Improved Regret Bounds of Bilinear Bandits using Action Space Analysis

Kyoungseok Jang

Dept. of Mathematical Science, KAIST

Kwang-sung Jun

Dept. of Computation, University of Arizona

Seyoung Yun

Graduate School of AI, KAIST

Wanmo Kang

Dept. of Mathematical Science, KAIST

ICML 2021

Bilinear bandit

$$r_t = \boxed{x_t^\top} \times \underset{\text{(Unknown)}}{\boxed{\Theta^*}} \times \boxed{z_t} + \underset{\text{(Noise)}}{\eta_t}$$

- For $t = 1, \cdots, T$
  - the agent selects $x_t \in \mathbb{R}^{d_1}, z_t \in \mathbb{R}^{d_2}$
  - Receives reward r_t as a noisy bilinear function.
- Objective: minimize the pseudo-regret $R_T = \sum_{t=1}^{T} \left[(x^*)^\top \Theta^* z^* - x_t^\top \Theta^* z_t\right]$

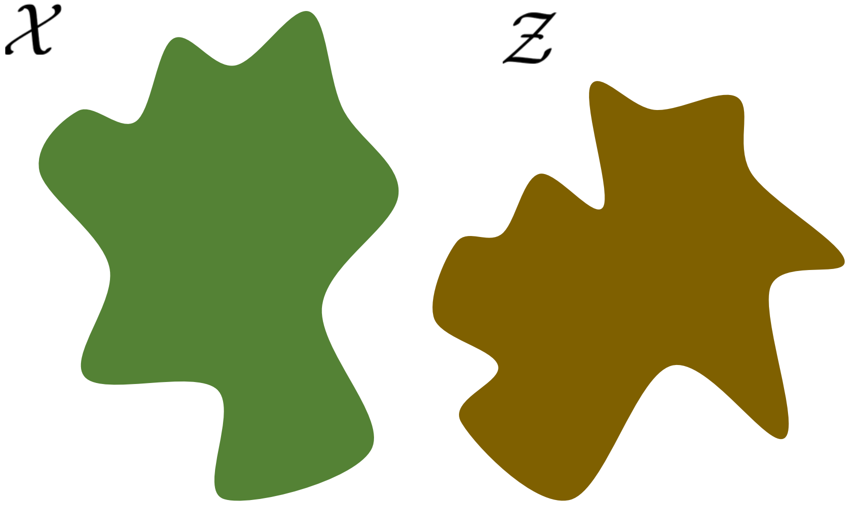
Our Contribution

- We reject the conjectured lower bound $\Omega(\sqrt{rd^3T})$
  - By proposing a new algorithm ϵ- FALB with upper bound $\tilde{O}(\sqrt{d^3T})$
  - Leverages the low-dimensional property of the action space

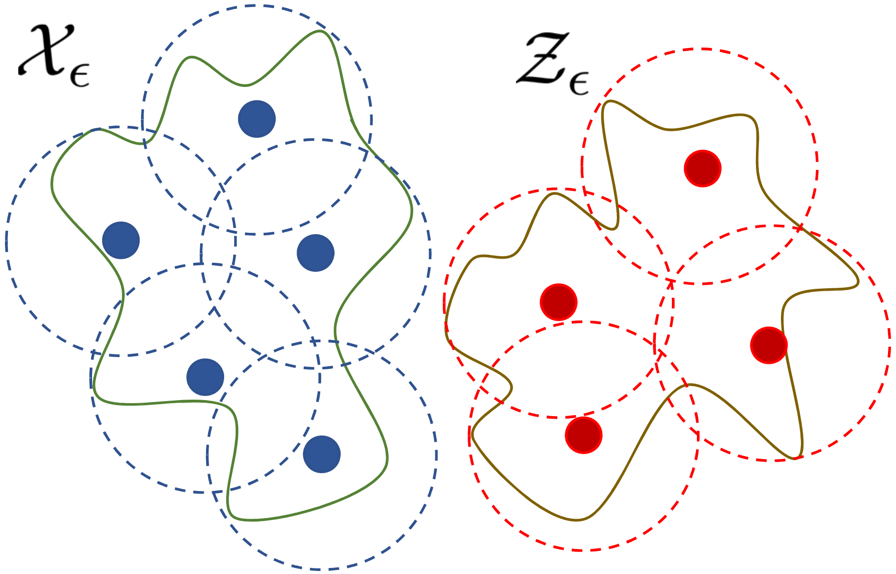
Result	**Regret UB**
LinUCB (Abbasi et al. (2011))	$\tilde{O}(d^2\sqrt{T})$
ESTR (Jun et al. (2019))	$\tilde{O}(\sqrt{rd^3T})$
LowGLOC (Lu et al. (2020))	$\tilde{O}(\sqrt{rd^3T})$
ϵ-FALB (Ours)	$\tilde{O}(\sqrt{d^3T})$

- We additionally proposed a practical algorithm – rO-UCB.

Algorithm: ϵ-FALB

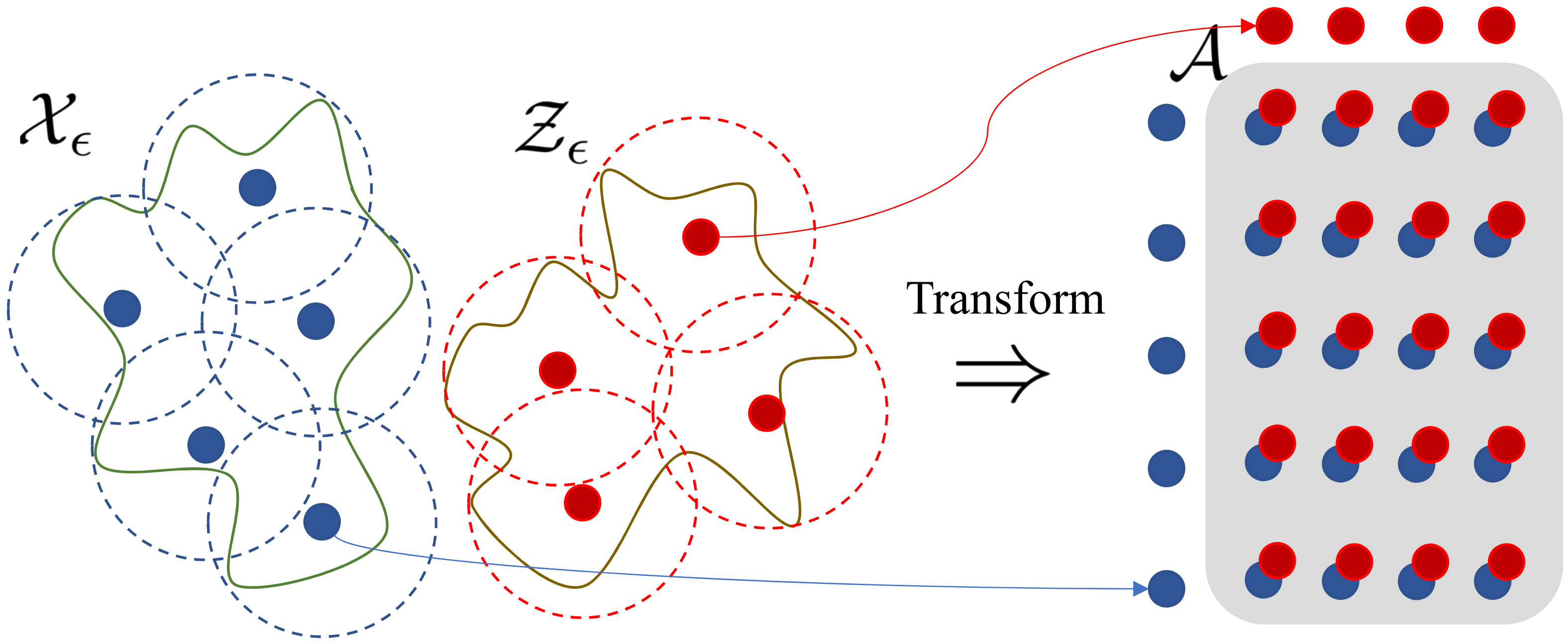


Algorithm: ϵ-FALB



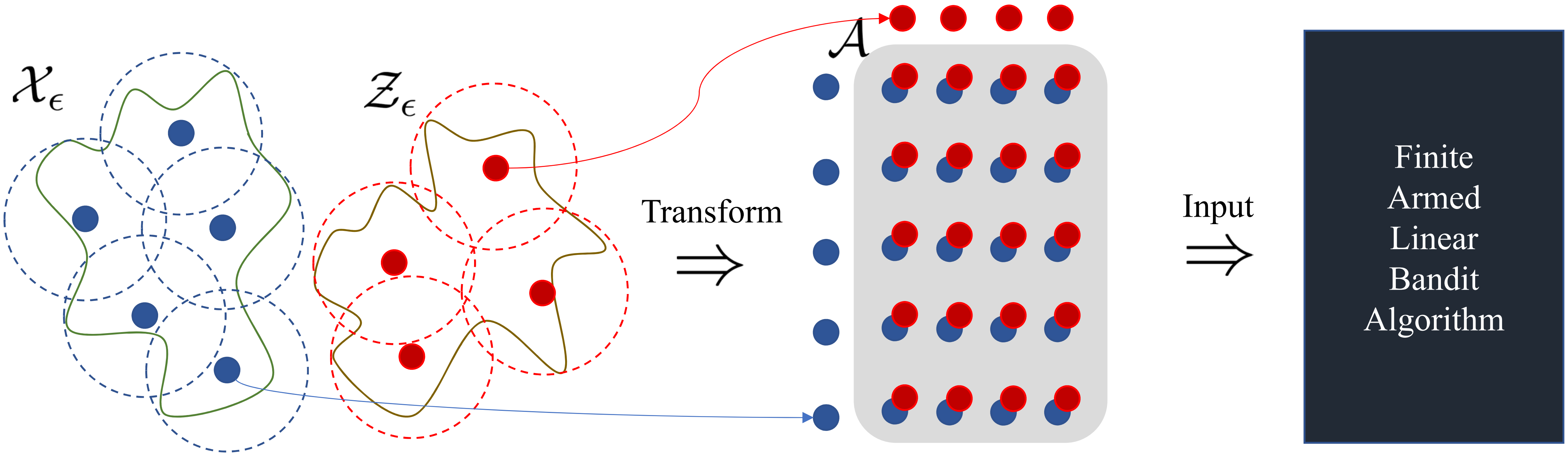
- Discretize the each side of the (possibly **infinite**) action space

Algorithm: ϵ-FALB



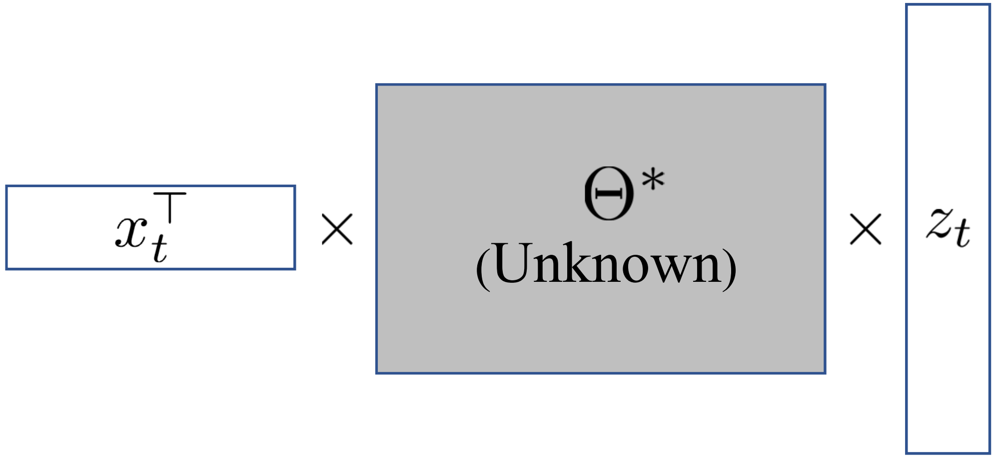
- Discretize the each side of the (possibly **infinite**) action space
- Set a new action space $\mathcal{A} = \{vec(xz^\top) : x \in \mathcal{X}_\epsilon, z \in \mathcal{Z}_\epsilon\}$

Algorithm: ϵ-FALB



- Discretize the each side of the (possibly **infinite**) action space
- Set a new action space $\mathcal{A} = \{vec(xz^\top) : x \in \mathcal{X}_\epsilon, z \in \mathcal{Z}_\epsilon\}$
- Apply $d_1 d_2$-dimensional **finite** armed linear bandit algorithms

Motivation



- Bilinear bandit

Motivation

$$x_t^\top \times \underset{\text{(Unknown)}}{\Theta^*} \times z_t = \langle \underset{\text{(Unknown)}}{vec(\Theta^*)} , vec(x_t z_t^\top) \rangle$$

- Bilinear bandit → $d_1 d_2$-dimensional linear bandit

Motivation

$$x_t^\top \times \underset{\text{(Unknown)}}{\Theta^*} \times z_t = \left\langle \underset{\text{(Unknown)}}{vec(\Theta^*)}, vec(x_t z_t^\top) \right\rangle$$

- Bilinear bandit → $d_1 d_2$-dimensional linear bandit
- Previous studies mainly focused on low-rank conditions of Θ^*

Motivation

$$x_t^\top \times \underset{\text{(Unknown)}}{\Theta^*} \times z_t = \left\langle \underset{\text{(Unknown)}}{vec(\Theta^*)} , vec(x_t z_t^\top) \right\rangle$$

- Bilinear bandit → $d_1 d_2$-dimensional linear bandit
- Previous studies mainly focused on low-rank conditions of Θ^*
- In the above perspective, the actions are always rank-1 matrices.

Motivation

$$x_t^\top \times \underset{\text{(Unknown)}}{\Theta^*} \times z_t = \left\langle \underset{\text{(Unknown)}}{vec(\Theta^*)} , vec(x_t z_t^\top) \right\rangle$$

- Bilinear bandit → $d_1 d_2$-dimensional linear bandit
- Previous studies mainly focused on low-rank conditions of Θ^*
- In the above perspective, the actions are always rank-1 matrices.
- Analyzing action spaces might reduce the regret upper bound

Why this approach works

- We focus on the dimension of the action spaces
  - Rank-r matrix manifold has dimension $r(d_1 + d_2 - r)$
- Discretization is a good way to exploit this dimension
  - We prove that $\log|\mathcal{A}| \approx (d_1 + d_2)\log(1/\epsilon)$
  - Discretization error is ignorable
- $d_1 d_2$-dimension FALB algorithm regret: $\tilde{O}(\sqrt{d_1 d_2 T \log K})$
- Which leads in total: $\tilde{O}(\sqrt{d_1 d_2 (d_1 + d_2) T})$

Additional algorithm – rO-UCB

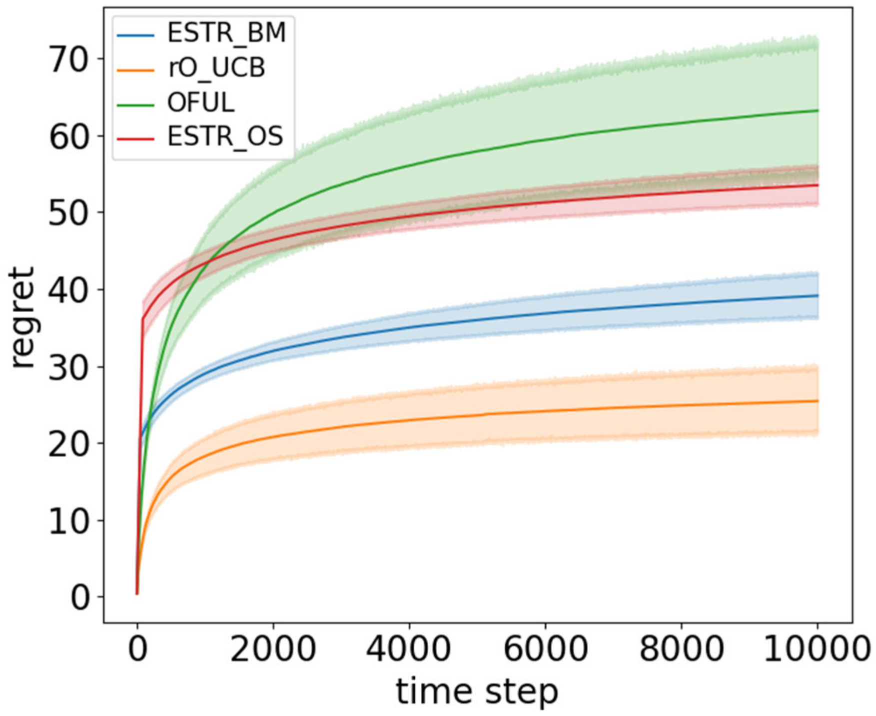
- Discretization is intractable in practice – requires $\tilde{O}(T^{d/2})$ points
- Instead, we propose a practical algorithm, rO-UCB
  - Based on the oracle about LSE with rank r constraint.

$$
(\mathsf{Opt}) \begin{cases} \min_{\Theta} & \sum_{s=1}^{t} (x_s^\top \Theta z_s - r_s)^2 \\ \text{subject to} & \mathsf{rank}(\Theta) \le r, \\ & \|\Theta\|_F \le C \end{cases}
$$

  - rO-UCB is an adapted algorithm of LinUCB, with our novel confidence set.
  - Regret upper bound: $\tilde{O}(\sqrt{rd^3T})$, but shows better empirical performance.

Experimental result

- Better performance compare to the state of the art algorithms.
- Does not depend on a force exploration phase



Thank you!

Contact e-mail: jajajang@kaist.ac.kr

Reference

- Abbasi-Yadkori et al., Improved algorithms for linear stochastic bandits. In NeurIPS 2011
- Jun et al, Bilinear bandits with low-rank structure. In ICML 2019
- Lu et al, Low-rank generalized linear bandit problem, In AISTATS 2020