Indexed Minimum Empirical Divergence-Based Algorithms for Linear Bandits

Anonymous authors
Paper under double-blind review

Abstract

The Indexed Minimum Empirical Divergence (IMED) algorithm is a highly effective approach that offers a stronger theoretical guarantee of the asymptotic optimality compared to the Kullback–Leibler Upper Confidence Bound (KL-UCB) algorithm for the multi-armed bandit problem. Additionally, it has been observed to empirically outperform UCB-based algorithms and Thompson Sampling. Despite its effectiveness, the generalization of this algorithm to contextual bandits with linear payoffs has remained elusive. In this paper, we present novel linear versions of the IMED algorithm, which we call the family of LinIMED algorithms. We demonstrate that LinIMED provides a $\widetilde{O}(d\sqrt{T})$ upper regret bound where d is the dimension of the context and T is the time horizon. Furthermore, extensive empirical studies reveal that LinIMED and its variants outperform widely-used linear bandit algorithms such as LinUCB and Linear Thompson Sampling in some regimes.

1 Introduction

The multi-armed bandit (MAB) problem (Lattimore & Szepesvári (2020)) is a classical topic in decision theory and reinforcement learning. Among the various subfields of bandit problems, the stochastic linear bandit is the most popular area due to its wide applicability in large-scale, real-world applications such as personalized recommendation systems (Li et al. (2010)), online advertising, and clinical trials. In the stochastic linear bandit model, at each time step t, the learner has to choose one arm A_t from the time-varying action set $\mathcal{A}_t$. Each arm $a \in \mathcal{A}_t$ has a corresponding context $x_{t,a} \in \mathbb{R}^d$, which is a d-dimensional vector. By pulling the arm $a \in \mathcal{A}_t$ at time step t, under the linear bandit setting, the learner will receive the reward $Y_{t,a}$, whose expected value satisfies $\mathbb{E}[Y_{t,a}|x_{t,a}] = \langle \theta^*, x_{t,a} \rangle$, where $\theta^* \in \mathbb{R}^d$ is an unknown parameter. The goal of the learner is to maximize his cumulative reward over a time horizon T, which also means minimizing the cumulative regret, defined as $R_T := \mathbb{E}\left[\sum_{t=1}^{T} \max_{a \in \mathcal{A}_t} Y_{t,a} - Y_{t,A_t}\right]$. The learner needs to balance the trade-off between the exploration of different arms (to learn their expected rewards) and the exploitation of the arm with the highest expected reward based on the available data.

1.1 Motivation and Related Work

The K-armed bandit setting is a special case of the linear bandit. There exist several good algorithms such as UCB1 (Auer et al. (2002)), Thompson Sampling (Agrawal & Goyal (2012)), and the Indexed Minimum Empirical Divergence (IMED) algorithm (Honda & Takemura (2015)) for this setting. There are three main families of asymptotically optimal multi-armed bandit algorithms based on different principles (Baudry et al. (2023)). However, among these algorithms, only IMED lacks an extension for contextual bandits with linear payoff. In the context of the varying arm setting of the linear bandit problem, the LinUCB algorithm in Li et al. (2010) is frequently employed in practice. It has a theoretical guarantee on the regret in the order of $O(d\sqrt{T}\log(T))$ using the confidence width as in OFUL (Abbasi-Yadkori et al. (2011)). Although the SupLinUCB algorithm introduced by Chu et al. (2011) uses phases to decompose the reward dependence of each time step and achieves an $\widetilde{O}(\sqrt{dT})$ (the $\widetilde{O}(\cdot)$ notation omits logarithmic factors in T) regret upper bound, its empirical performance falls short of both the algorithm in Li et al. (2010) and the Linear Thompson Sampling algorithm (Agrawal & Goyal (2013)) as mentioned in Lattimore & Szepesvári (2020, Chapter 22).

	Problem independent regret bound	Efficient for large finite arm sets?	Regret bound independent of K?	Principle that the algorithm is based on
OFUL (Abbasi-Yadkori et al. (2011))	$O(d\sqrt{T}\log(T))$	✗	✓	Optimism
LinUCB (Li et al. (2010))	Hard to analyze	Not Applicable	Unknown	Optimism
LinTS (Agrawal & Goyal (2013))	$O(d^{\frac{3}{2}}\sqrt{T}) \wedge O(d\sqrt{T\log(K)})$	✓	✓	Posterior sampling
SupLinUCB (Chu et al. (2011))	$O(\sqrt{dT\log^3(KT)})$	✓	✗	Optimism
LinUCB with OFUL's confidence bound	$O(d\sqrt{T}\log(T))$	✓	✓	Optimism
LinIMED-3 (this paper)	$O(d\sqrt{T}\log(T))$	✓	✓	Min. emp. divergence
SupLinIMED (this paper)	$O(\sqrt{dT\log^3(KT)})$	✓	✗	Min. emp. divergence

Table 1: Comparison of algorithms for linear bandits with varying arm sets

On the other hand, the Optimism in the Face of Uncertainty Linear (OFUL) bandit algorithm in Abbasi-Yadkori et al. (2011) achieves a regret upper bound of $\widetilde{O}(d\sqrt{T})$ through an improved analysis of the confidence bound using a martingale technique. However, it involves a bilinear optimization problem over the action set and the confidence ellipsoid when choosing the arm at each time. This is computationally expensive, unless the confidence ellipsoid is a convex hull of a finite set.

For randomized algorithms designed for the linear bandit problem, Agrawal & Goyal (2013) proposed the LinTS algorithm, which is in the spirit of Thompson Sampling (Thompson (1933)) and the confidence ellipsoid similar to that of LinUCB-like algorithms. This algorithm performs efficiently and achieves a regret upper bound of $O(d^{\frac{3}{2}}\sqrt{T} \wedge d\sqrt{T\log K})$, where K is the number of arms at each time step such that $|\mathcal{A}_t| = K$ for all t. Compared to LinUCB with OFUL's confidence width, it has an extra $O(\sqrt{d} \wedge \sqrt{\log K})$ term for the minimax regret upper bound.

Recently, MED-like (minimum empirical divergence) algorithms have come to the fore since these randomized algorithms have the property that the probability of selecting each arm is in closed form, which benefits downstream work such as offline evaluation with the inverse propensity score. Both MED in the sub-Gaussian environment and its deterministic version IMED have demonstrated superior performances over Thompson Sampling (Bian & Jun (2021), Honda & Takemura (2015)). Baudry et al. (2023) also shows MED has a close relation to Thompson Sampling. In particular, it is argued that MED and TS can be interpreted as two variants of the same exploration strategy. Bian & Jun (2021) also shows that probability of selecting each arm of MED in the sub-Gaussian case can be viewed as a closed-form approximation of the same probability as in Thompson Sampling. We take inspiration from the extension of Thompson Sampling to linear bandits and thus are motivated to extend MED-like algorithms to the linear bandit setting and prove regret bounds that are competitive vis-à-vis the state-of-the-art bounds.

Thus, this paper aims to answer the question of whether it is possible to devise an extension of the IMED algorithm for the linear bandit problem the varying arm set setting (for both infinite and finite arm sets) with a regret upper bound of $O(d\sqrt{T}\log T)$ which matches LinUCB with OFUL's confidence bound while being as efficient as LinUCB. The proposed family of algorithms, called LinIMED as well as SupLinIMED, can be viewed as generalizations of the IMED algorithm (Honda & Takemura (2015)) to the linear bandit setting. We prove that LinIMED and its variants achieve a regret upper bound of $\widetilde{O}(d\sqrt{T})$ and they perform efficiently, no worse than LinUCB. SupLinIMED has a regret bound of $\widetilde{O}(\sqrt{dT})$, but works only for instances with finite arm sets. In our empirical study, we found that the different variants of LinIMED perform better than LinUCB and LinTS for various synthetic and real-world instances under consideration.

Compared to OFUL, LinIMED works more efficiently. Compared to SupLinUCB, our LinIMED algorithm is significantly simpler, and compared to LinUCB with OFUL's confidence bound, our empirical performance is better. This is because in our algorithm, the exploitation term and exploration term are decoupling and this leads to a finer control while tuning the hyperparameters in the empirical study.

Compared to LinTS, our algorithm's (specifically LinIMED-3) regret bound is superior, by an order of $O(\sqrt{d} \wedge \sqrt{\log K})$. Since fixed arm setting is a special case of finite varying arm setting, our result is more general than other fixed-arm linear bandit algorithms like Spectral Eliminator (Valko et al. (2014)) and PEGOE (Lattimore & Szepesvári (2020, Chapter 22)). Finally, we observe that since the index used in LinIMED has a similar form to the index used in the Information Directed Sampling (IDS) procedure in Kirschner et al. (2021) (which is known to be asymptotically optimal but more difficult to compute), LinIMED performs significantly better on the "End of Optimism" example in Lattimore & Szepesvari (2017). We summarize the comparisons of LinIMED to other linear bandit algorithms in Table 1.

2 Problem Statement

Notations: For any d dimensional vector $x \in \mathbb{R}^d$ and a $d \times d$ positive definite matrix A, we use $\|x\|_A$ to denote the Mahalanobis norm $\sqrt{x^\top Ax}$. We use $a \wedge b$ (resp. $a \vee b$) to represent the minimum (resp. maximum) of two real numbers a and b.

The Stochastic Linear Bandit Model: In the stochastic linear bandit model, the learner chooses an arm A_t at each round t from the arm set $\mathcal{A}_t = \{a_{t,1}, a_{t,2}, \ldots\} \subseteq \mathbb{R}$, where we assume the cardinality of each arm set $\mathcal{A}_t$ can be potentially infinite such that $|\mathcal{A}_t| = \infty$ for all $t \geq 1$. Each arm $a \in \mathcal{A}_t$ at time t has a corresponding context $x_{t,a}$, which is known to the learner. After choosing arm A_t, the environment reveals the reward

$$Y_t = \langle \theta^*, X_t \rangle + \eta_t$$

to the learner where $X_t := x_{t,A_t}$ is the corresponding context of the arm A_t, $\theta^* \in \mathbb{R}^d$ is an unknown coefficient of the linear model, η_t is an R-sub-Gaussian noise conditioned on $\{A_1, A_2, \ldots, A_t, Y_1, Y_2, \ldots, Y_{t-1}\}$ such that for any $\lambda \in \mathbb{R}$, almost surely,

$$\mathbb{E}\left[\exp(\lambda \eta_t) \mid A_1, A_2, \ldots, A_t, Y_1, Y_2, \ldots, Y_{t-1}\right] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right) .$$

Denote $a_t^* := \arg\max_{a \in \mathcal{A}_t} \langle \theta^*, x_{t,a} \rangle$ as the arm with the largest reward at time t. The goal of the learner is to minimize the expected cumulative regret over the horizon T. The (expected) cumulative regret is defined as

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \langle \theta^*, x_{t,a_t^*} \rangle - \langle \theta^*, X_t \rangle\right] .$$

Assumption 1. *For each time t, we assume that $\|X_t\| \leq L$, and $\|\theta^*\| \leq S$ for some fixed $L, S > 0$. We also assume that $\Delta_{t,b} := \max_{a \in \mathcal{A}_t} \langle \theta^*, x_{t,a} \rangle - \langle \theta^*, x_{t,b} \rangle \leq 1$ for each arm $b \in \mathcal{A}_t$ and time t.*

3 Description of LinIMED Algorithms

In the pseudocode of Algorithm 1, for each time step t, in Line 4, we use the improved confidence bound of θ^* as in Abbasi-Yadkori et al. (2011) to calculate the confidence bound $\beta_{t-1}(\gamma)$. After that, for each arm $a \in \mathcal{A}_t$, in Lines 7 to 8, the empirical gap between the highest empirical reward and the empirical reward of arm a is estimated as

$$\hat{\Delta}_{t,a} = \begin{cases} \max_{j \in \mathcal{A}_t} \langle \hat{\theta}_{t-1}, x_{t,j} \rangle - \langle \hat{\theta}_{t-1}, x_{t,a} \rangle & \text{if LinIMED-1,2} \\ \max_{j \in \mathcal{A}_t} \mathrm{UCB}_t(j) - \mathrm{UCB}_t(a) & \text{if LinIMED-3} \end{cases}$$

Then, in Lines 10 to 12, with the use of the confidence width of $\beta_{t-1}(\gamma)$, we can compute the index $I_{t,a}$ for the empirical best arm $a = \arg\max_{j \in \mathcal{A}_t} \hat{\mu}_{t,a}$ (for LinIMED-1,2) or the highest UCB arm $a = \arg\max_{j \in \mathcal{A}_t} \mathrm{UCB}_j(a)$ (for LinIMED-3). The different versions of LinIMED encourage different amounts of exploitation. For the other arms, in Line 14, the index is defined and computed as

$$I_{t,a} = \frac{\hat{\Delta}_{t,a}^2}{\beta_{t-1}(\gamma)\|x_{t,a}\|^2_{V_{t-1}^{-1}}} + \log \frac{1}{\beta_{t-1}(\gamma)\|x_{t,a}\|^2_{V_{t-1}^{-1}}}.$$

Then with all the indices of the arms calculated, in Line 17, we choose the arm A_t with the minimum index such that $A_t = \arg\min_{a \in \mathcal{A}_t} I_{t,a}$ (where ties are broken arbitrarily) and the agent receives its reward. Finally,

Algorithm 1 LinIMED-x for $x \in \{1,2,3\}$

1: **Input:** Dimension d, Regularization parameter λ, Bound S on $\|\theta^*\|$, Sub-Gaussian parameter R, Concentration parameter γ of θ^*, Bound L on $\|x_{t,a}\|$ for all $t \geq 1$ and $a \in \mathcal{A}_t$, Constant $C \geq 1$.
2: **Initialize:** $V_0 = \lambda I_{d\times d}$, $W_0 = 0_{d\times 1}$(all zeros vector with d dimensions), $\hat{\theta}_0 = V_0^{-1}W_0$
3: **for** $t = 1, 2, \ldots T$ **do**
4: Receive the arm set $\mathcal{A}_t$ and compute $\beta_{t-1}(\gamma) = (R\sqrt{d\log(\frac{1+(t-1)L^2/\lambda}{\gamma})} + \sqrt{\lambda}S)^2$.
5: **for** $a \in \mathcal{A}_t$ **do**
6: Compute:
7: $\hat{\mu}_{t,a} = \langle \hat{\theta}_{t-1}, x_{t,a}\rangle$ and $\hat{\Delta}_{t,a} = \max_{j\in\mathcal{A}_t} \hat{\mu}_{t,j} - \hat{\mu}_{t,a}$ (LinIMED-1, 2)
8: $\mathrm{UCB}_t(a) = \langle \hat{\theta}_{t-1}, x_{t,a}\rangle + \sqrt{\beta_{t-1}(\gamma)}\|x_{t,a}\|_{V_{t-1}^{-1}}$ and
$\hat{\Delta}_{t,a} = \max_{j\in\mathcal{A}_t} \mathrm{UCB}_t(j) - \mathrm{UCB}_t(a)$ (LinIMED-3)
9: **if** $a = \arg\max_{j\in\mathcal{A}_t} \hat{\mu}_{t,a}$ (LinIMED-1,2) or $a = \arg\max_{j\in\mathcal{A}_t} \mathrm{UCB}_t(a)$ (LinIMED-3) **then**
10: $I_{t,a} = -\log(\beta_{t-1}(\gamma)\|x_{t,a}\|^2_{V_{t-1}^{-1}})$ (LinIMED-1)
11: $I_{t,a} = \log T \wedge (-\log(\beta_{t-1}(\gamma)\|x_{t,a}\|^2_{V_{t-1}^{-1}}))$ (LinIMED-2)
12: $I_{t,a} = \log \frac{C}{\max_{a\in\mathcal{A}_t} \hat{\Delta}^2_{t,a}} \wedge (-\log(\beta_{t-1}(\gamma)\|x_{t,a}\|^2_{V_{t-1}^{-1}}))$ (LinIMED-3)
13: **else**
14: $I_{t,a} = \frac{\hat{\Delta}^2_{t,a}}{\beta_{t-1}(\gamma)\|x_{t,a}\|^2_{V_{t-1}^{-1}}} - \log(\beta_{t-1}(\gamma)\|x_{t,a}\|^2_{V_{t-1}^{-1}})$
15: **end if**
16: **end for**
17: Pull the arm $A_t = \arg\min_{a\in\mathcal{A}_t} I_{t,a}$ (ties are broken arbitrarily) and receive its reward Y_t.
18: **Update:**
19: $V_t = V_{t-1} + X_t X_t^\top$
20: $W_t = W_{t-1} + Y_t X_t$
21: $\hat{\theta}_t = V_t^{-1}W_t$
22: **end for**

in Lines 19 to 21, we use ridge regression to estimate the unknown θ^* as $\hat{\theta}_t$ and update the matrix V_t and the vector W_t. After that, the algorithm iterates to the next time step until the time horizon T. From the pseudo-code, we observe that the only differences between the three algorithms are the way that the square gap, which plays the role of the empirical divergence, is estimated and the index of the empirically best arm. The latter point implies that we encourage the empirically best arm to be selected more often in LinIMED-2 and LinIMED-3 compared to LinIMED-1; in other words, we encourage more exploitation in LinIMED-2 and LinIMED-3. Similar to the core spirit of IMED algorithm Honda & Takemura (2015), the first term of our index $I_{t,a}$ for LinIMED-1 algorithm is $\hat{\Delta}^2_{t,a}/(\beta_{t-1}(\gamma)\|x_{t,a}\|^2_{V_{t-1}^{-1}})$, this is the term controls the exploitation, while the second term $-\log(\beta_{t-1}(\gamma)\|x_{t,a}\|^2_{V_{t-1}^{-1}})$ controls the exploration in our algorithm.

3.1 Description of the SupLinIMED Algorithm

Now we consider the case in which the arm set $\mathcal{A}_t$ at each time t is finite but still time-varying. In particular, $\mathcal{A}_t = \{a_{t,1}, a_{t,2}, \ldots, a_{t,K}\} \subseteq \mathbb{R}$ are sets of constant size K such that $|\mathcal{A}_t| = K < \infty$. In the pseudocode of Algorithm 3, we apply the SupLinUCB framework (Chu et al., 2011), leveraging Algorithm 2 as a subroutine within each phase. This ensures the independence of the choice of the arm from past observations of rewards, thereby yielding a concentration inequality in the estimated reward (see Lemma 1 in Chu et al. (2011)) that converges to within $\sqrt{d}$ proximity of the unknown expected reward in a finite arm setting. As a result, the regret yields an improvement of $\sqrt{d}$ ignoring the logarithmic factor. At each time step t and phase s, in Line 5, we utilize the BaseLinUCB Algorithm as a subroutine to calculate the sample mean and confidence width since we also need these terms to calculate the IMED-style indices of each arm. In Lines 6–9 (Step 1), if the width of each arm is smaller than $\frac{1}{\sqrt{T}}$, then we choose the arm with the smaller IMED-style index. In Lines 10–12 (Step 2), the framework is the same as in SupLinUCB (Chu et al. (2011)), if the width of each arm is

Algorithm 2 BaseLinUCB

Input: $\gamma = \frac{1}{2t^2}$, $\alpha = \sqrt{\frac{1}{2}\ln\frac{2TK}{\gamma}}$, $\Psi_t \subseteq \{1, 2, ..., t-1\}$
$V_t = I_d + \sum_{\tau\in\Psi_t} x_{\tau,A_\tau}^T x_{\tau,A_\tau}$
$b_t = \sum_{\tau\in\Psi_t} Y_{\tau,A_\tau} x_{\tau,A_\tau}$
$\hat{\theta}_t = V_t^{-1} b_t$
Observe K arm features $x_{t,1}, x_{t,2}, \ldots, x_{t,K} \in \mathbb{R}^d$
for $a \in [K]$ **do**
  $w_{t,a} = \alpha\sqrt{x_{t,a}^T V_t^{-1} x_{t,a}}$
  $\hat{Y}_{t,a} = \langle\hat{\theta}_t, x_{t,a}\rangle$
end for

Algorithm 3 SupLinIMED

1: **Input:** $T \in \mathbb{N}$, $S = \log T$, $\Psi_t^s = \emptyset$ for all $s \in [\log T], t \in [T]$
2: **for** $t = 1, 2, ..., T$ **do**
3:   $s \leftarrow 1$ and $\hat{\mathcal{A}}_1 \leftarrow [K]$
4:   **repeat**
5:     Use BaseLinUCB with Ψ_t^s to calculate the width $w_{t,a}^s$ and sample mean $\hat{Y}_{t,a}^s$ for all $a \in \hat{\mathcal{A}}_s$.
6:     **if** $w_{t,a}^s \leq \frac{1}{\sqrt{T}}$ for all $a \in \hat{\mathcal{A}}_s$ **then**
7:       choose $A_t = \arg\min_{a\in\hat{\mathcal{A}}_s} I_{t,a}$ where $I_{t,a}$ is the same index function as in LinIMED algorithm:
8:       Calculate the index

$$I_{t,a} = \begin{cases} \log(2T) \wedge (-\log((w_{t,a}^s)^2)) & \textbf{If } a = \arg\max_{b\in\hat{\mathcal{A}}_s} \hat{Y}_{t,b}^s \\ (\frac{\hat{\Delta}_{t,a}^s}{w_{t,a}^s})^2 - \log((w_{t,a}^s)^2) & \textbf{otherwise} \end{cases} \quad \text{where} \quad \hat{\Delta}_{t,a}^s := \max_{b\in\hat{\mathcal{A}}_s} \hat{Y}_{t,b}^s - \hat{Y}_{t,a}^s \ .$$

9:       Keep the same index sets at all levels: $\Psi_{t+1}^{s'} \leftarrow \Psi_t^{s'}$ for all $s' \in [S]$. ← **Step 1**
10:     **else if** $w_{t,a}^s \leq 2^{-s}$ for all $a \in \hat{\mathcal{A}}_s$ **then**
11:       $\hat{\mathcal{A}}_{s+1} \leftarrow \left\{a \in \hat{\mathcal{A}}_s \ : \ \hat{Y}_{t,a}^s + w_{t,a}^s \geq \max_{a'\in\hat{\mathcal{A}}_s}(\hat{Y}_{t,a'}^s + w_{t,a'}^s) - 2^{1-s}\right\}$
12:       $s \leftarrow s+1$ ← **Step 2**
13:     **else**
14:       Choose $A_t \in \hat{\mathcal{A}}_s$ such that $w_{t,A_t}^s > 2^{-s}$
15:       Update the index sets at all levels: $\Psi_{t+1}^{s'} \leftarrow \Psi_t^{s'} \cup \{t\}$ if $s = s'$; $\Psi_{t+1}^{s'} \leftarrow \Psi_t^{s'}$ if $s \neq s'$ ← **Step 3**
16:     **end if**
17:   **until** an action A_t is found
18: **end for**

smaller than 2^{-s} but there exist some arms with widths larger than $\frac{1}{\sqrt{T}}$, then in Line 11 the "unpromising" arms will be eliminated until the width of each arm is smaller enough to satisfy the condition in Line 6. Otherwise, if there exist any arms with widths that are larger than 2^{-s}, in Lines 14–15 (Step 3), we choose one such arm and record the context and reward of this arm to the next layer Ψ_{t+1}^s.

3.2 Relation to the IMED algorithm of Honda & Takemura (2015)

The IMED algorithm is a deterministic algorithm for the K-armed bandit problem. At each time step t, it chooses the arm a with the minimum index, i.e.,

$$a = \arg\min_{i\in[K]} \left\{T_i(t) D_{\inf}(\hat{F}_i(t), \hat{\mu}^*(t)) + \log T_i(t)\right\} \ , \tag{1}$$

where $T_i(t) = \sum_{s=1}^{t-1} \mathbb{1}\{A_t = a\}$ is the total arm pulls of the arm i until time t and $D_{\inf}(\hat{F}_i(t), \hat{\mu}^*(t))$ is some divergence measure between the empirical distribution of the sample mean for arm i and the arm with the highest sample mean. More precisely, $D_{\inf}(F,\mu) := \inf_{G\in\mathcal{G}:\mathbb{E}(G)\leq\mu} D(F\|G)$ and $\mathcal{G}$ is the family of distributions supported on $(-\infty, 1]$. As shown in Honda & Takemura (2015), its asymptotic bound is even better than KL-UCB (Garivier & Cappé (2011)) algorithm and can be extended to semi-bounded support models such as $\mathcal{G}$. Also, this algorithm empirically outperforms the Thompson Sampling algorithm as shown in Honda & Takemura (2015). However, the linear extension of IMED algorithm was, prior to our work, still unknown. In our design of LinIMED algorithm, we replace the optimized KL-divergence measure in IMED in Eqn. (1) with the squared gap between the sample mean of the arm i and the arm with the maximum sample mean. This choice simplifies our analysis and does not adversely affect the regret bound. On the other hand, we view the term $1/T_i(t)$ as the variance of the sample mean of arm i at time t; then in this spirit, we use $\beta_{t-1}(\gamma)\|x_{t,a}\|^2_{V_{t-1}^{-1}}$ as the variance of the sample mean (which is $\langle\hat{\theta}_{t-1}, x_{t,a}\rangle$) of arm a at time t. We choose $\hat{\Delta}^2_{t,a}/(\beta_{t-1}(\gamma)\|x_{t,a}\|^2_{V_{t-1}^{-1}})$ instead of the KL-divergence approximation for the index since in the classical linear bandit setting, the noise is sub-Gaussian and it is known that the KL-divergence of two Gaussian random variables with the same variance ($\mathrm{KL}(\mathcal{N}(\mu_1,\sigma^2),\mathcal{N}(\mu_2,\sigma^2)) = \frac{(\mu_1-\mu_2)^2}{2\sigma^2}$) has a closed form expression similar to $\hat{\Delta}^2_{t,a}/(\beta_{t-1}(\gamma)\|x_{t,a}\|^2_{V_{t-1}^{-1}})$ ignoring the constant $\frac{1}{2}$.

4 Theorem Statements

Theorem 1. *Under Assumption 1, the assumption that $\langle\theta^*, x_{t,a}\rangle \geq 0$ for all $t \geq 1$ and $a \in \mathcal{A}_t$, and the assumption that $\sqrt{\lambda}S \geq 1$, the regret of the LinIMED-1 algorithm is upper bounded as follows:*

$$R_T \leq O\big(d\sqrt{T}\log^{\frac{3}{2}}(T)\big).$$

A proof sketch of Theorem 1 is provided in Section 5.

Theorem 2. *Under Assumption 1, and the assumption that $\sqrt{\lambda}S \geq 1$, the regret of the LinIMED-2 algorithm is upper bounded as follows:*

$$R_T \leq O\big(d\sqrt{T}\log^{\frac{3}{2}}(T)\big).$$

Theorem 3. *Under Assumption 1, the assumption that $\sqrt{\lambda}S \geq 1$, and that C in Line 12 is a constant, the regret of the LinIMED-3 algorithm is upper bounded as follows:*

$$R_T \leq O\big(d\sqrt{T}\log(T)\big).$$

Theorem 4. *Under Assumption 1, the assumption that $L = S = 1$, the regret of the SupLinIMED algorithm (which is applicable to linear bandit problems with $K < \infty$ arms) is upper bounded as follows:*

$$R_T \leq O\big(\sqrt{dT\log^3(KT)}\big).$$

The upper bounds on the regret of LinIMED and its variants are all of the form $\widetilde{O}(d\sqrt{T})$, which, ignoring the logarithmic term, is the same as OFUL algorithm (Abbasi-Yadkori et al. (2011)). Compared to LinTS, it has an advantage of $O(\sqrt{d}\wedge\sqrt{\log K})$. Also, these upper bounds do not depend on the number of arms K, which means it can be applied to linear bandit problems with a large arm set (including infinite arm sets). One observes that LinIMED-2 and LinIMED-3 do not require the additional assumption that $\langle\theta^*, x_{t,a}\rangle \geq 0$ for all $t \geq 1$ and $a \in \mathcal{A}_t$ to achieve the $\widetilde{O}(d\sqrt{T})$ upper regret bound. It is difficult to prove the regret bound for the LinIMED-1 algorithm without this assumption since in our proof we need to use that $\langle\theta^*, X_t\rangle \geq 0$ for any time t to bound the F_1 term. On the other hand, LinIMED-2 and LinIMED-3 encourage more exploitations in terms of the index of the empirically best arm at each time without adversely influencing the regret bound; this will accelerate the learning with well-preprocessed datasets. The regret bound of LinIMED-3, in fact, matches that of LinUCB with OFUL's confidence bound. In the proof, we will extensively use a technique known as the "peeling device" (Lattimore & Szepesvári, 2020, Chapter 9). This analytical technique, commonly used in the theory of bandit algorithms, involves the partitioning of the range of some random variable into several pieces, then using the basic fact that $\mathbb{P}(A \cap (\cup_{i=1}^{\infty} B_i)) \leq \sum_{i=1}^{\infty} \mathbb{P}(A \cap B_i)$, we can utilize the more *refined range* of the random variable to derive desired bounds.

Finally, Theorem 4 says that when the arm set is finite, we can use the framework of SupLinUCB (Chu et al., 2011) with our LinIMED index $I_{t,a}$ to achieve a regret bound of the order of $\widetilde{O}(\sqrt{dT})$, which is $\sqrt{d}$ better than the regret bounds yielded by the family of LinIMED algorithms (ignoring the logarithmic terms). The proof is provided in Appendix D.

5 Proof Sketch of Theorem 1

We choose to present the proof sketch of Theorem 1 since it contains the main ingredients for all the theorems in the preceding section. Before presenting the proof, we introduce the following lemma and corollary.

Lemma 1. *(Abbasi-Yadkori et al. (2011, Theorem 2)) Under Assumption 1, for any time step $t \geq 1$ and any $\gamma > 0$, we have*

$$\mathbb{P}\big(\|\hat{\theta}_{t-1} - \theta^*\|_{V_{t-1}} \leq \sqrt{\beta_{t-1}(\gamma)}\big) \geq 1 - \gamma.$$

This lemma illustrates that the true parameter θ^* lies in the ellipsoid centered at $\hat{\theta}_{t-1}$ with high probability, which also states the width of the confidence bound.

The second is a corollary of the elliptical potential count lemma in Abbasi-Yadkori et al. (2011):

Corollary 1. *(Corollary of Lattimore & Szepesvári (2020, Exercise 19.3)) Assume that $V_0 = \lambda I$ and $\|X_t\| \leq L$ for $t \in [T]$, for any constant $0 < m \leq 2$, the following holds:*

$$\sum_{t=1}^{T} \mathbb{1}\left\{\|X_t\|^2_{V_{t-1}^{-1}} \geq m\right\} \leq \frac{6d}{m} \log\left(1 + \frac{2L^2}{\lambda m}\right).$$

We remark that this lemma is slightly stronger than the classical elliptical potential lemma since it reveals information about the upper bound of frequency that $\|X_t\|^2_{V_{t-1}^{-1}}$ exceeds some value m. Equipped with this lemma, we can perform the peeling device on $\|X_t\|^2_{V_{t-1}^{-1}}$ in our proof of the regret bound, which is a novel technique to the best of our knowledge.

Proof. First we define a_t^* as the best arm in time step t such that $a_t^* = \arg\max_{a \in \mathcal{A}_t} \langle \theta^*, x_{t,a} \rangle$, and use $x_t^* := x_{t,a_t^*}$ denote its corresponding context. Let $\Delta_t := \langle \theta^*, x_t^* \rangle - \langle \theta^*, X_t \rangle$ denote the regret in time t. Define the following events:

$$B_t := \big\{\|\hat{\theta}_{t-1} - \theta^*\|_{V_{t-1}} \leq \sqrt{\beta_{t-1}(\gamma)}\big\}, \quad C_t := \big\{\max_{b \in \mathcal{A}_t} \langle \hat{\theta}_{t-1}, x_{t,b} \rangle > \langle \theta^*, x_t^* \rangle - \delta\big\}$$
$$D_t := \big\{\hat{\Delta}_{t,A_t} \geq \varepsilon\big\}.$$

where δ and ε are free parameters set to be $\delta = \frac{\Delta_t}{\sqrt{\log T}}$ and $\varepsilon = (1 - \frac{2}{\sqrt{\log T}})\Delta_t$ in this proof sketch.

Then the expected regret $R_T = \mathbb{E} \sum_{t=1}^{T} \Delta_t$ can be partitioned by events B_t, C_t, D_t such that:

$$R_T = \underbrace{\mathbb{E} \sum_{t=1}^{T} \Delta_t \cdot \mathbb{1}\left\{B_t, C_t, D_t\right\}}_{=:F_1} + \underbrace{\mathbb{E} \sum_{t=1}^{T} \Delta_t \cdot \mathbb{1}\left\{B_t, C_t, \overline{D}_t\right\}}_{=:F_2} + \underbrace{\mathbb{E} \sum_{t=1}^{T} \Delta_t \cdot \mathbb{1}\left\{B_t, \overline{C}_t\right\}}_{=:F_3} + \underbrace{\mathbb{E} \sum_{t=1}^{T} \Delta_t \cdot \mathbb{1}\left\{\overline{B}_t\right\}}_{=:F_4}.$$

For F_1, from the event C_t and the fact that $\langle \theta^*, x_t^* \rangle = \Delta_t + \langle \theta^*, X_t \rangle \geq \Delta_t$ (here is where we use that $\langle \theta^*, x_{t,a} \rangle \geq 0$ for all t and a), we obtain $\max_{b \in \mathcal{A}_t} \langle \hat{\theta}_{t-1}, x_{t,b} \rangle > (1 - \frac{1}{\sqrt{\log T}})\Delta_t$. For convenience, define $\hat{A}_t := \arg\max_{b \in \mathcal{A}_t} \langle \hat{\theta}_{t-1}, x_{t,b} \rangle$ as the empirically best arm at time step t, where ties are broken arbitrarily, then use $\hat{X}_t$ to denote the corresponding context of the arm $\hat{A}_t$. Therefore from the Cauchy–Schwarz inequality, we have $\|\hat{\theta}_{t-1}\|_{V_{t-1}}\|\hat{X}_t\|_{V_{t-1}^{-1}} \geq \langle\hat{\theta}_{t-1}, \hat{X}_t\rangle > (1-\frac{1}{\sqrt{\log T}})\Delta_t$. This implies that

$$\|\hat{X}_t\|_{V_{t-1}^{-1}} \geq \frac{(1-\frac{1}{\sqrt{\log T}})\Delta_t}{\|\hat{\theta}_{t-1}\|_{V_{t-1}}} \ . \tag{2}$$

On the other hand, we claim that $\|\hat{\theta}_{t-1}\|_{V_{t-1}}$ can be upper bounded as $O(\sqrt{T})$. This can be seen from the fact that $\|\hat{\theta}_{t-1}\|_{V_{t-1}} = \|\hat{\theta}_{t-1}-\theta^*+\theta^*\|_{V_{t-1}} \leq \|\hat{\theta}_{t-1}-\theta^*\|_{V_{t-1}} + \|\theta^*\|_{V_{t-1}}$. Since the event B_t holds, we know the first term is upper bounded by $\sqrt{\beta_{t-1}(\gamma)}$, and since the largest eigenvalue of the matrix V_{t-1} is upper bounded by $\lambda+TL$ and $\|\theta^*\| \leq S$, the second term is upper bounded by $S\sqrt{\lambda+TL^2}$. Hence, $\|\hat{\theta}_{t-1}\|_{V_{t-1}}$ is upper bounded by $O(\sqrt{T})$. Then one can substitute this bound back into Eqn. (2), and this yields

$$\|\hat{X}_t\|_{V_{t-1}^{-1}} \geq \Omega\Big(\frac{1}{\sqrt{T}}\Big(1-\frac{1}{\sqrt{\log T}}\Big)\Delta_t\Big) \ . \tag{3}$$

Furthermore, by our design of the algorithm, the index of A_t is not larger than the index of the arm with the largest empirical reward at time t. Hence,

$$I_{t,A_t} = \frac{\hat{\Delta}_{t,A_t}^2}{\beta_{t-1}(\gamma)\|X_t\|_{V_{t-1}^{-1}}^2} + \log\frac{1}{\beta_{t-1}(\gamma)\|X_t\|_{V_{t-1}^{-1}}^2} \leq \log\frac{1}{\beta_{t-1}(\gamma)\|\hat{X}_t\|_{V_{t-1}^{-1}}^2} \ . \tag{4}$$

In the following, we set γ as well as another free parameter Γ as follows:

$$\Gamma = \frac{d\log^{\frac{3}{2}} T}{\sqrt{T}} \quad \text{and} \quad \gamma = \frac{1}{t^2}, . \tag{5}$$

If $\|X_t\|_{V_{t-1}^{-1}}^2 \geq \frac{\Delta_t^2}{\beta_{t-1}(\gamma)}$, by using Corollary 1 with the choice in Eqn. (5), the upper bound of F_1 in this case is $O\big(d\sqrt{T\log T}\big)$. Otherwise, using the event D_t and the bound in (3), we deduce that for all T sufficiently large, we have $\|X_t\|_{V_{t-1}^{-1}}^2 \geq \Omega\big(\frac{\Delta_t^2}{\beta_{t-1}(\gamma)\log(T/\Delta_t^2)}\big)$. Therefore by using Corollary 1 and the "peeling device" (Lattimore & Szepesvári, 2020, Chapter 9) on Δ_t such that $2^{-l} < \Delta_t \leq 2^{-l+1}$ for $l = 1, 2, \ldots, \lceil Q\rceil$ where $Q = -\log_2\Gamma$ and Γ is chosen as in Eqn. (5). Now consider,

$$\begin{aligned}
F_1 &\leq O(1) + \mathbb{E}\sum_{t=1}^{T}\Delta_t\cdot\mathbb{1}\left\{\|X_t\|_{V_{t-1}^{-1}}^2 \geq \Omega\Big(\frac{\Delta_t^2}{\beta_{t-1}(\gamma)\log(T/\Delta_t^2)}\Big)\right\} \\
&\leq O(1)+T\Gamma+\mathbb{E}\sum_{t=1}^{T}\sum_{l=1}^{\lceil Q\rceil}\Delta_t\cdot\mathbb{1}\left\{\|X_t\|_{V_{t-1}^{-1}}^2 \geq\Omega\Big(\frac{\Delta_t^2}{\beta_{t-1}(\gamma)\log(T/\Delta_t^2)}\Big)\right\}\mathbb{1}\{2^{-l}<\Delta_t\leq 2^{-l+1}\} \\
&\leq O(1)+T\Gamma+\mathbb{E}\sum_{t=1}^{T}\sum_{l=1}^{\lceil Q\rceil}2^{-l+1}\cdot\mathbb{1}\left\{\|X_t\|_{V_{t-1}^{-1}}^2 \geq \Omega\Big(\frac{2^{-2l}}{\beta_{t-1}(\gamma)\log(T\cdot 2^{2l})}\Big)\right\} \\
&\leq O(1)+T\Gamma+\mathbb{E}\sum_{l=1}^{\lceil Q\rceil}2^{-l+1}O\left(2^{2l}d\beta_T(\gamma)\log(2^{2l}T)\log\left(1+\frac{2L^2\cdot 2^{2l}\beta_T(\gamma)\log(T\cdot 2^{2l})}{\lambda}\right)\right) &(6)\\
&\leq O(1)+T\Gamma+\sum_{l=1}^{\lceil Q\rceil}2^{l+1}\cdot O\left(d\beta_T(\gamma)\log(\frac{T}{\Gamma^2})\log\left(1+\frac{L^2\beta_T(\gamma)\log(\frac{T}{\Gamma^2})}{\lambda\Gamma^2}\right)\right) \\
&\leq O(1)+T\Gamma+O\left(\frac{d\beta_T(\gamma)\log(\frac{T}{\Gamma^2})}{\Gamma}\log\left(1+\frac{L^2\beta_T(\gamma)\log(\frac{T}{\Gamma^2})}{\lambda\Gamma^2}\right)\right) \ , &(7)
\end{aligned}$$

where in Inequality (6) we used Corollary 1. Substituting the choices of Γ and γ in (5) into (7) yields the upper bound on $\mathbb{E}\sum_{t=1}^{T}\Delta_t\cdot\mathbb{1}\{B_t, C_t, D_t\}\cdot\mathbb{1}\big\{\|X_t\|_{V_{t-1}^{-1}}^2 < \frac{\Delta_t^2}{\beta_{t-1}(\gamma)}\big\}$ of the order $O(d\sqrt{T}\log^{\frac{3}{2}} T)$. Hence $F_1 \leq O(d\sqrt{T}\log^{\frac{3}{2}} T)$. Other details are fleshed out in Appendix A.2.

For F_2, since C_t and $\overline{D}_t$ together imply that $\langle \theta^*, x_t^* \rangle - \delta < \varepsilon + \langle \hat{\theta}_{t-1}, X_t \rangle$, then using the choices of δ and ε, we have $\langle \hat{\theta}_{t-1} - \theta^*, X_t \rangle > \frac{\Delta_t}{\sqrt{\log T}}$. Substituting this into the event B_t and using the Cauchy–Schwarz inequality, we have

$$\|X_t\|^2_{V_{t-1}^{-1}} \geq \frac{\Delta_t^2}{\beta_{t-1}(\gamma) \log T}.$$

Again applying the "peeling device" on Δ_t and Corollary 1, we can upper bound F_2 as follows:

$$F_2 \leq T\Gamma + O\left(\frac{d\beta_T(\gamma) \log T}{\Gamma} \right) \log \left(1 + \frac{L^2 \beta_T(\gamma) \log T}{\lambda \Gamma^2} \right) . \tag{8}$$

Then with the choice of Γ and γ as stated in (5), the upper bound of the F_2 is also of order $O(d\sqrt{T} \log^{\frac{3}{2}} T)$. More details of the calculation leading to Eqn. (8) are in Appendix A.3.

For F_3, this is the case when the best arm at time t does not perform sufficiently well so that the empirically largest reward at time t is far from the highest expected reward. One observes that minimizing F_3 results in a tradeoff with respect to F_1. On the event $\overline{C}_t$, we can again apply the "peeling device" on $\langle \theta^*, x_t^* \rangle - \langle \hat{\theta}_{t-1}, x_t^* \rangle$ such that $\frac{q+1}{2}\delta \leq \langle \theta^*, x_t^* \rangle - \langle \hat{\theta}_{t-1}, x_t^* \rangle < \frac{q+2}{2}\delta$ where $q \in \mathbb{N}$. Then using the fact that $I_{t,A_t} \leq I_{t,a_t^*}$, we have

$$\log \frac{1}{\beta_{t-1}(\gamma)\|X_t\|^2_{V_{t-1}^{-1}}} < \frac{q^2\delta^2}{4\beta_{t-1}(\gamma)\|x_t^*\|^2_{V_{t-1}^{-1}}} + \log \frac{1}{\beta_{t-1}(\gamma)\|x_t^*\|^2_{V_{t-1}^{-1}}} . \tag{9}$$

On the other hand, using the event B_t and the Cauchy–Schwarz inequality, it holds that

$$\|x_t^*\|_{V_{t-1}^{-1}} \geq \frac{(q+1)\delta}{2\sqrt{\beta_{t-1}(\gamma)}} . \tag{10}$$

If $\|X_t\|^2_{V_{t-1}^{-1}} \geq \frac{\Delta_t^2}{\beta_{t-1}(\gamma)}$, the regret in this case is bounded by $O(d\sqrt{T \log T})$. Otherwise, combining Eqn. (9) and Eqn. (10) implies that

$$\|X_t\|^2_{V_{t-1}^{-1}} \geq \frac{(q+1)^2\delta^2}{4\beta_{t-1}(\gamma)} \exp\left(-\frac{q^2}{(q+1)^2} \right) .$$

Using Corollary 1, we can now conclude that F_3 is upper bounded as

$$F_3 \leq T\Gamma + O\left(\frac{d\beta_T(\gamma) \log T}{\Gamma} \right) \log \left(1 + \frac{L^2 \beta_T(\gamma) \log T}{\lambda \Gamma^2} \right) . \tag{11}$$

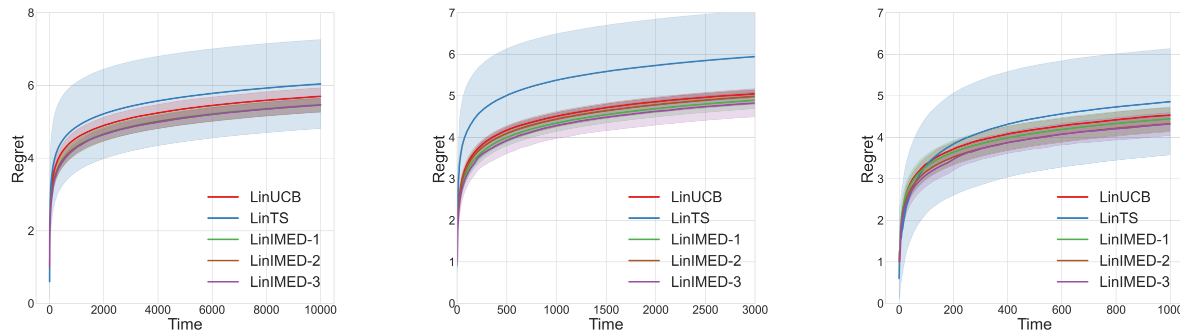
Substituting Γ and γ in Eqn. (5) into Eqn. (11), we can upper bound F_3 by $O(d\sqrt{T} \log^{\frac{3}{2}} T)$. Complete details are provided in Appendix A.4.

For F_4, using Lemma 1 with the choice of $\gamma = 1/t^2$ and $Q = -\log \Gamma$, we have

$$\begin{aligned} F_4 = \mathbb{E} \sum_{t=1}^{T} \Delta_t \cdot \mathbb{1}\left\{ \overline{B}_t \right\} &\leq T\Gamma + \mathbb{E} \sum_{t=1}^{T} \sum_{l=1}^{\lceil Q \rceil} \Delta_t \cdot \mathbb{1}\left\{ 2^{-l} < \Delta_t \leq 2^{-l+1} \right\} \mathbb{1}\left\{ \overline{B}_t \right\} \\ &\leq T\Gamma + \sum_{t=1}^{T} \sum_{l=1}^{\lceil Q \rceil} 2^{-l+1} \mathbb{P}\big(\overline{B}_t\big) \leq T\Gamma + \sum_{t=1}^{T} \sum_{l=1}^{\lceil Q \rceil} 2^{-l+1} \gamma < T\Gamma + \frac{\pi^2}{3} . \end{aligned}$$

Thus, $F_4 \leq O(d\sqrt{T} \log^{\frac{3}{2}} T)$. In conclusion, with the choice of Γ and γ in Eqn. (5), we have shown that the expected regret of LinIMED-1 $R_T = \sum_{i=1}^{4} F_i$ is upper bounded by $O(d\sqrt{T} \log^{\frac{3}{2}} T)$. $\square$

For LinIMED-2, the proof is similar but the assumption that $\langle \theta^*, x_{t,a} \rangle \geq 0$ is not required. For LinIMED-3, by directly using the UCB in Line 8 of Algorithm 1, we improve the regret bound to match the state-of-the-art $O(d\sqrt{T} \log T)$, which matches that of LinUCB with OFUL's confidence bound.



(a) $K = 10$ (b) $K = 100$ (c) $K = 1000$

Figure 1: Simulation results (expected regrets) on the synthetic dataset with different K's

6 Empirical Studies

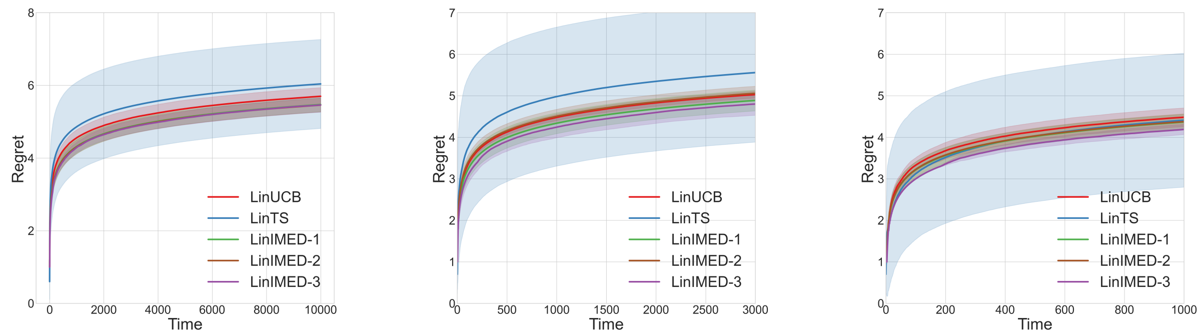
This section aims to justify the utility of the family of LinIMED algorithms we developed and to demonstrate their effectiveness through quantitative evaluations in simulated environments and real-world datasets such as the MovieLens dataset.

We compare our LinIMED algorithms with LinTS and LinUCB with the choice $\lambda = L^2$. We set $\beta_t(\gamma) = (R\sqrt{3d\log(1+t)} + \sqrt{2})^2$ (here $\gamma = \frac{1}{(1+t)^2}$ and $L = \sqrt{2}$) for the synthetic dataset with varying and finite arm set and $\beta_t(\gamma) = (R\sqrt{d\log((1+t)t^2)} + \sqrt{20})^2$ (here $\gamma = \frac{1}{t^2}$ and $L = \sqrt{20}$) for the MovieLens dataset respectively. The confidence widths $\sqrt{\beta_t(\gamma)}$ for each algorithm are multiplied by a factor α and we tune α by searching over the grid $\{0.1, 0.2, \ldots, 1.0\}$ and report the **best performance** for each algorithm; see Appendix E. Both γ's are of order $O(\frac{1}{t^2})$ as suggested by our proof sketch in Eqn. (5). We set $C = 30$ in LinIMED-3 throughout. The sub-Gaussian noise level is $R = 0.1$. We choose LinUCB and LinTS as competing algorithms since they are paradigmatic examples of deterministic and randomized contextual linear bandit algorithms respectively. We omit the SupLinUCB and SupLinIMED algorithms in our comparisons even though these algorithms "based on the Sup structure" (including SupLinRel) enjoy a better $\widetilde{O}(\sqrt{dT})$ regret bound. This is because it is well known that there is a substantial performance degradation compared to established methodologies like LinUCB or Linear Thompson Sampling (as mentioned in Lattimore & Szepesvári (2020, Chapter 22)).

6.1 Experiments on a Synthetic Dataset in the Varying Arm Set Setting

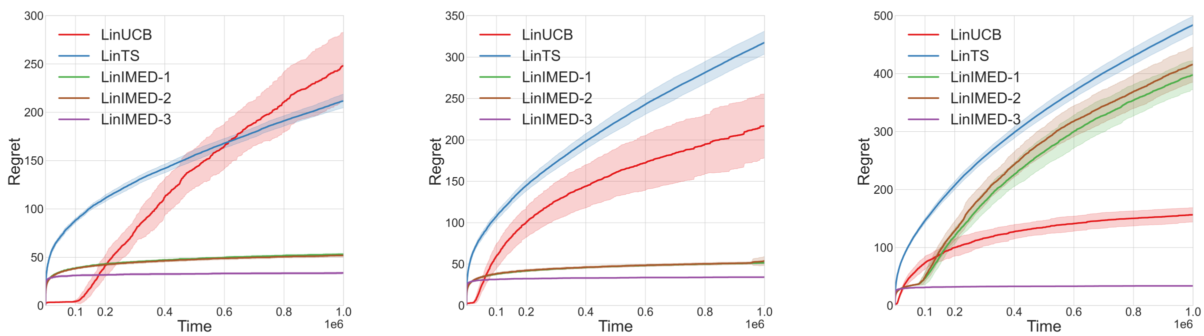
We perform an empirical study on a varying arm setting. We evaluate the performance with different dimensions d and different number of arms K. We set the unknown parameter vector and the best context vector as $\theta^* = x_t^* = [\frac{1}{\sqrt{d-1}}, \ldots, \frac{1}{\sqrt{d-1}}, 0]^\top \in \mathbb{R}^d$. There are $K-2$ suboptimal arms vectors, which are all the same (i.e., repeated) and share the context $[\frac{t}{(t+1)\sqrt{d-1}}, \ldots, \frac{t}{(t+1)\sqrt{d-1}}, \frac{t}{t+1}]^\top \in \mathbb{R}^d$. Finally, there is also one "worst" arm vector with context $[0, 0, \ldots, 0, 1]^\top$. This synthetic dataset is inspired by the synthetic varying arm set in Gales et al. (2022). First we fix $d = 20$. The results for different numbers of arms such as $K = 10, 100, 1000$ are shown in Fig. 1. Note that each plot is repeated 20 times to obtain the mean and standard deviation of the regret. From Fig. 1, we observe that LinIMED and its variants outperform LinTS and LinUCB regardless of the number of the arms K. Second, we set $K = 10$ with the dimension $d = 20, 50, 100$. Each trial is again repeated 10 times and the regret over time is shown in Fig. 2. Again, we see that LinIMED and its variants clearly perform better than LinUCB and LinTS.

The experimental results on synthetic data demonstrate that the performances of all three versions of the LinIMED algorithms are largely similar but LinIMED-3 is slightly superior (corroborating our theoretical findings). More importantly, they outperform both the LinTS and LinUCB algorithms in a statistically significant manner, regardless of the number of arms K or the dimension d of the data.



(a) $d = 20$ (b) $d = 50$ (c) $d = 100$

Figure 2: Simulation results (expected regrets) on the synthetic dataset with different d's



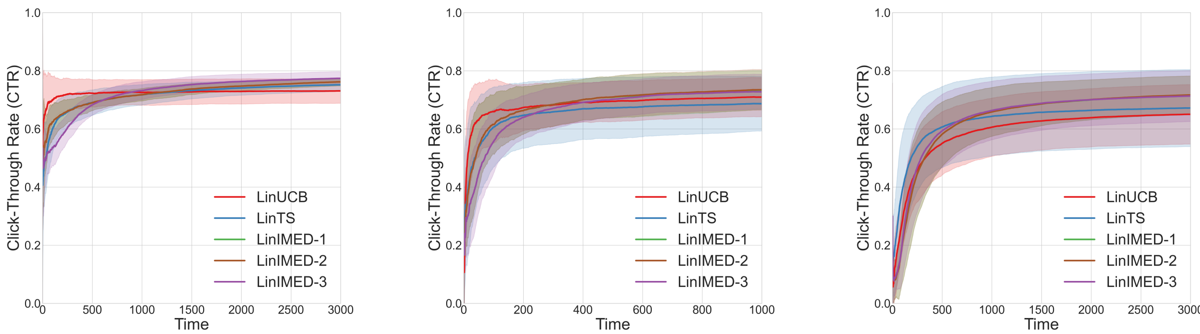
(a) $\varepsilon = 0.005$ (b) $\varepsilon = 0.01$ (c) $\varepsilon = 0.02$

Figure 3: Simulation results (expected regrets) on the "End of Optimism" instance with different ε's

6.2 Experiments on the “End of Optimism” instance

Algorithms based on the optimism principle such as LinUCB and Linear Thompson Sampling have been shown to be not asymptotically optimal. A paradigmatic example is the case known as the “End of Optimism” (Lattimore & Szepesvari, 2017; Kirschner et al., 2021)). In this two-dimensional case in which the true parameter vector $\theta^* = [1; 0]$, there are three arms represented by the arm vectors: $[1; 0], [0; 1]$ and $[1 - \varepsilon; 2\varepsilon]$, where ε is a small positive number. In this example, it is observed that even pulling a highly suboptimal arm (the second one) provides a lot of information about the best arm (the first one). We perform experiments with the same confidence parameter $\beta_t(\gamma) = (R\sqrt{3d \log(1 + t)} + \sqrt{2})^2$ as in Section 6.1 (where the noise level $R = 0.1$, dimension $d = 2$). Each algorithm is run over 10 independent trials. The regrets of all competing algorithms are shown in Fig. 3 with $\varepsilon = 0.05, 0.01, 0.02$ and for a fixed horizon $T = 10^6$.

From Fig 3 we observe our proposed LinIMED algorithms perform much better than LinUCB and Linear Thompson Sampling in this “End of Optimism” instance. In particular, LinIMED-3 performs significantly better than LinUCB and LinTS even when ε is of a moderate value such as $\varepsilon = 0.02$. We surmise that the reason behind the superior performance of our LinIMED algorithms on the "End of Optimism" instance is that the first term of our LinIMED index is $\hat{\Delta}_{t,a}^2/(\beta_{t-1}(\gamma)\|x_{t,a}\|^2_{V_{t-1}^{-1}})$, which can be viewed as an approximate and simpler version of the information ratio as mentioned as the Information-Directed Sampling (IDS) algorithm for linear bandits which attains asymptotically optimality (Kirschner et al., 2021).



(a) $K = 50$ (b) $K = 100$ (c) $K = 500$

Figure 4: Simulation results (CTRs) of the MovieLens dataset with different K's

6.3 Experiments on the MovieLens Dataset

The MovieLens dataset (Cantador et al. (2011)) is a widely-used benchmark dataset for research in recommendation systems. We specifically use the MovieLens 10M dataset, which contains 10 million ratings (from 0 to 5) and 100,000 tag applications applied to 10,000 movies by 72,000 users. To preprocess the dataset, we choose the best $K \in \{50, 100, 500\}$ movies for consideration. At each iteration t, one random user visits the website and is recommended one of the best K movies. We assume that the user will click on the recommended movie if the user's rating of this movie is at least 3; otherwise, the user will not click. We implement the three versions of LinIMED, LinUCB, and LinTS on this dataset. Each trial is repeated over 50 runs and the averages and standard deviations of the click-through rates (CTRs) as functions of time are reported in Fig. 4. One observes that on the MovieLens dataset, the variants of LinIMED also perform similarly. Furthermore, they significantly outperform LinUCB and LinTS for all $K \in \{50, 100, 500\}$ when time horizon T is sufficiently large.

7 Future Work

In the future, a fruitful direction of research is to further modify the LinIMED algorithm to make it also asymptotically optimal; we believe that in this case, the analysis would be more challenging, but the theoretical and empirical performances might be superior to our three LinIMED algorithms. In addition, one can generalize the family of IMED-style algorithms to generalized linear bandits or neural contextual bandits.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24, 2011.

Shipra Agrawal and Navin Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pp. 39.1–39.26. JMLR Workshop and Conference Proceedings, 2012.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pp. 127–135. PMLR, 2013.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

Dorian Baudry, Kazuya Suzuki, and Junya Honda. A general recipe for the analysis of randomized multi-armed bandit algorithms. *arXiv preprint arXiv:2303.06058*, 2023.

Jie Bian and Kwang-Sung Jun. Maillard sampling: Boltzmann exploration done optimally. *arXiv preprint arXiv:2111.03290*, 2021.

Iván Cantador, Peter Brusilovsky, and Tsvi Kuflik. Second workshop on information heterogeneity and fusion in recommender systems (HetRec2011). In *Proceedings of the Fifth ACM Conference on Recommender Systems*, pp. 387–388, 2011.

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214. JMLR Workshop and Conference Proceedings, 2011.

Spencer B Gales, Sunder Sethuraman, and Kwang-Sung Jun. Norm-agnostic linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 73–91. PMLR, 2022.

Aurélien Garivier and Olivier Cappé. The KL-UCB algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 359–376. JMLR Workshop and Conference Proceedings, 2011.

Junya Honda and Akimichi Takemura. Non-asymptotic analysis of a new bandit algorithm for semi-bounded rewards. *J. Mach. Learn. Res.*, 16:3721–3756, 2015.

Johannes Kirschner, Tor Lattimore, Claire Vernade, and Csaba Szepesvári. Asymptotically optimal information-directed sampling. In *Conference on Learning Theory*, pp. 2777–2821. PMLR, 2021.

Tor Lattimore and Csaba Szepesvari. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pp. 728–737. PMLR, 2017.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pp. 661–670, 2010.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933. ISSN 00063444. URL `http://www.jstor.org/stable/2332286`.

Michal Valko, Rémi Munos, Branislav Kveton, and Tomáš Kocák. Spectral bandits for smooth graph functions. In *International Conference on Machine Learning*, pp. 46–54. PMLR, 2014.