Linear Bandits: From Theory to Applications

Claire Vernade

DeepMind – Foundations Team
credits: Csaba Szepesvári, Tor Lattimore for their blog
Sequential Decision Making

\[ A_t \in A_t \subset \mathbb{R}^d \]

\[ X_t = f_\theta(A_t) + \eta_t \]
Real World Sequential Decision Making

\[ A_t \in \mathcal{A}_t \subset \mathbb{R}^d \]

Agent

Environment

\[ X_t, A_t, D_t \]

\[ X_t = f_\theta(A_t) + \eta_t \]

next step

delay buffer
1. Linear Bandits

2. Real-World Setting: Delayed Feedback
Linear Bandits
1. In round $t$, observe action set $\mathcal{A}_t \subset \mathbb{R}^d$.
2. The learner chooses $A_t \in \mathcal{A}_t$ and receives $X_t$, satisfying

$$\mathbb{E}[X_t | A_1, A_1, \ldots, A_t, A_t] = \langle A_t, \theta_* \rangle := f_{\theta_*}(A_t)$$

for some unknown $\theta_*$. 
3. Light-tailed noise:

$$X_t - \langle A_t, \theta_* \rangle = \eta_t \sim \mathcal{N}(0, 1)$$

Goal: Keep regret

$$R_n = \mathbb{E} \left[ \sum_{t=1}^{n} \max_{a \in \mathcal{A}_t} \langle a, \theta_* \rangle - X_t \right]$$

small.
Real-World setting

Typical setting: a user, represented by its feature vector $u_t$, shows up and we have a finite set of (correlated) actions $(a_1, \ldots, a_K)$.

Some function $\Phi$ joins these vectors pairwise to create a contextualized action set:

$$\forall i \in [K], \quad \Phi(u_t, a_i) = a_{t,i} \in \mathbb{R}^d \quad A_t = \{a_{t,1}, \ldots, a_{t,K}\}.$$

No assumption is to be made on the joining function $\Phi$ as the bandit may take over the decision step from that contextualized action set.

So, it is equivalent to $A_t \sim \prod(\mathbb{R}^d)$ some arbitrary distribution, or $A_1, \ldots, A_n$ fixed arbitrarily by the environment.
Say, reward in round $t$ is $X_t$, action in round $t$ is $A_t \in \mathbb{R}^d$:

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

We want to estimate $\theta^*$: regularized least-squares estimator:

$$\hat{\theta}_t = V_t^{-1} \sum_{s=1}^{t} A_s X_s,$$

$$V_0 = \lambda I, \quad V_t = V_0 + \sum_{s=1}^{t} A_s A_s^\top.$$

Choice of confidence regions (ellipsoids) $C_t$:

$$C_t \triangleq \left\{ \theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_{t-1}\|_{V_{t-1}}^2 \leq \beta_t \right\}.$$

where, for $A$ positive definite, $\|x\|_A^2 = x^\top A x$. 
“Choose the best action in the best environment amongst the plausible ones.”

Choose $C_t$ with suitable $(\beta_t)_t$ and let

$$A_t = \arg\max_{a \in A} \max_{\theta \in C_t} \langle a, \theta \rangle.$$  

Or, more concretely, for each action $a \in A$, compute the ”optimistic index”

$$U_t(a) = \max_{\theta \in C_t} \langle a, \theta \rangle.$$  

Maximising a linear function over a convex closed set, the solution is explicit:

$$A_t = \arg\max_a U_t(a) = \arg\max_a \langle a, \hat{\theta}_t \rangle + \sqrt{\beta_t} \| a \|_{V_{t-1}}^{-1}.$$ 

Optimism in the Face of Uncertainty Principle

\[ a_{1,t} = A_t \]

\[ a_{2,t} \]

\[ C_t \]

\[ \theta \]

\[ \hat{\theta}_t \]

Regret: 11.00
Timestep: 19
Regret Bound

Assumptions:

1. *Bounded scalar mean reward*: $|\langle a, \theta_* \rangle| \leq 1$ for any $a \in \bigcup_t A_t$.
2. *Bounded actions*: for any $a \in \bigcup_t A_t$, $\|a\|_2 \leq L$.
3. *Honest confidence intervals*: There exists a $\delta \in (0, 1)$ such that with probability $1 - \delta$, for all $t \in [n]$, $\theta_* \in C_t$ for some choice of $(\beta_t)_{t \leq n}$.

**Theorem (LinUCB Regret)**

Let the conditions listed above hold. Then with probability $1 - \delta$ the regret of LinUCB satisfies

$$\hat{R}_n \leq \sqrt{8dn\beta_n \log \left( \frac{d\lambda + nL^2}{d\lambda} \right)}.$$
Proof

Jensen’s inequality shows that

\[
\hat{R}_n = \sum_{t=1}^{n} \langle A_t^* - A_t, \theta \rangle := \sum_{t=1}^{n} r_t \leq \sqrt{n \sum_{t=1}^{n} r_t^2}
\]

where \( A_t^* = \text{argmax}_{a \in A_t} \langle a, \theta^*_t \rangle \).

Let \( \tilde{\theta}_t \) be the vector that realizes the maximum over the ellipsoid:

\( \tilde{\theta}_t \in C_t \) s.t. \( \langle A_t, \tilde{\theta}_t \rangle = U_t(A_t) \).

From the definition of LinUCB,

\[
\langle A_t^*, \theta^*_t \rangle \leq U_t(A_t^*) \leq U_t(A_t) = \langle A_t, \tilde{\theta}_t \rangle.
\]

Then,

\[
r_t \leq \langle A_t, \tilde{\theta}_t - \theta^* \rangle \leq \|A_t\|_{V_t^{-1}} \|\tilde{\theta}_t - \theta^*_t\|_{V_{t-1}} \leq 2 \|A_t\|_{V_t^{-1}} \sqrt{\beta_t}.
\]
So we now have a new upper bound,

\[
\hat{R}_n = \sum_{t=1}^{n} r_t \leq \sqrt{n \sum_{t=1}^{n} r_t^2} \leq 2 \sqrt{n \beta_n \sum_{t=1}^{n} (1 \wedge \|A_t\|_{V_t^{-1}}^2)}.
\]

**Lemma (Abbasi-Yadkori et al. (2011))**

Let \(x_1, \ldots, x_n \in \mathbb{R}^d\), \(V_t = V_0 + \sum_{s=1}^{t} x_s x_s^\top\), \(t \in [n]\), and \(L \geq \max_t \|x_t\|_2\). Then,

\[
\sum_{t=1}^{n} \left(1 \wedge \|x_t\|_{V_t^{-1}}^2\right) \leq 2 \log \left(\frac{\det V_n}{\det V_0}\right) \leq d \log \left(\frac{\text{trace}(V_0) + nL^2}{d \det^{1/d}(V_0)}\right).
\]
Confidence Ellipsoids

Assumptions: $\|\theta_\ast\| \leq S$, and let $(A_s)_s, (\eta_s)_s$ be so that for any $1 \leq s \leq t$, $\eta_s |\mathcal{F}_{s-1} \sim \text{subG}(1)$, where $\mathcal{F}_s = \sigma(A_1, \eta_1, \ldots, A_{s-1}, \eta_{s-1}, A_s)$

Fix $\delta \in (0, 1)$. Let

$$\beta_{t+1} = \sqrt{\lambda} S + \sqrt{2 \log \left( \frac{1}{\delta} \right)} + \log \left( \frac{\det V_t(\lambda)}{\lambda^d} \right)$$

$$\leq \sqrt{\lambda} S + \sqrt{2 \log \left( \frac{1}{\delta} \right)} + \log \left( \frac{\lambda d + nL^2}{d^2} \right),$$

and

$$C_{t+1} = \left\{ \theta \in \mathbb{R}^d : \|\hat{\theta}_t - \theta_\ast\|_{V_t(\lambda)} \leq \beta_{t+1} \right\}.$$

**Theorem**

$C_{t+1}$ is a confidence set for $\theta_\ast$ at level $1 - \delta$:

$$\mathbb{P} (\theta_\ast \in C_{t+1}) \geq 1 - \delta.$$

**Proof**: See Chapter 20 of *Bandit Algorithms* (www.banditalgs.com)

• Auer [6] was the first to consider optimism for linear bandits (LinRel, SupLinRel). Main restriction: $|A_t| < +\infty$.


• The name LinUCB comes from Chu et al. [7].

• Alternative routes:
  • Explore then commit for action sets with smooth boundary. Abbasi-Yadkori [1], Abbasi-Yadkori et al. [2], Rusmevichientong and Tsitsiklis [11].
  • Phased elimination
  • Thompson sampling
Theorem (LinUCB Regret)

Let the conditions listed above hold. Then with probability $1 - \delta$ the regret of LinUCB satisfies

$$\hat{R}_n \leq \sqrt{8dn\beta_n \log \left( \frac{\text{trace}(V_0) + nL^2}{d \det^{\frac{1}{d}}(V_0)} \right)} = O(d\sqrt{n}).$$

Linear bandits are an elegant model of the exploration-exploitation dilemma when actions are correlated.

The main ingredients of the regret analysis are:

- bounding the instantaneous regret using the definition of optimism;
- a maximal concentration inequality holding for a randomized, sequential design;
- the Elliptical Potential Lemma.
Real-World Setting: Delayed Feedback
In a real-world application, rewards are delayed ...
In a real-world application, rewards are delayed ... and censored.
Modified setting: at round $t \geq 1$,

- receive contextualized action set $A_t = \{a_1, \ldots, a_K\}$ and choose action $A_t \in A_t$,
- two random variables are generated but not observed: $X_t \sim B(\theta^T A_t)$ and $D_t \sim \mathcal{D}(\tau)$,
- at $t + D_t$ the reward $X_t$ of action $A_t$ is disclosed ...
- ...unless $D_t > m$: If the delay is too long, the reward is discarded.

New parameter: $0 < m < T$ is the cut-off time of the system. If the delay is longer, the reward is never received. The delay distribution $\mathcal{D}(\tau)$ characterizes the proportion of converting actions: $\tau_m = p(D_t \leq m)$. 
A new estimator

We now have:

\[ V_t = \sum_{s=1}^{t-1} A_s A_s^\top \quad \tilde{b}_t = \sum_{s=1}^{t-1} A_s X_s 1\{D_s \leq m\} \]

where \( \tilde{b}_t \) contains additional non-identically distributed samples:

\[ \tilde{b}_t = \sum_{s=1}^{t-m} A_s X_s 1\{D_s \leq m\} + \sum_{s=t-m+1}^{t-1} A_s X_s 1\{D_s \leq t - s\} \]

"Conditionally biased" least squares estimator includes every received feedback

\[ \hat{\theta}_t^b = V_t^{-1} \tilde{b}_t \]

Baseline: use previous estimator but discard last \( m \) steps

\[ \hat{\theta}_t^{\text{disc}} = V_{t-m}^{-1} b_{t-m} \quad \text{with} \quad \mathbb{E}[\hat{\theta}_t^{\text{disc}}|F_t] \approx \tau_m \theta \]
Confidence interval and the D-LinUCB policy

We remark that

\[
\hat{\theta}_t^b - \tau_m \theta = \hat{\theta}_t^b - \hat{\theta}_{t+m}^{\text{disc}} + \hat{\theta}_{t+m}^{\text{disc}} - \tau_m \theta
\]

\[
= \hat{\theta}_t^b - \hat{\theta}_{t+m}^{\text{disc}} + \underbrace{\hat{\theta}_{t+m}^{\text{disc}} - \tau_m \theta}_{\text{finite bias}} + \underbrace{\hat{\theta}_{t+m}^{\text{disc}} - \tau_m \theta}_{\text{same as before}}
\]

For the new \( C_t \), we have new optimistic indices

\[
A_t = \arg\max_{a \in A} \max_{\theta \in C_t} \langle a, \theta \rangle.
\]

But now, the solution has an extra (vanishing) bias term

\[
A_t = \arg\max_a \langle a, \hat{\theta}_t \rangle + \sqrt{\beta_t} \| a \|_{V_{t-1}}^{-1} + m \| a \|_{V_{t-1}}^{-2}.
\]

D-LinUCB: Easy, straightforward, harmless modification of LinUCB, with regret guarantees in the delayed feedback setting.
Theorem (D-LinUCB Regret)

Under the same conditions as before, with $V_0 = \lambda I$, with probability $1 - \delta$ the regret of D-LinUCB satisfies

$$\hat{R}_n \leq \tau_m^{-1} \sqrt{8dn\beta_n \log \left( \frac{\text{trace}(V_0) + nL^2}{d \det^{\frac{1}{d}}(V_0)} \right)} + \frac{dm}{(\lambda - 1)\tau_m^{-1}} \log \left( 1 + \frac{n}{d(\lambda - 1)} \right).$$
We fix $n = 3000$ and generate geometric delays with $\mathbb{E}[D_t] = 100$. In a real setting, this would correspond to an experiment that lasts 3h, with average delays of 6 minutes.

Then, we let the cut off vary $m \in 250, 500, 1000$, i.e. waiting time of 15min, 30min and 1h, respectively.

**Figure 1:** Comparison of the simulated behaviors of D-LinUCB and (waiting)LinUCB
Conclusions

- Linear Bandits are a powerful and well-understood way of solving the exploration-exploitation trade-off in a metric space;
- The techniques have been extended to Generalized Linear models by Filippi et al. [9]
- and to kernel regression Valko et al. [12, 13].
- Yet, including constraints and external sources of noise in real-world application is challenging.
- Some use cases challenge the bandit model assumptions...
- ... and then it’s time to open the box of MDP’s (e.g. UCRL abd KL-UCRL Auer et al. [5], Filippi et al. [10]).
Conclusions

• Linear Bandits are a powerful and well-understood way of solving the exploration-exploitation trade-off in a metric space;
• The techniques have been extended to Generalized Linear models by Filippi et al. [9]
• and to kernel regression Valko et al. [12, 13].
• Yet, including constraints and external sources of noise in real-world application is challenging.
• Some use cases challenge the bandit model assumptions...
• ... and then it’s time to open the box of MDP’s (e.g. UCRL abd KL-UCRL Auer et al. [5], Filippi et al. [10]).

Thanks!


