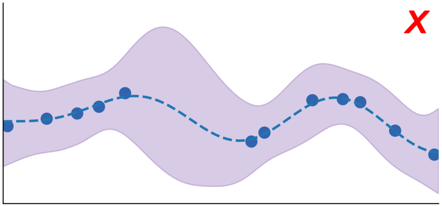
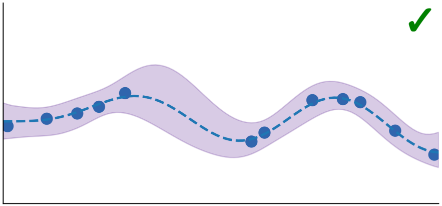
Improved Algorithms for Stochastic Linear Bandits Using Tail Bounds for Martingale Mixtures

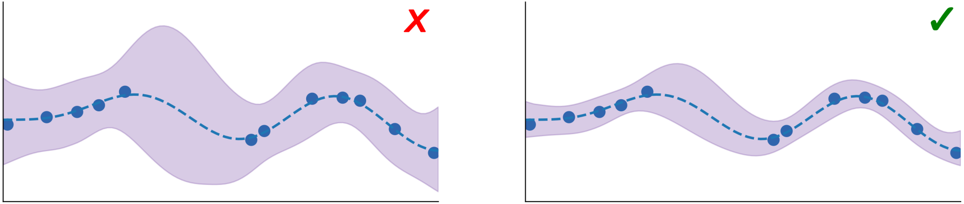
Hamish Flynn David Reeb Melih Kandemir Jan Peters

Tighter Confidence Bounds Are Better Confidence Bounds



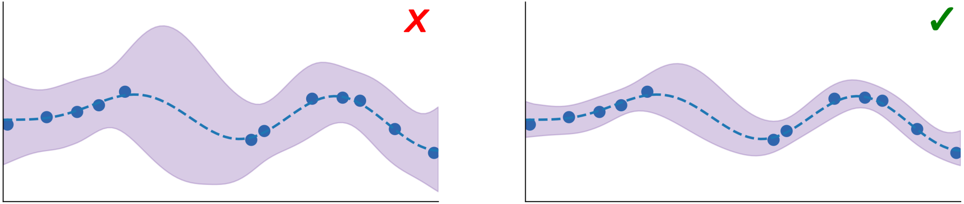


Tighter Confidence Bounds Are Better Confidence Bounds



Confidence bounds → Upper Confidence Bound (UCB) algorithms for bandits.

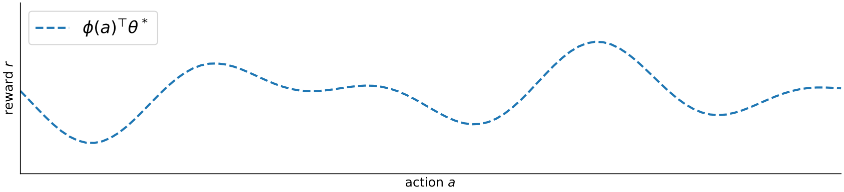
Tighter Confidence Bounds Are Better Confidence Bounds



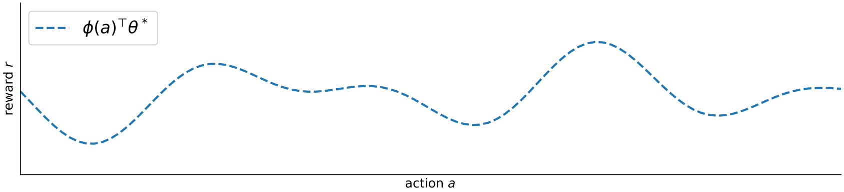
Confidence bounds $\rightarrow$ Upper Confidence Bound (UCB) algorithms for bandits.

Tighter confidence bounds $\rightarrow$ better UCB algorithms.

Stochastic Linear Bandits

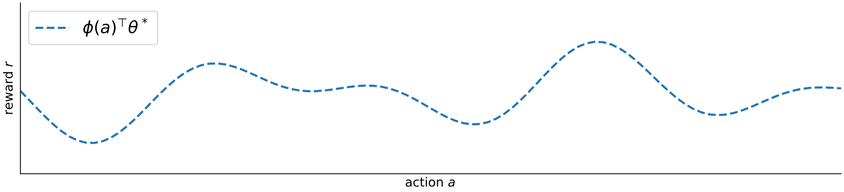


Stochastic Linear Bandits



At round t, query any action $a_t \in \mathcal{A}_t$, receive a noisy reward $r_t = \phi(a_t)^\top \boldsymbol{\theta}^* + \epsilon_t$.

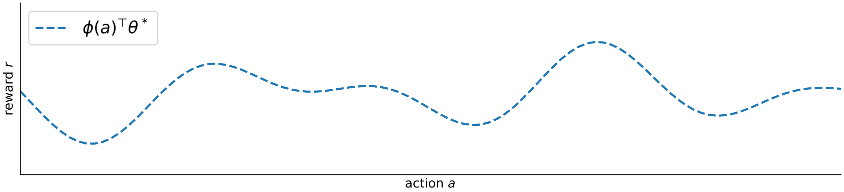
Stochastic Linear Bandits



At round t, query any action $a_t \in \mathcal{A}_t$, receive a noisy reward $r_t = \phi(a_t)^\top \boldsymbol{\theta}^* + \epsilon_t$.

Goal: Maximise total reward/minimise cumulative regret.

Stochastic Linear Bandits



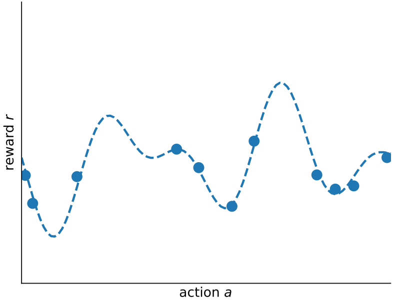
At round t, query any action $a_t \in \mathcal{A}_t$, receive a noisy reward $r_t = \phi(a_t)^\top \boldsymbol{\theta}^* + \epsilon_t$.

Goal: Maximise total reward/minimise cumulative regret.

Assumptions: $\epsilon_1, \epsilon_2, \ldots$ are (conditionally) σ-sub-Gaussian and $\|\boldsymbol{\theta}^*\|_2 \leq B$.

$\boldsymbol{\theta}^* \in \mathbb{R}^d$ is unknown, ϕ is known and upper bounds on σ and B are known.

UCB Algorithms for Stochastic Linear Bandits (e.g. OFUL[1])

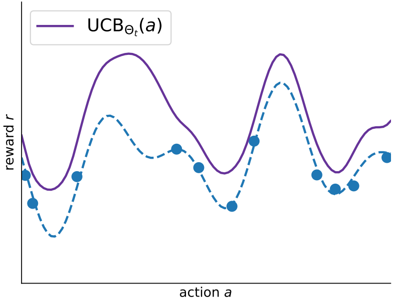


LinUCB:

For $t = 0, 1, 2, \ldots$

[1] Y. Abbasi-Yadkori et al. (2011) Improved algorithms for linear stochastic bandits. NeurIPS

UCB Algorithms for Stochastic Linear Bandits (e.g. OFUL[1])



LinUCB:

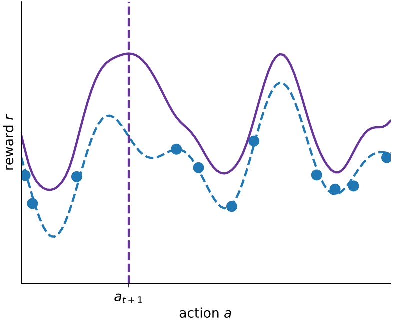
For $t = 0, 1, 2, \ldots$

- Use $\{(a_k, r_k)\}_{k=1}^{t}$ to construct a confidence set Θ_t and the corresponding upper confidence bound

$$\mathrm{UCB}_{\Theta_t}(a) := \max_{\theta \in \Theta_t} \{\phi(a)^\top \boldsymbol{\theta}\}$$

[1] Y. Abbasi-Yadkori et al. (2011) Improved algorithms for linear stochastic bandits. NeurIPS

UCB Algorithms for Stochastic Linear Bandits (e.g. OFUL[1])



LinUCB:

For $t = 0, 1, 2, \ldots$

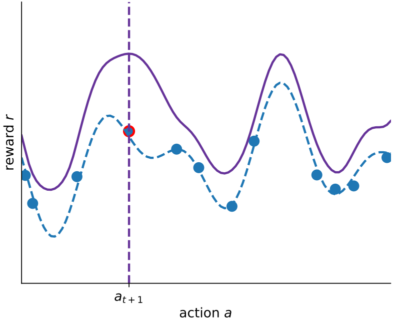
- Use $\{(a_k, r_k)\}_{k=1}^{t}$ to construct a confidence set Θ_t and the corresponding upper confidence bound

$$\mathrm{UCB}_{\Theta_t}(a) := \max_{\theta \in \Theta_t} \{\phi(a)^\top \boldsymbol{\theta}\}$$

- Play $a_{t+1} = \mathrm{argmax}_{a \in \mathcal{A}_{t+1}} \{\mathrm{UCB}_{\Theta_t}(a)\}$

[1] Y. Abbasi-Yadkori et al. (2011) Improved algorithms for linear stochastic bandits. NeurIPS

UCB Algorithms for Stochastic Linear Bandits (e.g. OFUL[1])



LinUCB:

For $t = 0, 1, 2, \ldots$

- Use $\{(a_k, r_k)\}_{k=1}^{t}$ to construct a confidence set Θ_t and the corresponding upper confidence bound

$$\mathrm{UCB}_{\Theta_t}(a) := \max_{\theta \in \Theta_t} \{\phi(a)^\top \boldsymbol{\theta}\}$$

- Play $a_{t+1} = \mathrm{argmax}_{a \in \mathcal{A}_{t+1}} \{\mathrm{UCB}_{\Theta_t}(a)\}$
- Observe reward $r_{t+1} = \phi(a_{t+1})^\top \boldsymbol{\theta}^* + \epsilon_{t+1}$

[1] Y. Abbasi-Yadkori et al. (2011) Improved algorithms for linear stochastic bandits. NeurIPS

In This Work

- New confidence sets Θ_t for stochastic linear bandits using a new tail bound for martingale mixtures

In This Work

- New confidence sets Θ_t for stochastic linear bandits using a new tail bound for martingale mixtures
- Provably tighter upper/lower confidence bounds than previous state-of-the-art (OFUL)

In This Work

- New confidence sets Θ_t for stochastic linear bandits using a new tail bound for martingale mixtures
- Provably tighter upper/lower confidence bounds than previous state-of-the-art (OFUL)
- LinUCB with our tighter confidence bounds leads to improved performance in hyperparameter tuning problems

In This Work

- New confidence sets Θ_t for stochastic linear bandits using a new tail bound for martingale mixtures
- Provably tighter upper/lower confidence bounds than previous state-of-the-art (OFUL)
- LinUCB with our tighter confidence bounds leads to improved performance in hyperparameter tuning problems
- LinUCB with our confidence sets has an $O(d\sqrt{T}\ln(T))$ worst-case cumulative regret bound (like OFUL)

Baseline Tail Bound

General Plan. Derive a data-dependent constraint for $\boldsymbol{\theta}^*$ using tail bounds for non-i.i.d. data.

Baseline Tail Bound

General Plan. Derive a data-dependent constraint for $\boldsymbol{\theta}^*$ using tail bounds for non-i.i.d. data.

We want a bound on the sum of squared errors

$$\|\Phi_t \boldsymbol{\theta}^* - \boldsymbol{r}_t\|_2^2 = \sum_{k=1}^{t} (\phi(a_k)^\top \boldsymbol{\theta}^* - r_k)^2 = \sum_{k=1}^{t} \epsilon_k^2.$$

Baseline Tail Bound

General Plan. Derive a data-dependent constraint for $\boldsymbol{\theta}^*$ using tail bounds for non-i.i.d. data.

We want a bound on the sum of squared errors

$$\|\Phi_t \boldsymbol{\theta}^* - \boldsymbol{r}_t\|_2^2 = \sum_{k=1}^{t} (\phi(a_k)^\top \boldsymbol{\theta}^* - r_k)^2 = \sum_{k=1}^{t} \epsilon_k^2.$$

Baseline tail bound. Use the σ-sub-Gaussian property of ϵ_k: with probability $\geq 1 - \delta$

$$\forall t \geq 1: \qquad \|\Phi_t \boldsymbol{\theta}^* - \boldsymbol{r}_t\|_2^2 \leq \sigma^2 t + 2\sigma^2 \sqrt{t \ln\left(\frac{t^2\pi^2}{6\delta}\right)} + 2\sigma^2 \ln\left(\frac{t^2\pi^2}{6\delta}\right).$$

Martingale Mixture Tail Bound for Linear Bandits

Choose a sequence of incrementally updated mean vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots$ and covariance matrices $\boldsymbol{T}_1, \boldsymbol{T}_2, \ldots$

$$\boldsymbol{\mu}_t = \left[\begin{array}{ccc|c} - & \boldsymbol{\mu}_{t-1} & - & \mu_t \end{array}\right]^\top, \quad \boldsymbol{T}_t = \left[\begin{array}{ccc|c} & & & T_1 \\ & \boldsymbol{T}_{t-1} & & \vdots \\ & & & T_{t-1} \\ \hline T_1 & \cdots & T_{t-1} & T_t \end{array}\right].$$

Martingale Mixture Tail Bound for Linear Bandits

Choose a sequence of incrementally updated mean vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots$ and covariance matrices $\boldsymbol{T}_1, \boldsymbol{T}_2, \ldots$

$$\boldsymbol{\mu}_t = \left[\begin{array}{c|c} - \quad \boldsymbol{\mu}_{t-1} \quad - & \mu_t \end{array}\right]^\top, \quad \boldsymbol{T}_t = \left[\begin{array}{ccc|c} & & & T_1 \\ & \boldsymbol{T}_{t-1} & & \vdots \\ & & & T_{t-1} \\ \hline T_1 & \cdots & T_{t-1} & T_t \end{array}\right].$$

μ_t and $T_1, \ldots, T_t$ can depend on the previous data $a_1, r_1, \ldots, a_{t-1}, r_{t-1}, a_t$.

Martingale Mixture Tail Bound for Linear Bandits

Choose a sequence of incrementally updated mean vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots$ and covariance matrices $\boldsymbol{T}_1, \boldsymbol{T}_2, \ldots$

$$\boldsymbol{\mu}_t = \left[\begin{array}{c|c} - \quad \boldsymbol{\mu}_{t-1} \quad - & \mu_t \end{array}\right]^\top, \quad \boldsymbol{T}_t = \left[\begin{array}{ccc|c} & & & T_1 \\ & \boldsymbol{T}_{t-1} & & \vdots \\ & & & T_{t-1} \\ \hline T_1 & \cdots & T_{t-1} & T_t \end{array}\right].$$

μ_t and $T_1, \ldots, T_t$ can depend on the previous data $a_1, r_1, \ldots, a_{t-1}, r_{t-1}, a_t$.

Standard choice. $\boldsymbol{\mu}_t = \mathbf{0}$, $\boldsymbol{T}_t = \Phi_t \Phi_t^\top = (\phi(a_i)^\top \phi(a_j))_{1 \le i,j \le t}$.

Martingale Mixture Tail Bound for Linear Bandits

Choose a sequence of incrementally updated mean vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots$ and covariance matrices $\boldsymbol{T}_1, \boldsymbol{T}_2, \ldots$

$$\boldsymbol{\mu}_t = \left[\begin{array}{ccc|c} - & \boldsymbol{\mu}_{t-1} & - & \mu_t \end{array} \right]^\top, \quad \boldsymbol{T}_t = \left[\begin{array}{ccc|c} & & & T_1 \\ & \boldsymbol{T}_{t-1} & & \vdots \\ & & & T_{t-1} \\ \hline T_1 & \cdots & T_{t-1} & T_t \end{array} \right].$$

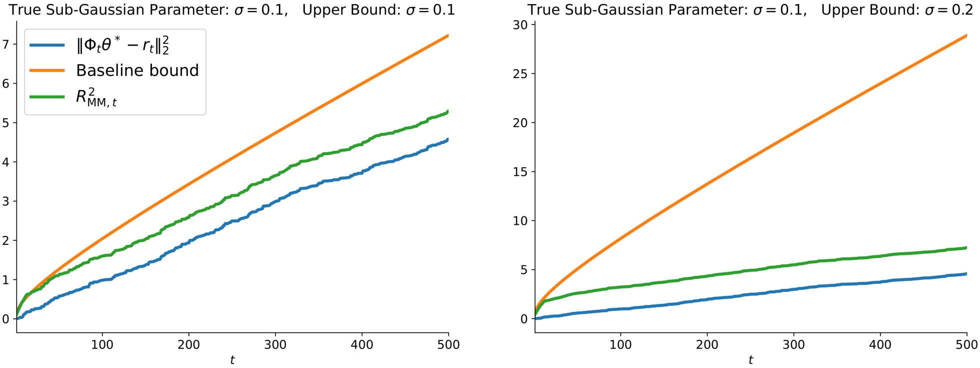
μ_t and $T_1, \ldots, T_t$ can depend on the previous data $a_1, r_1, \ldots, a_{t-1}, r_{t-1}, a_t$.

Standard choice. $\boldsymbol{\mu}_t = \mathbf{0}$, $\boldsymbol{T}_t = \Phi_t \Phi_t^\top = (\phi(a_i)^\top \phi(a_j))_{1 \le i,j \le t}$.

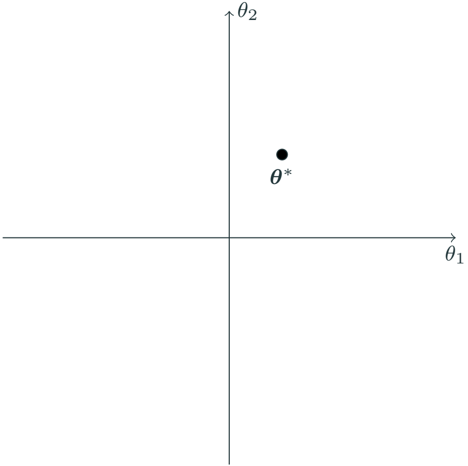
Martingale mixture tail bound. With probability $\ge 1 - \delta$, for all $t \ge 1$

$$\|\Phi_t \boldsymbol{\theta}^* - \boldsymbol{r}_t\|_2^2 \le (\boldsymbol{\mu}_t - \boldsymbol{r}_t)^\top \left(\boldsymbol{I} + \frac{\boldsymbol{T}_t}{\sigma^2} \right)^{-1} (\boldsymbol{\mu}_t - \boldsymbol{r}_t) + \sigma^2 \ln \left(\det \left(\boldsymbol{I} + \frac{\boldsymbol{T}_t}{\sigma^2} \right) \right) + 2\sigma^2 \ln(1/\delta) =: R_{\mathrm{MM},t}^2.$$

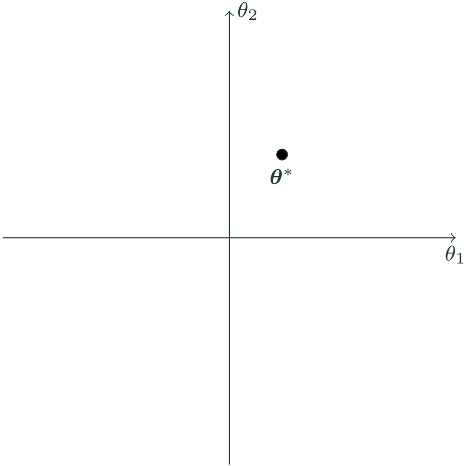
Our Tail Bound Against The Baseline Tail Bound



Confidence Sets For Linear Bandits



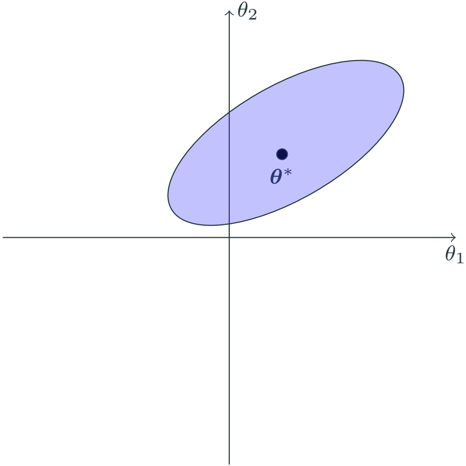
Confidence Sets For Linear Bandits



Using our martingale mixture tail bound, we have

$$\|\Phi_t \boldsymbol{\theta}^* - \boldsymbol{r}_t\|_2 \leq R_{\mathrm{MM},t},$$

Confidence Sets For Linear Bandits



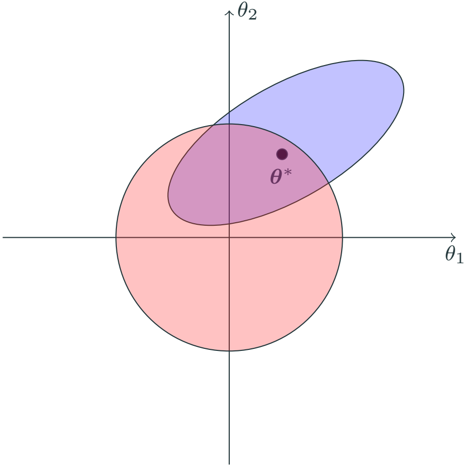
Using our martingale mixture tail bound, we have

$$\|\Phi_t \boldsymbol{\theta}^* - \boldsymbol{r}_t\|_2 \leq R_{\mathrm{MM},t},$$

This means that $\boldsymbol{\theta}^*$ lies within the set

$$\{\boldsymbol{\theta} \in \mathbb{R}^d : \|\Phi_t \boldsymbol{\theta} - \boldsymbol{r}_t\|_2 \leq R_{\mathrm{MM},t}\}.$$

Confidence Sets For Linear Bandits



Using our martingale mixture tail bound, we have

$$\|\Phi_t \boldsymbol{\theta}^* - \boldsymbol{r}_t\|_2 \leq R_{\mathrm{MM},t},$$

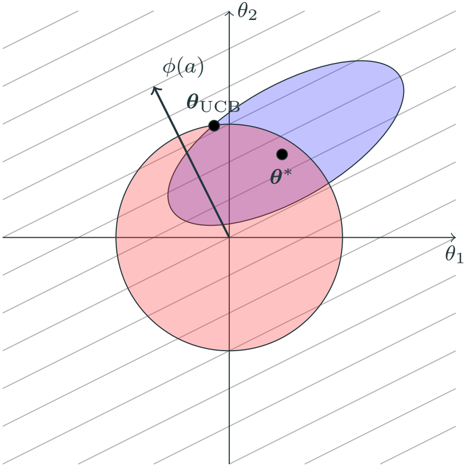
This means that $\boldsymbol{\theta}^*$ lies within the set

$$\{\boldsymbol{\theta} \in \mathbb{R}^d : \|\Phi_t \boldsymbol{\theta} - \boldsymbol{r}_t\|_2 \leq R_{\mathrm{MM},t}\}.$$

Incorporating the smoothness assumption, we obtain

$$\Theta_t = \{\boldsymbol{\theta} \in \mathbb{R}^d : \|\Phi_t \boldsymbol{\theta} - \boldsymbol{r}_t\|_2 \leq R_{\mathrm{MM},t}, \|\boldsymbol{\theta}\|_2 \leq B\}.$$

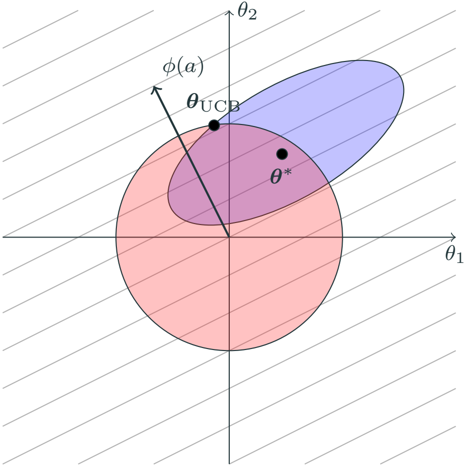
Convex Martingale Mixture UCB Algorithm



To run LinUCB with our confidence sets, we need to maximise $\mathrm{UCB}_{\Theta_t}(a)$ w.r.t. a, where

$$
\begin{aligned}
\mathrm{UCB}_{\Theta_t}(a) \;&=\; \max_{\boldsymbol{\theta}\in\mathbb{R}^d} \;\; \phi(a)^\top \boldsymbol{\theta} \\
&\qquad \text{s.t. } \; \|\Phi_t \boldsymbol{\theta} - \boldsymbol{r}_t\|_2 \le R_{\mathrm{MM},t} \\
&\qquad \text{and } \; \|\boldsymbol{\theta}\|_2 \le B \\
&=\; \phi(a)^\top \boldsymbol{\theta}_{\mathrm{UCB}}.
\end{aligned}
$$

Convex Martingale Mixture UCB Algorithm

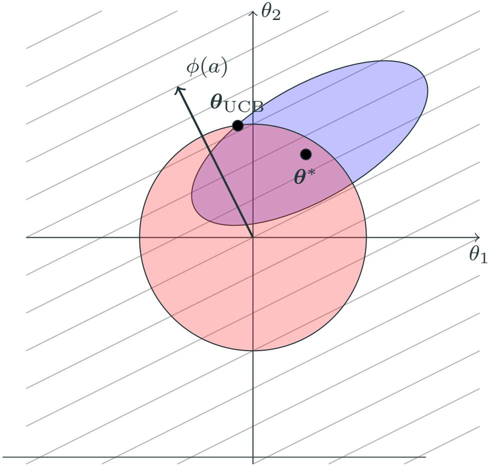


To run LinUCB with our confidence sets, we need to maximise $\mathrm{UCB}_{\Theta_t}(a)$ w.r.t. a, where

$$\begin{aligned} \mathrm{UCB}_{\Theta_t}(a) &= \max_{\boldsymbol{\theta} \in \mathbb{R}^d} \ \phi(a)^\top \boldsymbol{\theta} \\ &\qquad \text{s.t. } \|\Phi_t \boldsymbol{\theta} - \boldsymbol{r}_t\|_2 \leq R_{\mathrm{MM},t} \\ &\qquad \text{and } \|\boldsymbol{\theta}\|_2 \leq B \\ &= \phi(a)^\top \boldsymbol{\theta}_{\mathrm{UCB}}. \end{aligned}$$

For continuous action sets, we approximately maximise $\mathrm{UCB}_{\Theta_t}(a)$ w.r.t a using gradient-based methods.

Convex Martingale Mixture UCB Algorithm



To run LinUCB with our confidence sets, we need to maximise $\mathrm{UCB}_{\Theta_t}(a)$ w.r.t. a, where

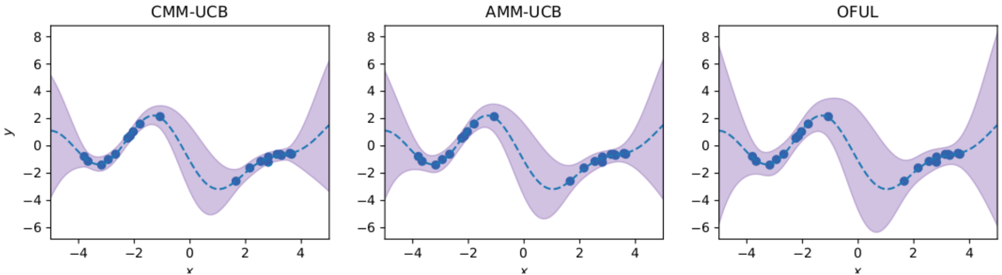
$$\begin{aligned}\mathrm{UCB}_{\Theta_t}(a) &= \max_{\boldsymbol{\theta}\in\mathbb{R}^d} \ \phi(a)^\top\boldsymbol{\theta} \\ &\qquad \text{s.t. } \|\Phi_t\boldsymbol{\theta}-\boldsymbol{r}_t\|_2 \le R_{\mathrm{MM},t} \\ &\qquad \text{and } \|\boldsymbol{\theta}\|_2 \le B \\ &= \phi(a)^\top\boldsymbol{\theta}_{\mathrm{UCB}}.\end{aligned}$$

For continuous action sets, we approximately maximise $\mathrm{UCB}_{\Theta_t}(a)$ w.r.t a using gradient-based methods.

We calculate $\mathrm{UCB}_{\Theta_t}(a)$ and $\nabla_a \mathrm{UCB}_{\Theta_t}(a)$ numerically using differentiable convex optimisation (cvxpylayers)[2].

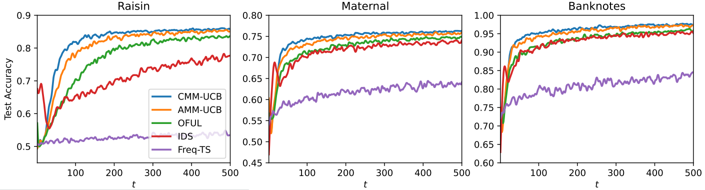
[2] A. Agrawal et al. (2019) Differentiable convex optimization layers. NeurIPS

Tighter Confidence Bounds



Hyperparameter Tuning

	Raisin		Maternal		Banknotes	
	Mean Acc	Max Acc	Mean Acc	Max Acc	Mean Acc	Max Acc
CMM-UCB (Ours)	**0.818** ± 0.018	**0.893** ± 0.019	**0.744** ± 0.020	**0.829** ± 0.023	**0.954** ± 0.005	**1.000** ± 0.000
AMM-UCB (Ours)	0.800 ± 0.017	0.892 ± 0.020	0.736 ± 0.020	**0.829** ± 0.023	0.948 ± 0.005	**1.000** ± 0.000
OFUL	0.764 ± 0.019	0.891 ± 0.019	0.722 ± 0.019	0.827 ± 0.022	0.929 ± 0.006	**1.000** ± 0.000
IDS[3]	0.706 ± 0.048	0.891 ± 0.020	0.714 ± 0.019	0.827 ± 0.024	0.926 ± 0.007	**1.000** ± 0.000
Freq-TS[4]	0.527 ± 0.022	0.884 ± 0.019	0.616 ± 0.018	0.823 ± 0.022	0.808 ± 0.012	**1.000** ± 0.000



[3] J. Kirschner and A. Krause. (2018) Information directed sampling and bandits with heteroscedastic noise, COLT

[4] S. Agrawal and N. Goyal. (2013) Thompson sampling for contextual bandits with linear payoffs, ICML

Regret Analysis

Step 1. Cumulative regret is bounded by the confidence bound widths (UCB minus LCB).

$$\sum_{t=1}^{T} \phi(a_t^*)^\top \boldsymbol{\theta}^* - \phi(a_t)^\top \boldsymbol{\theta}^* \leq \sum_{t=1}^{T} \mathrm{UCB}_{\Theta_{t-1}}(a_t^*) - \mathrm{LCB}_{\Theta_{t-1}}(a_t) \leq \sum_{t=1}^{T} \mathrm{UCB}_{\Theta_{t-1}}(a_t) - \mathrm{LCB}_{\Theta_{t-1}}(a_t).$$

Regret Analysis

Step 1. Cumulative regret is bounded by the confidence bound widths (UCB minus LCB).

$$\sum_{t=1}^{T} \phi(a_t^*)^\top \boldsymbol{\theta}^* - \phi(a_t)^\top \boldsymbol{\theta}^* \leq \sum_{t=1}^{T} \mathrm{UCB}_{\Theta_{t-1}}(a_t^*) - \mathrm{LCB}_{\Theta_{t-1}}(a_t) \leq \sum_{t=1}^{T} \mathrm{UCB}_{\Theta_{t-1}}(a_t) - \mathrm{LCB}_{\Theta_{t-1}}(a_t).$$

Step 2. Via weak duality for the convex program $\max_{\boldsymbol{\theta} \in \Theta_t} \{\phi(a)^\top \boldsymbol{\theta}\}$, we obtain

$$\sum_{t=1}^{T} \mathrm{UCB}_{\Theta_{t-1}}(a_t) - \mathrm{LCB}_{\Theta_{t-1}}(a_t) \leq \sum_{t=1}^{T} 2R_{\mathrm{AMM},t-1} \sqrt{\phi(a_t)^\top \left(\Phi_{t-1}^\top \Phi_{t-1} + \alpha \boldsymbol{I}\right)^{-1} \phi(a_t)}.$$

Regret Analysis

Step 1. Cumulative regret is bounded by the confidence bound widths (UCB minus LCB).

$$\sum_{t=1}^{T} \phi(a_t^*)^\top \boldsymbol{\theta}^* - \phi(a_t)^\top \boldsymbol{\theta}^* \leq \sum_{t=1}^{T} \mathrm{UCB}_{\Theta_{t-1}}(a_t^*) - \mathrm{LCB}_{\Theta_{t-1}}(a_t) \leq \sum_{t=1}^{T} \mathrm{UCB}_{\Theta_{t-1}}(a_t) - \mathrm{LCB}_{\Theta_{t-1}}(a_t).$$

Step 2. Via weak duality for the convex program $\max_{\boldsymbol{\theta} \in \Theta_t} \{\phi(a)^\top \boldsymbol{\theta}\}$, we obtain

$$\sum_{t=1}^{T} \mathrm{UCB}_{\Theta_{t-1}}(a_t) - \mathrm{LCB}_{\Theta_{t-1}}(a_t) \leq \sum_{t=1}^{T} 2R_{\mathrm{AMM},t-1} \sqrt{\phi(a_t)^\top \left(\Phi_{t-1}^\top \Phi_{t-1} + \alpha \boldsymbol{I}\right)^{-1} \phi(a_t)}.$$

Step 3. Separately upper bound $R_{\mathrm{AMM},T-1}$ and $\sum_{t=1}^{T} \sqrt{\phi(a_t)^\top \left(\Phi_{t-1}^\top \Phi_{t-1} + \alpha \boldsymbol{I}\right)^{-1} \phi(a_t)}$, to obtain

$$\sum_{t=1}^{T} \phi(a_t^*)^\top \boldsymbol{\theta}^* - \phi(a_t)^\top \boldsymbol{\theta}^* \leq \mathcal{O}(d\sqrt{T} \ln(T)).$$

Poster Session Information

Our paper title: "Improved Algorithms for Stochastic Linear Bandits Using Tail Bounds for Martingale Mixtures"

Poster #1801

Poster Session 6, Thu 14 December, 17:00 - 19:00 CST

We'll be happy to talk at the poster!