Directional Optimism for Safe Linear Bandits

AISTATS 2024

Spencer Hutchinson, Berkay Turan, Mahnoosh Alizadeh

University of California, Santa Barbara

CONTRIBUTIONS

We study the *safe linear bandit problem*, a version of the stochastic linear bandit problem where the learner must satisfy uncertain constraints in every round. For this problem, we give a novel algorithm (ROFUL), a novel generalization (linked convex constraints), and improved regret guarantees for specific settings (problem-dependent and finite star-convex action sets).

Algorithm	General	Problem-Dependent	Finite star-convex	Linked convex constraints
Safe-LTS [1]	$d^{3/2}\sqrt{T}$	-	-	-
GenOP [2][3]	$d\sqrt{T}$	$\frac{d^2}{\Delta}+\sqrt{T}$	-	$d\sqrt{T}$
ROFUL	$d\sqrt{T}$	$\frac{d^2}{\Delta}+\sqrt{T}$	-	$d\sqrt{T}$
Safe-PE	-	-	$\sqrt{dT}$	$\sqrt{dT}$

Novel algorithms and regret bounds, where prior work shown in gray. Regret bounds are $\tilde{\mathcal{O}}(\cdot)$.

SAFE LINEAR BANDIT PROBLEM

Interaction Model:

At each round $t \in [T]$:

1. Play action $x_t \in \mathcal{X}$.
2. Observe reward $y_t = \theta^\top x_t + \epsilon_t$ (unknown θ).
3. Observe constraint feedback $z_t = a^\top x_t + \eta_t$ (unknown a).

Learning Goals:

- Minimize pseudo-regret:

$$R_T = \sum_t \theta^\top (x_* - x_t), \qquad x_* = \text{argmax}_{x\in\mathcal{Y}} \theta^\top x, \qquad \mathcal{Y} = \{x \in \mathcal{X} : a^\top x \le b\}$$

- Satisfy constraint in all rounds: $a^\top x_t \le b \quad \forall t \in [T]$

Assumptions:

- Action set ($\mathcal{X}$) is star-convex and bounded ($\|x\| \le 1 \ \ \forall x \in \mathcal{X}$).
- Optimal reward is positive ($\theta^\top x_* > 0$).
- Reward and constraint are bounded ($\|\theta\| \le S_\theta$, $\|a\| \le S_a$).
- Noise (ϵ_t, η_t) is subgaussian.

TECHNICAL APPROACH

*Fundamentally, this is a problem of choosing **directions**, where the uncertainty in each direction comes from both the reward and constraint functions.*

- In each direction ($u \in \mathbb{S}$), the only viable action is that with the largest scaling.
- Therefore, the challenge is to choose *directions* to efficiently balance exploration-exploitation given that the reward and maximum scaling in each direction is unknown.
- Then, the best action is the one in this direction with the largest *safe scaling*.

Our algorithms use this idea:

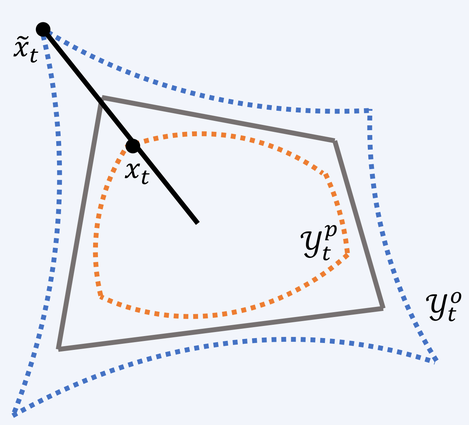
- ROFUL chooses the *optimistic direction* in each round.
- Problem-dependent analysis considers *directionally well-separated instances*.
- Safe-PE eliminates *low-reward directions* in each phase.

ROFUL ALGORITHM

Optimistic and pessimistic sets: The algorithm makes use of sets that overestimate the feasible set (optimistic set $\mathcal{Y}_t^o$) and underestimate the feasible set (pessimistic set $\mathcal{Y}_t^p$).

Optimistic direction selection: The optimistic direction is identified by finding the optimistic action ($\tilde{x}_t$) that maximizes the upper confidence bound of the reward over the optimistic set (line 4).

Safe scaling: The safe scaling of the optimistic direction is found by finding the largest scaling of the optimistic direction that is in the pessimistic set, while incorporating the assumed bound on constraint (line 6).



Algorithm 1: Restrained OFUL (ROFUL)

Input: $\mathcal{X}, \nu, b, \beta_t, \delta \in (0,1), \lambda \ge 1$

for $t = 1$ **to** T **do**

2. Update $\hat{a}_t := V_t^{-1}\sum_{k=1}^{t-1} x_k z_k$ and $\hat{\theta}_t := V_t^{-1}\sum_{k=1}^{t-1} x_k y_k$, where $V_t = \sum_{k=1}^{t-1} x_k x_k^\top + \lambda I$.
3. Update $\mathcal{Y}_t^p := \left\{x \in \mathcal{X} : \hat{a}_t^\top x + \beta_t \|x\|_{V_t^{-1}} \le b\right\}$ and $\mathcal{Y}_t^o := \left\{x \in \mathcal{X} : \hat{a}_t^\top x - \beta_t \|x\|_{V_t^{-1}} \le b\right\}$.
4. Find a $\tilde{x}_t \in \arg\max_{x\in\mathcal{Y}_t^o} \left(\hat{\theta}_t^\top x + \beta_t \|x\|_{V_t^{-1}}\right)$.
5. Set $\tilde{b}_t = \begin{cases} \min\left(\frac{\nu}{\|\tilde{x}_t\|}, 1\right) & \text{if } \tilde{x}_t \ne \mathbf{0}, \\ 1 & \text{else.} \end{cases}$
6. Set $\mu_t = \max\{\mu \in [0,1] : \mu\tilde{x}_t \in \mathcal{Y}_t^p\}$ and $\gamma_t = \max\left(\tilde{b}_t, \mu_t\right)$.
7. Play $x_t = \gamma_t \tilde{x}_t$ and observe y_t, z_t.

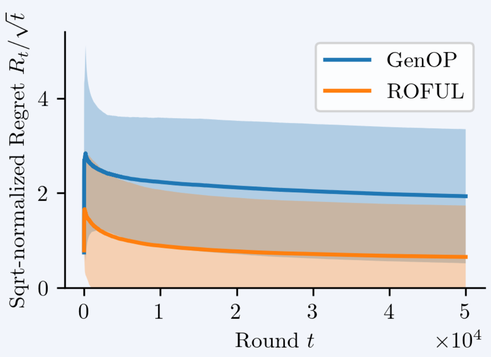
end

Theorem (General Regret Bound)**.** With probability at least $1-\delta$, ROFUL ensures that

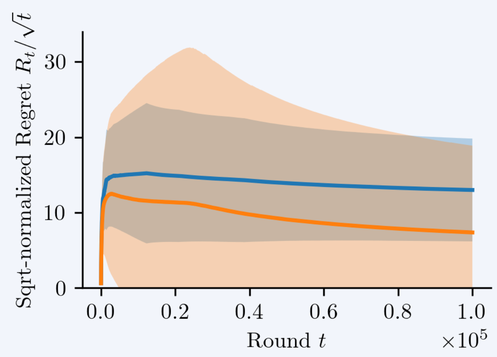
$$R_T \le 2\frac{\|\theta\| + S_a}{b}\beta_T\sqrt{2dT\log\left(1+\frac{T}{\lambda d}\right)}.$$

Comparison to existing approaches:

- Existing approaches (e.g. [1][2][3]) choose actions directly from the pessimistic set using an expanded upper confidence bound.
- Unlike ROFUL, these existing approaches rely on a fixed constant that needs to be chosen ahead of time and therefore chosen with worst-case quantities.
- ROFUL enjoys better empirical performance when constraint is less tight, i.e. b is large.



Simulation results with large b.



Simulation results with small b.

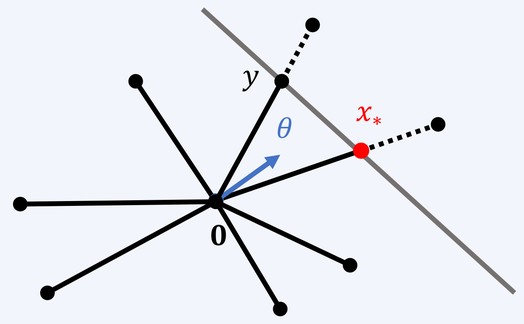
PROBLEM-DEPENDENT ANALYSIS

Directional reward gap: We consider the gap in reward between the best and second-best directions,

$$\Delta := \inf_{x\in\mathcal{Y}\,:\, x\ne\alpha x_* \ \forall\alpha>0} \theta^\top(x_* - x).$$

Theorem (Wrong Directions)**.** When $\Delta > 0$, the number of wrong directions chosen by ROFUL is $\mathcal{O}\left(\frac{1}{\Delta^2} d^2 \log^2 T\right)$.

Nearly dimension-free regret: When $\Delta > 0$ and Δ is known, it is possible to achieve $\tilde{\mathcal{O}}\left(\frac{d^2}{\Delta} + \sqrt{T}\right)$ regret by first identifying the correct direction and then playing in only that direction.

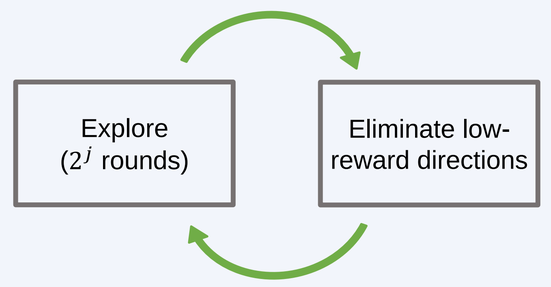


$\Delta = \theta^\top(x_* - y)$

SAFE-PE ALGORITHM

Finite directions: When the action set has finite directions (i.e. finite star-convex), then the confidence set can be constructed over each direction, reducing width of confidence set by $\sqrt{d}$.

Theorem (Regret of Safe-PE)**.** When the action set is finite star-convex, Safe-PE enjoys regret $\tilde{\mathcal{O}}\left(\sqrt{dT}\right)$.



LINKED CONVEX CONSTRAINTS

Convex constraints: We extend the problem to convex constraints with linear uncertainty, i.e.

$$Ax_t \in \mathcal{G}$$

for convex $\mathcal{G}$ and feedback $z_t = Ax_t + \eta_t$.

Analysis approach: We take a convex analysis-based approach, using the fact that

$$\alpha\mathcal{G} + (1-\alpha)\mathrm{r}\mathbb{B} \subseteq \mathcal{G},$$

when $\mathrm{r}\mathbb{B} \subseteq \mathcal{G}$.

$\mathcal{H}(x) := \hat{A}_t x + \beta_t \|x\|_{V_t^{-1}}\mathbb{B}$

$\mathcal{P}_\alpha := \alpha u + 2\beta_t \|\alpha u\|_{V_t^{-1}}\mathbb{B}$

FUTURE DIRECTIONS

Our approach might yield similar gains when applied to related safe learning problems such as,

- Constrained MDPs
- Safe Kernelized (or GP) Bandits

ACKNOWLEDGEMENTS

This work was supported by NSF award #1847096.

REFERENCES

1. Moradipari et al. (2021) Safe linear thompson sampling with side information.
2. Pacchiano et al. (2021) Stochastic bandits with linear constraints.
3. Amani et al. (2022) Decentralized multi-agent linear bandits with safety constraints.