexity.

Considering the difference between $G$-optimal allocation and $\mathcal{XY}$-optimal allocation, the upper bound analysis based on $G$-optimal design can't match the lower bound in any sense. Here we rearrange the clean event and give the proof process for the $\mathcal{XY}$-optimal design in line 14 of the LinFACTE algorithm.

The initial point of the algorithm design and the proof is the definition of the events $\mathcal{E}_3$ below.

$$\mathcal{E}_3 = \bigcap_{i\in\mathcal{A}_I} \bigcap_{\substack{j\in\mathcal{A}_I \\ j\neq i}} \bigcap_{r\in\mathbb{N}} |(\hat{\mu}_j(r) - \hat{\mu}_i(r)) - (\mu_j - \mu_i)| \leq 2\varepsilon_r. \tag{EC.3.1}$$

~~Define $g_{\mathcal{XY}}(\mathcal{S}) = \min_{\pi_r\in\mathcal{P}(\mathcal{A}(r))} \max_{s\in\mathcal{S}} \|s\|^2_{\left(\sum_{\boldsymbol{a}\in\mathcal{A}} \pi_r(\boldsymbol{a})\boldsymbol{a}\boldsymbol{a}^\top\right)^{-1}}$ and the set $\mathcal{Y}(\mathcal{A})$ is defined as $\mathcal{Y}(\mathcal{A}) \triangleq \{\boldsymbol{a}-\boldsymbol{a}' : \forall \boldsymbol{a},\boldsymbol{a}'\in\mathcal{A}, \boldsymbol{a}\neq\boldsymbol{a}'\}$.~~ (ZK L: The content has been stated in the section of Problem Formulation.)Considering that the arm is sampled based on the preset allocation, i.e. the fixed design, here we introduce the following claim to give a simple confidence region of estimated parameter $\boldsymbol{\theta}$.

**Claim 0: Let $\delta\in(0,1)$, it holds that for each vector $\boldsymbol{y}\in\mathcal{Y}(\mathcal{A})$, we have**

$$P(\mathcal{E}_3) \geq 1-\delta. \tag{EC.3.2}$$

Proof. Conditioned on a specific choice of $\mathcal{Y}(\mathcal{A})$, since $\hat{\boldsymbol{\theta}}_r$ is a least squares estimator of $\boldsymbol{\theta}$ and the noise is i.i.d., we know that $\boldsymbol{y}^\top \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_r\right)$ is $\|\boldsymbol{y}\|^2_{\boldsymbol{V}_r^{-1}}$-sub-Gaussian for all $\boldsymbol{y} \in \mathcal{Y}(\mathcal{A})$. Furthermore, due to the guarantees of the rounding procedure, we have

$$\|\boldsymbol{y}\|^2_{\boldsymbol{V}_r^{-1}} \leq (1+\varepsilon)\, g_{\mathcal{X}\mathcal{Y}}(\mathcal{Y}(\mathcal{A}))/T_r \leq \frac{2^{-2r-1}}{\log \frac{2K(K-1)r(r+1)}{\delta}} \tag{EC.3.3}$$

for all $\boldsymbol{y} \in \mathcal{Y}(\mathcal{A})$ by our choice of $T_r$. Since the right-hand side is deterministic, independent of random reward of arms, for any $\rho > 0$, we have that

$$\mathbb{P}\left\{ |\boldsymbol{y}^\top \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_r\right)| > \sqrt{2^{-2r} \frac{\log(2/\rho)}{\log \frac{2K(K-1)r(r+1)}{\delta}}} \right\} \leq \rho \tag{EC.3.4}$$

for all $\boldsymbol{y} \in \mathcal{Y}(\mathcal{A})$. Taking $\rho = \frac{\delta}{K(K-1)r(r+1)}$ and and union bounding over all the possible $\boldsymbol{y} \in \mathcal{Y}(\mathcal{A})$, where we have $|\mathcal{Y}(\mathcal{A})| \leq |\mathcal{Y}(\mathcal{A}(0))| \leq K(K-1)$. Thus, to give the probability guarantee, we have

$$\begin{aligned}
\mathbb{P}(\mathcal{E}_3^c) &= \mathbb{P}\left\{ \bigcup_{i\in\mathcal{A}_I} \bigcup_{\substack{j\in\mathcal{A}_I \\ j\neq i}} \bigcup_{r\in\mathbb{N}} |(\hat{\mu}_j(r) - \hat{\mu}_i(r)) - (\mu_j - \mu_i)| > \varepsilon_r \right\} \\
&\leq \sum_{r=1}^{\infty} \mathbb{P}\left\{ \bigcup_{i\in\mathcal{A}_I} \bigcup_{\substack{j\in\mathcal{A}_I \\ j\neq i}} |(\hat{\mu}_j(r) - \hat{\mu}_i(r)) - (\mu_j - \mu_i)| > \varepsilon_r \right\} \\
&\leq \sum_{r=1}^{\infty} \sum_{i=1}^{K} \sum_{\substack{j=1 \\ j\neq i}}^{K} \frac{\delta}{K(K-1)r(r+1)} \\
&= \delta.
\end{aligned} \tag{EC.3.5}$$

Considering the union bounds over the rounds $r \in \mathbb{N}$, we have

$$P(\mathcal{E}_3) \geq 1 - \delta. \tag{EC.3.6}$$

□

Thus with the standard result of the $\mathcal{X}\mathcal{Y}$-optimal design, we have

$$C_{\delta/K}(r) \triangleq \varepsilon_r. \tag{EC.3.7}$$

**Claim 1: On the newly designed clean event $\mathcal{E}_2$, $1 \in \mathcal{A}_I$ for all $r \in \mathbb{N}$**

Proof. Firstly we show $1 \in \mathcal{A}_I$ for all $r \in \mathbb{N}$, that is, the best arm is never removed from $\mathcal{A}$. If the event $\mathcal{E}_2$ holds, note for any arm $i$

$$\hat{\mu}_i(r) - \hat{\mu}_1(r) \leq \mu_i - \mu_1 + 2C_{\delta/K}(r) \leq 2C_{\delta/K}(r) < 2C_{\delta/K}(r) + \varepsilon \tag{EC.3.8}$$

which particularly shows that $\hat{\mu}_1(r) + C_{\delta/K}(r) > \max_{i\in\mathcal{A}_I} \hat{\mu}_i - C_{\delta/K}(r) - \varepsilon = L_r$ and $\hat{\mu}_1(r) + C_{\delta/K}(r) \geq \max_{i\in\mathcal{A}} \hat{\mu}_i(r) - C_{\delta/K}$ showing that arm 1 will never exit $\mathcal{A}_I$ in line 22 or line 24. □

**Claim 2: Proof of Lemma 3**

Proof. Recall that in **Claim 0** and **Claim 1** of Section EC.2.0.11, we always have $(\mathcal{A}_I \cap G_\varepsilon) \subset G'_r$ and $(\mathcal{A}_I \cap G_\varepsilon^c) \subset B'_r$, where $G'_r \triangleq \{i : \boldsymbol{y}_{1,i}\boldsymbol{\theta} - \varepsilon > -4\varepsilon_r\}$ and $B'_r \triangleq \left\{i : \boldsymbol{y}_{1,i}^\top \boldsymbol{\theta} - \varepsilon < 4\varepsilon_r\right\}$. Furthermore, it means that $\mathcal{A}_I \subset (G'_r \cup B'_r)$, and thus

$$\begin{aligned} T_r &= \max\left\{\left\lceil \frac{2g_{\mathcal{XY}}\left(\mathcal{Y}(\mathcal{A})\right)(1+\epsilon)}{\varepsilon_r^2}\log\left(\frac{2K(K-1)r(r+1)}{\delta}\right)\right\rceil, r\left(\epsilon\right)\right\} \\ &\leq \max\left\{\left\lceil \frac{2g_{\mathcal{XY}}\left(\mathcal{Y}(G'_r \cup B'_r)\right)(1+\epsilon)}{\varepsilon_r^2}\log\left(\frac{2K(K-1)r(r+1)}{\delta}\right)\right\rceil, r\left(\epsilon\right)\right\}, \end{aligned} \tag{EC.3.9}$$

where we note that the quantity on the right-hand side is deterministic. □

**Claim 3: With probability greater than $1-\delta$, using an $\epsilon$-efficient rounding procedure. The proposed LinFACTE algorithm correctly identifies all $\varepsilon$-best arms and is instance-optimal up to logarithmic factors, given by**

$$T \leq C_4\left[R_{\text{upper}}\log\left(\frac{2K(R_{\text{upper}}+1)}{\delta}\right)\right](\Gamma^*)^{-1} + r\left(\epsilon\right)R_{\text{upper}}, \tag{EC.3.10}$$

where $C_4$ is a universal constant, $R_{\text{upper}} = \max\left\{\left\lceil \log_2 \frac{4}{\alpha_\varepsilon}\right\rceil, \left\lceil \log_2 \frac{4}{\beta_\varepsilon}\right\rceil\right\}$, and $(\Gamma^*)^{-1}$ is the critical term in the lower bound of linear bandit with All $\varepsilon$-Best pure exploration task.

Proof. Combining the result of **Claim 0** and **Claim 2** in this section, with probability as least $1-\delta$, we introduce the term $(\Gamma^*)^{-1}$, the lower bound in Theorem 3 to assist our derivation.

$$\begin{aligned} T &\leq \sum_{r=1}^{R_{\text{upper}}} \max\left\{\left\lceil \frac{2g_{\mathcal{XY}}\left(\mathcal{Y}(\mathcal{A})\right)(1+\epsilon)}{\varepsilon_r^2}\log\left(\frac{2K(K-1)r(r+1)}{\delta}\right)\right\rceil, r\left(\epsilon\right)\right\} \\ &\leq \sum_{r=1}^{R_{\text{upper}}} 2\cdot 2^{2r} g_{\mathcal{XY}}\left(\mathcal{Y}(\mathcal{A})\right)(1+\epsilon)\log\left(\frac{2K(K-1)r(r+1)}{\delta}\right) + (1+r\left(\epsilon\right))R_{\text{upper}} \\ &\leq \left[64\left(1+\epsilon\right)\log\left(\frac{2K(K-1)R_{\text{upper}}(R_{\text{upper}}+1)}{\delta}\right)\frac{R_{\text{upper}}dL_1}{\mathcal{G}_{\mathcal{Y}}{}^2 L_2}\right](\Gamma^*)^{-1} + (1+r\left(\epsilon\right))R_{\text{upper}} \\ &\leq \left[128\left(1+\epsilon\right)\log\left(\frac{2K(R_{\text{upper}}+1)}{\delta}\right)\frac{R_{\text{upper}}dL_1}{\mathcal{G}_{\mathcal{Y}}{}^2 L_2}\right](\Gamma^*)^{-1} + (1+r\left(\epsilon\right))R_{\text{upper}} \\ &\leq C_4\left[dR_{\text{upper}}\log\left(\frac{2K(R_{\text{upper}}+1)}{\delta}\right)\right](\Gamma^*)^{-1} + r\left(\epsilon\right)R_{\text{upper}}, \end{aligned} \tag{EC.3.11}$$

where $C_4$ is a universal constant, $R_{\text{upper}} = \max\left\{\left\lceil \log_2 \frac{4}{\alpha_\varepsilon}\right\rceil, \left\lceil \log_2 \frac{4}{\beta_\varepsilon}\right\rceil\right\}$, and the third inequality comes from Equation (EC.2.57) □

## EC.4. Technical Skills

(ZK L: It seems that this section can be removed, or we can find some other technical skills that have been used in our proof.) (CH: Did we use these two lemmas?)(ZK L: The first one is used in Section EC.2.0.4 and the second one is used in equation (EC.2.4).) (CH: Let's keep them for now.)

LEMMA EC.4.1.

$$r \geq \log_2 \frac{1}{|\Delta|} \Rightarrow C_{\delta/K}(r) = \varepsilon_r = 2^{-r} \leq |\Delta|. \tag{EC.4.1}$$

LEMMA EC.4.2. *(Inversion Reverses Loewner orders) Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{d\times d}$. Suppose that $\boldsymbol{A} \succeq \boldsymbol{B}$ and $\boldsymbol{B}$ is invertible, we have*

$$\boldsymbol{A}^{-1} \preceq \boldsymbol{B}^{-1}. \tag{EC.4.2}$$

Proof. Note that it suffices to show that $\|\boldsymbol{x}\|^2_{\boldsymbol{B}^{-1}} - \|\boldsymbol{x}\|^2_{\boldsymbol{A}^{-1}} = \|\boldsymbol{x}\|^2_{\boldsymbol{B}^{-1}-\boldsymbol{A}^{-1}} \geq 0$ for any $\boldsymbol{x} \in \mathbb{R}^d$. Let $\boldsymbol{x} \in \mathbb{R}^d$. Then, by the Cauchy-Schwarz inequality,

$$\|\boldsymbol{x}\|^2_{\boldsymbol{A}^{-1}} = \left\langle \boldsymbol{x}, \boldsymbol{A}^{-1}\boldsymbol{x} \right\rangle \leq \|\boldsymbol{x}\|_{\boldsymbol{B}^{-1}} \left\|\boldsymbol{A}^{-1}\boldsymbol{x}\right\|_{\boldsymbol{B}} \leq \|\boldsymbol{x}\|_{\boldsymbol{B}^{-1}} \left\|\boldsymbol{A}^{-1}\boldsymbol{x}\right\|_{\boldsymbol{A}} = \|\boldsymbol{x}\|_{\boldsymbol{B}^{-1}} \|\boldsymbol{x}\|_{\boldsymbol{A}^{-1}}. \tag{EC.4.3}$$

Hence $\|\boldsymbol{x}\|_{\boldsymbol{A}^{-1}} \leq \|\boldsymbol{x}\|_{\boldsymbol{B}^{-1}}$ for all $\boldsymbol{x}$, which completes the claim. □

LEMMA EC.4.3. *For any arm $i \in \mathcal{A}_I \cap G_\varepsilon \backslash \{1\}$, we have*

$$\frac{\min_{\boldsymbol{p}\in S_K} \max_{i\in\mathcal{A}_I\cap G_\varepsilon\backslash\{1\}} \|\boldsymbol{y}_{1,i}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}}{\min_{\boldsymbol{p}\in S_K} \min_{i\in\mathcal{A}_I\cap G_\varepsilon\backslash\{1\}} \|\boldsymbol{y}_{1,i}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}} \leq \frac{dL_1}{{\mathcal{G}_{\mathcal{Y}}}^2 L_2}. \tag{EC.4.4}$$

Proof. For any arm $i \in [K]$, from a perspective of geometry quantity, let $\text{conv}(\mathcal{A} \cup -\mathcal{A})$ denote the convex hull of symmetric $\mathcal{A} \cup -\mathcal{A}$. Then for any set $\mathcal{Y} \subset \mathbb{R}^d$ define the gauge of $\mathcal{Y}$ as

$$\mathcal{G}_{\mathcal{Y}} = \max\left\{c > 0 : c\mathcal{Y} \subset \text{conv}(\mathcal{A} \cup -\mathcal{A})\right\}. \tag{EC.4.5}$$

We can provide a natural upper bound for $\min_{\boldsymbol{p}\in S_K} \max_{i\in\mathcal{A}_I\cap G_\varepsilon\backslash\{1\}} \|\boldsymbol{y}_{1,i}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}}$, given by

$$\begin{aligned}
\min_{\boldsymbol{p}\in S_K} \max_{i\in\mathcal{A}_I\cap G_\varepsilon\backslash\{1\}} \|\boldsymbol{y}_{1,i}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}} &\leq \min_{\boldsymbol{p}\in S_K} \max_{y\in\mathcal{Y}(\mathcal{A}_I)} \|\boldsymbol{y}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}} \\
&= \frac{1}{{\mathcal{G}_{\mathcal{Y}}}^2} \min_{\boldsymbol{p}\in S_K} \max_{y\in\mathcal{Y}(\mathcal{A}_I)} \|\boldsymbol{y}\mathcal{G}_{\mathcal{Y}}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}} \\
&\leq \frac{1}{{\mathcal{G}_{\mathcal{Y}}}^2} \min_{\boldsymbol{p}\in S_K} \max_{\boldsymbol{a}\in\text{conv}(\mathcal{A}\cup-\mathcal{A})} \|\boldsymbol{a}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}} \\
&= \frac{1}{{\mathcal{G}_{\mathcal{Y}}}^2} \min_{\boldsymbol{p}\in S_K} \max_{i\in\mathcal{A}_I} \|\boldsymbol{a}_i\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}} \\
&\leq \frac{d}{{\mathcal{G}_{\mathcal{Y}}}^2},
\end{aligned} \tag{EC.4.6}$$

where the third line follows from the fact that the maximum value of a convex function on a convex set must occur at a vertex. With the Kiefer-Wolfowitz Theorem for the $G$-optimal design, the last inequality is achieved. Furthermore, for any arm $i \in \mathcal{A}_I \cap G_\varepsilon \backslash \{1\}$, we have

$$\begin{aligned}
\min_{\boldsymbol{p}\in S_K} \min_{i\in\mathcal{A}_I\cap G_\varepsilon\backslash\{1\}} \|\boldsymbol{y}_{1,i}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}} &\geq \min_{\boldsymbol{p}\in S_K} \min_{i\in\mathcal{A}_I\cap G_\varepsilon\backslash\{1\}} \text{eig}_{\min}(\boldsymbol{V}_{\boldsymbol{p}}^{-1}) \|\boldsymbol{y}_{1,i}\|_2^2 \\
&= \min_{\boldsymbol{p}\in S_K} \min_{i\in\mathcal{A}_I\cap G_\varepsilon\backslash\{1\}} \frac{1}{\text{eig}_{\max}(\boldsymbol{V}_{\boldsymbol{p}})} \|\boldsymbol{y}_{1,i}\|_2^2 \\
&\geq \frac{1}{\max_{i\in\mathcal{A}_I} \|\boldsymbol{a}_i\|_2} \min_{\boldsymbol{p}\in S_K} \min_{i\in\mathcal{A}_I\cap G_\varepsilon\backslash\{1\}} \|\boldsymbol{y}_{1,i}\|_2^2,
\end{aligned} \tag{EC.4.7}$$

where $\text{eig}_{\max}(\cdot)$ and $\text{eig}_{\min}(\cdot)$ are respectively the largest and smallest eigenvalue operators of a matrix. The first line follows from the Rayleigh Quotient and Rayleigh Theorem. The last line is derived by the relationship $\text{eig}_{\max}(\boldsymbol{V}_{\boldsymbol{p}}) \leq \max_{i\in\mathcal{A}_I} \|\boldsymbol{a}_i\|_2$. Recall the assumption in Theorem 5 that $\min_{i\in G_\varepsilon} \|\boldsymbol{a}_1 - \boldsymbol{a}_i\|^2 \geq L_2$ and the assumption ~~of realizable model in equation (??)~~ in Section 2.1 that $\|\boldsymbol{a}_i\|_2 \leq L_1$ for $\forall i \in [K]$, we have

$$\min_{\boldsymbol{p}\in S_K} \min_{i\in\mathcal{A}_I\cap G_\varepsilon\setminus\{1\}} \|\boldsymbol{y}_{1,i}\|^2_{\boldsymbol{V}_{\boldsymbol{p}}^{-1}} \geq \frac{L_2}{L_1}. \tag{EC.4.8}$$

Finally, combining inequalities (EC.4.6) and (EC.4.8) just completes the claim. □