Minimum Empirical Divergence for Sub-Gaussian Linear Bandits

Kapilan Balagopalan, Kwang-Sung Jun

kapilanbgp@arizona.edu, kjun@cs.arizona.edu

THE UNIVERSITY OF ARIZONA

Preliminary

Protocol: For $t \le n$:

- Choose an arm $A_t \in \mathcal{A}_t \subset \mathbb{R}^d$ and receive the reward Y_t.

Model:

- Reward $Y_t = \langle \theta^*, A_t \rangle + \eta_t$, where $\theta^* \in \mathbb{R}^d$ is an unknown.
- Noise: η_t is σ_*^2-sub-Gaussian.

Goal: Minimize cumulative regret,

- $\mathrm{Reg}_n := \sum_{t=1}^n \langle a_t^*, \theta^* \rangle - \langle A_t, \theta^* \rangle$ where $a_t^* := \max_{a \in \mathcal{A}_t} \langle a, \theta^* \rangle$.

Assumption 1: For all $t \ge 1$, every arm $a \in \mathcal{A}_t$ satisfies $\|a\|_2 \le 1$ and for some constant B, $\Delta_{a,t} := \langle \theta^*, a_t^* \rangle - \langle \theta^*, a \rangle \le B$. Furthermore $\|\theta^*\|_2 = S_*$.

Main results

Theorem 1 (Instance-dependent bound). *Under Assumption 1, with $\delta_t = \frac{1}{t+1}$, LinMED satisfies, $\forall n \ge 1$,*

$$\mathbb{E}\,\mathrm{Reg}_n = \left(\frac{1}{\Delta} d \log(n) \left(\left(\sigma^2 d \log(n) + \lambda S^2\right) \log(\log n) + \left(\sigma_*^2 d \log(n) + \lambda S_*^2\right) H_{\max}\right)\right).$$

Our algorithm achieves an instance dependent bound of $\hat{O}\left(\frac{1}{\Delta} d^2 \left(\log^2 n\right)\right)$. (Symbol $\hat{O}$ ignores $\log(\log(n))$ factor)

Theorem 2 (Minimax Bound). *Under Assumption 1, with $\delta_t = \frac{1}{t+1}$, LinMED satisfies, $\forall n \ge 1$,*

$$\mathbb{E}\,\mathrm{Reg}_n = \left(\sqrt{n}\left(\log^{\frac{1}{2}}(n)\left(d\sigma \log(n) + \frac{\lambda S^2}{\sigma}\right) + \frac{H_{\max}}{\sigma \log^{\frac{3}{2}}(n)}\left(d\sigma_*^2 \log(n) + \lambda S_*^2\right)\right)\right).$$

S and σ^2 are the guesses for S_* and σ_*^2 respectively. Our algorithm achieves a near-optimal minimax bound ($\tilde{O}(d\sqrt{n})$) and a state-of-the-art instance dependent bound ($\frac{1}{\Delta} d^2 \log^2 n$), even when S_* and σ_*^2 are misspecified. (Many state-of-the-art algorithms including OFUL lacks an analysis when they are underspecified)

Comparison

Algorithms	Minimax regret	Instance dependent regret	Closed form probability	Probability assigned for all arms
OFUL(Abbasi-Yadkori et al., 2011)	$\tilde{O}(d\sqrt{n})$	$O(\frac{d^2}{\Delta}\log^2 n)$	N/A	✗
LinIMED(Bian and Tan V.F, 2024)	$\tilde{O}(d\sqrt{n})$	Unknown	N/A	✗
LinTS(Agrawal and Goyal, 2014)	$\tilde{O}(d^{\frac{3}{2}}\sqrt{n})$	Unknown	✗	✗
RandUCB(Vaswanit et al., 2020)	$\tilde{O}(d\sqrt{n})$	Unknown	✗	✗
SquareCB(Foster and Rakhlin, 2020)	$\tilde{O}(\sqrt{K d n})$	Unknown	✓	✗
E2D(Foster et al., 2023)	$\tilde{O}(d\sqrt{n})$	Unknown	✓	✗*
SpannerIGW(Zhu et al., 2022)	$\tilde{O}(d\sqrt{n})$	$\Omega(\Delta\sqrt{n})$	✓	✗*
EXP2(Bubeck and Cesa-Bianchi, 2012)	$O(\sqrt{dn\log K})$	$\Omega(\Delta\sqrt{n})$	✓	✓
LinMED(ours)	$\tilde{O}(d\sqrt{n})$	$\hat{O}(\frac{d^2}{\Delta}\log^2 n)$	✓	✓

Symbol '✗*' means that the algorithm can be modified to assign probability to all arms. Symbol $\tilde{O}$ ignores the $\log(n)$ factor. Symbol $\hat{O}$ ignores the $\log(\log(n))$ factor.

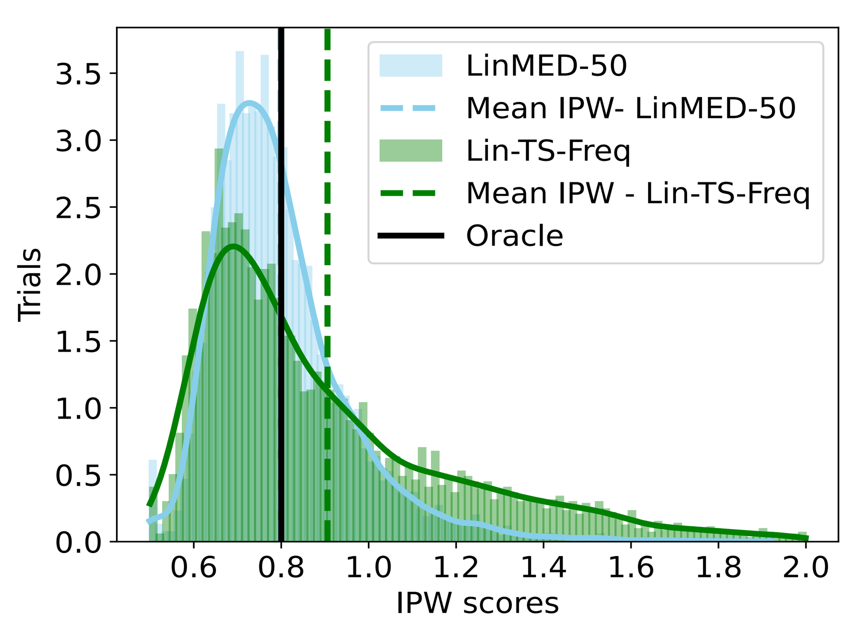
Supplementary result

Theorem 3. *There exists a linear bandit problem for which the EXP2 and SpannerIGW algorithms satisfy*

$$\mathbb{E}\,\mathrm{Reg}_n \ge \Omega(\Delta\sqrt{n}).$$

EXP2 (Bubeck and Cesa-Bianchi, 2012) and SpannerIGW (Zhu et al., 2022) have polynomial instance dependent lower bound. LinMED achieves polylog instance dependent upper bound.

OPE-friendliness



$\mathcal{A} = \{a_1 = (1,0)^\top, a_2 = (0.6, 0.8)^\top\}$, $\theta^* = (1,0)^\top$, LinTS's mean = 0.906, Oracle's mean ≈ LinMED's mean = 0.800

$$\text{IPW score} = \frac{1}{n}\sum_{t=1}^{n} \frac{\pi_t^{\text{target}}(A_t)}{p_t(A_t)} \cdot Y_t.$$

$$(\pi_t^{\text{target}}(A_t) = \tfrac{1}{|\mathcal{A}|})$$

IPW scores (expected regret equivalent) of the uniform policy when the logging policy is LinMED and LinTS respectively. We used 10^3 Monte Carlo samples to estimate the sampling probabilities of LinTS. Oracle denotes the expected reward of the uniform policy. LinTS shows a nontrivial amount of bias while LinMED is exactly aligned with the oracle.

Highlights and contributions

- LinMED: Linear extension for MED algorithm. (Bian and Jun, 2022, Honda and Takemura, 2011)
- Near-optimal minimax bound and logarithmic instance dependent bound even with noise misspecification.
- Polynomial instance dependent lower bounds for SpannerIGW (Zhu et al., 2022) and EXP2 (Bubeck and Cesa-Bianchi, 2012) ($\Omega(\Delta\sqrt{n})$), which are strictly worse than LinMED.
- Offline policy evaluation friendly algorithm: Our algorithm assigns a closed from probability for each arm, hence it can be used as a logging policy for offline performance evaluation of other policies.

MED vs LinMED

$$p_{t,a} \propto \exp\left(-\frac{N_{t-1,a}}{2} \cdot \left(\max_{a' \in \mathcal{A}} \hat{\mu}_{t-1,a'} - \hat{\mu}_{t-1,a}\right)^2\right). \qquad \text{(MED)}$$

Exploration	Empirical Divergence(ED)	Effective arm set size
$\underbrace{N_{t-1,a}}_{\#\text{ arm pull}} \Leftrightarrow \frac{1}{\|a\|^2_{V_{t-1}^{-1}}}$	$\underbrace{(\max_{a' \in \mathcal{A}} \hat{\mu}_{t-1,a'} - \hat{\mu}_{t-1,a})^2}_{\text{Estimated mean gap}} \Leftrightarrow \hat{\Delta}^2_{a,t}$	$\underbrace{K}_{\lvert\mathcal{A}\rvert} \Leftrightarrow \tilde{O}(d)$

LinMED algorithm

Algorithm 1 LinMED

Input: Regularization λ, failure rates $\{\delta_t\}_{t=0}^{\infty}$, optimal design fraction α_{opt}, empirical best fraction α_{emp}, S (guess for $\|\theta^*\|_2$), and σ^2 (guess for σ_*^2)

```
Initialize θ̂_0 = 0, V_0 = λI.
for t = 1, 2, ... do
    Observe arm set A_t.
    Estimate â_t = max_{a' ∈ A_t} ⟨θ̂_{t-1}, a'⟩.
    Estimate Δ̂_{a,t} := ⟨θ̂_{t-1}, â_t − a⟩   ∀a ∈ A_t.
    Define ∀a ∈ Ā_t
```

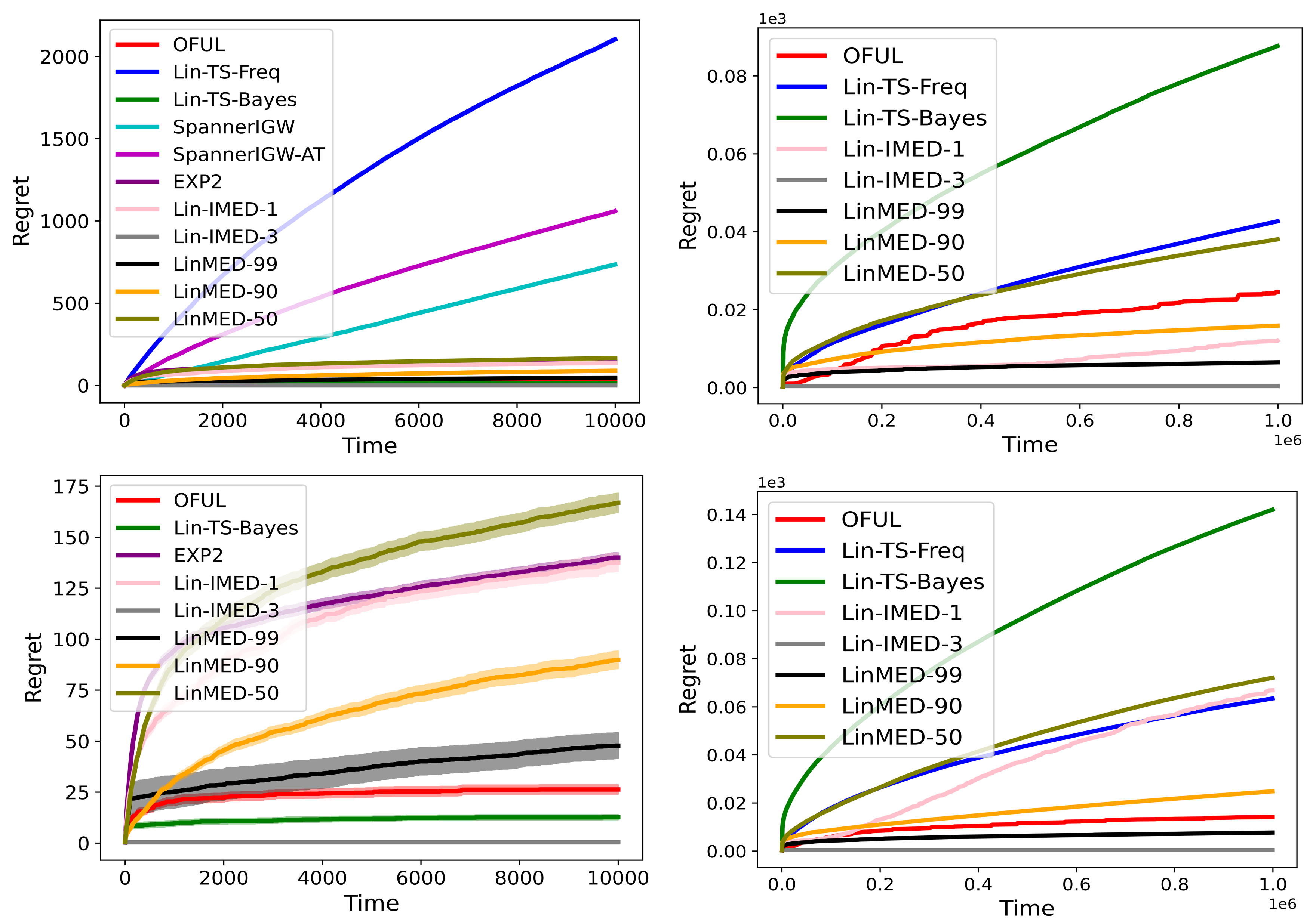
$$f_t(a) = \exp\left(-\frac{\hat{\Delta}^2_{a,t}}{\beta_{t-1}(\delta_{t-1})\|\hat{a}_t - a\|^2_{V_{t-1}^{-1}}}\right)$$

where we take $\frac{0}{0} = 0$ and $\beta_t(\delta_t) := \left(\sigma\sqrt{\log\left(\frac{\det V_t}{\det V_0}\right) + 2\log\frac{1}{\delta_t}} + \sqrt{\lambda}S\right)^2$

- Re-scale the arms: $\overline{\mathcal{A}}_{(t)} = \{\sqrt{f_t(a)} \cdot a \mid a \in \mathcal{A}_t\}$.
- Compute $q_t^{\text{opt}} = \text{ApproxDesign}(\bar{\mathcal{A}}_t)$ such that $\|b\|^2_{V^{-1}(q_t^{\text{opt}})} \le \tilde{O}(d), \forall b \in \bar{\mathcal{A}}_t$.
- Let $\forall a \in \mathcal{A}_t \quad q_t(a) = \alpha_{\text{opt}} \cdot q_t^{\text{opt}}(a) + \alpha_{\text{emp}} \cdot \mathbf{1}\{a = \hat{a}_t\} + (1 - \alpha_{\text{opt}} - \alpha_{\text{emp}}) \cdot \frac{1}{|\mathcal{A}_t|}$.
- Compute $p'_t(a) = \frac{q_t(a) f_t(a)}{\sum_{b \in \mathcal{A}_t} q_t(b) f_t(b)}$.
- Define $\mathcal{B}_t = \{a \in \mathcal{A}_t : \|a\|^2_{V_{t-1}^{-1}} > 1\}$.
- **if** $|\mathcal{B}_t| > 0$ **then**
  - $\forall a \in \mathcal{A}_t, \quad p_t(a) = \frac{1}{2}p'_t(a) + \frac{1}{2}\mathbf{1}\{a = B_t\}$ where $B_t \in \mathcal{B}_t$.
- **else**
  - $\forall a \in \mathcal{A}_t \quad p_t(a) = p'_t(a)$.
- **end if**
- Take action $A_t \sim p_t$.
- Observe the reward Y_t and update $V_t = V_{t-1} + A_t A_t^\top$ and $\hat{\theta}_t = V_t^{-1}\sum_{s=1}^{t} A_s Y_s$.

end for

Experiments



Large gap instance

Model 1 ⇐ $\theta^* = (1,0)$, $\eta_t \sim \mathcal{N}(0, \sigma^2 = 1)$

$\mathcal{A} = \{(1,0), (0,1)\}$

End of optimism instance

Model 1, $\epsilon \in \{0.005, 0.01\}$

$\mathcal{A} = \{a_1 = (1,0), a_2 = (0,1), a_3 = (1-\epsilon, 2\epsilon)\}$

LinMED shows logarithmic growth for Large gap instance. Optimistic algorithms like OFUL and Thompson sampling fail under End of optimism experiments (Lattimore and Szepesvári, 2017).