Improved Algorithms for Linear Stochastic Bandits
a NIPS article by Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári

Émilie Kaufmann

SMILE NIPS’11 Defrief session, February 16th 2012
The linear bandit problem

Setting

- set of arms $\mathcal{D} \subset \mathbb{R}^d$
- unknown parameter $\theta^* \in \mathbb{R}^d$
- best arm (unknown) $x^* = \arg\max_{x \in \mathcal{D}} \langle x | \theta^* \rangle$
- At time $t$ pick an arm $X_t \in \mathcal{D}$ and observe reward:

$$Y_t = \langle X_t | \theta^* \rangle + \eta_t$$

where the noise $\eta_t$ is centered ($\mathbb{E}[\eta_t | \mathcal{F}_{t-1}] = 0$) and subgaussian:

$$\mathbb{E} \left[ e^{\lambda \eta_t} | \mathcal{F}_{t-1} \right] \leq \exp \left( \frac{\lambda^2}{2} \right)$$  \hspace{1cm} (1)

- with $\mathcal{F}_t = \sigma(X_1, \ldots, X_t, X_{t+1}, \eta_1, \ldots, \eta_{t-1}, \eta_t)$
The linear bandit problem

**Goal:** Design an algorithm (ie a sequential choice of the arms) minimizing the *regret*:

\[
R_n = \sum_{t=1}^{n} \langle x^* \mid \theta^* \rangle - \sum_{t=1}^{n} \langle X_t \mid \theta^* \rangle = \sum_{t=1}^{n} \langle x^* - X_t \mid \theta^* \rangle
\]

**In the article:** a regret bound for the regret \( R_n \) of the OFUL algorithm, holding with high probability

**Some applications:**
- \( \mathcal{D} \subset \{0, 1\}^d \) : online shortest path problem
- \( \mathcal{D} \) is finite : marketing ($$)
A special case: the multiarmed bandit problem

The multiarmed bandit problem in the gaussian case

- $K$ independent arms
- arm $j$ is an i.i.d. sequence of $\mathcal{N}(\mu_j, 1)$ (known variance)
- if arm $I_t$ is chosen, the reward is drawn from $\mathcal{N}(\mu_j, 1)$

Why does it fit our setting?

- $\mathcal{D} = \{e_1, \ldots, e_K\}$ the canonical basis of $\mathbb{R}^K$
- $\theta^* = (\mu_1, \ldots, \mu_K)^T$
- if arm $I_t$ is drawn at time $t$ we get the reward

$$Y_t = \langle e_{I_t} | \theta^* \rangle + \eta_t$$

with $\eta_t$ an i.i.d sequence of $\mathcal{N}(0, 1)$
A special case: the multiarmed bandit problem

The multiarmed bandit problem in the gaussian case

- $K$ independent arms
- arm $j$ is an i.i.d. sequence of $\mathcal{N}(\mu_j, 1)$ (known variance)
- if arm $I_t$ is chosen, the reward is drawn from $\mathcal{N}(\mu_j, 1)$

Why does it fit our setting?

- $D = \{e_1, ..., e_K\}$ the canonical basis of $\mathbb{R}^K$
- $\theta^* = (\mu_1, ..., \mu_K)^T$
- if arm $I_t$ is drawn at time $t$ we get the reward

$$Y_t = \langle e_{I_t} \mid \theta^* \rangle + \eta_t$$

with $\eta_t$ an i.i.d sequence of $\mathcal{N}(0, 1)$

$\Rightarrow$ Generalize UCB optimistic heuristic for the Linear Case
Some ideas to solve the problem

With $X_{1:t} = \begin{pmatrix} X_1^T \\ \vdots \\ X_t^T \end{pmatrix} \in \mathbb{R}^{t \times d}$, $Y_{1:t} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_t \end{pmatrix}$, $\eta_t = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_t \end{pmatrix} \in \mathbb{R}^t$:

$$Y_{1:t} = X_{1:t}\theta^* + \eta_t$$

- **First idea:** use a regularized least-square estimate of $\theta^*$:

$$\hat{\theta}_t = (V_t)^{-1} X_{1:t}^T Y_{1:t}$$

where

$$V_t = \lambda I + X_{1:t}^T X_{1:t}$$

and choose $X_{t+1} = \arg\max_{x \in \mathcal{D}} \langle x \mid \hat{\theta}_t \rangle$
Some ideas to solve the problem

With $X_{1:t} = \begin{pmatrix} X_1^T \\ \vdots \\ X_t^T \end{pmatrix} \in \mathbb{R}^{t \times d}$, $Y_{1:t} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_t \end{pmatrix}$, $\eta_t = \begin{pmatrix} \eta_1 \\ \vdots \\ \eta_t \end{pmatrix} \in \mathbb{R}^t$:

$$Y_{1:t} = X_{1:t} \theta^* + \eta_t$$

- First idea: use a regularized least-square estimate of $\theta^*$:

$$\hat{\theta}_t = (V_t)^{-1} X_{1:t}^T Y_{1:t}$$

where

$$V_t = \lambda I + X_{1:t}^T X_{1:t}$$

and choose $X_{t+1} = \arg\max_{x \in D} \langle x | \hat{\theta}_t \rangle$. $\Rightarrow$ Exploitation only!
A generic optimistic algorithm for the linear bandit:

- Build a confidence ellipsoid around $\hat{\theta}_t$:

$$C_t(\delta) = \left\{ \theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_t\|_{V_t} \leq \beta_t(\delta) \right\}$$

such that $\mathbb{P} \left( \forall t \in \mathbb{N}, \theta^* \in C_t(\delta) \right) \geq 1 - \delta$

- at time $t$ choose $X_t$ such that:

$$(X_t, \tilde{\theta}_t) = \arg\max_{(x, \theta) \in \mathcal{D} \times C_{t-1}(\delta)} \langle x | \theta \rangle$$
The previous optimization problem rewrites:

$$X_t = \arg\max_{X \in D} \langle X, \hat{\theta}_t \rangle + \beta_t(\delta) \|X\|\sqrt{V_t^{-1}}$$

For the multiarmed bandit case:

- $V_t = \text{Diag}(N_t(1), ..., N_t(K))$ where $N_t(j)$ is the number of draws of arm $j$ up to time $t$
- Denotes $R_j(t)$ the sum of reward gathered from arm $j$ up to time $t$
- Arm chosen at time $t$ is the one with highest index:

$$\frac{R_t(j)}{N_t(j)} + \frac{\beta_t(\delta)}{\sqrt{N_t(j)}}$$
The previous optimization problem rewrites:

\[ X_t = \arg\max_{X \in D} \langle X, \hat{\theta}_t \rangle + \beta_t(\delta) \|X\|_{V_t^{-1}} \]

For the mutiarmed bandit case:

- \( V_t = \text{Diag}(N_t(1), \ldots, N_t(K)) \) where \( N_t(j) \) is the number of draws of arm \( j \) up to time \( t \)
- Denotes \( R_j(t) \) the sum of reward gathered from arm \( j \) up to time \( t \)
- Arm chosen at time \( t \) is the one with highest index:

\[ \frac{R_t(j)}{N_t(j)} + \frac{\beta_t(\delta)}{\sqrt{N_t(j)}} \]

\( \Rightarrow \beta_t(\delta) \) is an exploration rate... and should be logarithmic!
Building the confidence region

Previously: (in a non-regularized setting)
- Dani, Kakade 2008:
  \[
  \beta_t(\delta) = \sqrt{\max \left( 128d \ln(t) \ln\left(\frac{t^2}{\delta}\right), \frac{64}{9} \ln^2\left(\frac{t^2}{\delta}\right) \right)}
  \]
- Rusmeviechentong, Tsitsiklis 2009:
  \[
  \beta_t(\delta) = K \sqrt{\log(t)} \sqrt{d \log(t) + \log(t^2/\delta)}
  \]
- Empirically smaller confidence region seems to work!

In the article:
- a new (elegant) technique for building a smaller confidence region
- better empirical performance!
Finding the martingale

Let $S_t = X^T_{1:t} \eta_{1:t} = \sum_{k=1}^{t} \eta_k X_k$. $S_t$ is a $\mathcal{F}_t$ martingale and:

$$\forall x \in \mathbb{R}^d, \ |x'(\hat{\theta}_t - \theta^*)| \leq ||x||_{\mathcal{V}_{t-1}} \left( ||S_t||_{\mathcal{V}_{t-1}} + \sqrt{\lambda}||\theta^*|| \right)$$

Therefore, finding $\beta_t(\delta)$ such that

$$\mathbb{P} \left( \forall t \in \mathbb{N}, ||S_t||_{\mathcal{V}_{t-1}} \leq \beta_t(\delta) \right) \geq 1 - \delta$$

leads to:

$$\mathbb{P} \left( \forall t \in \mathbb{N}, ||\hat{\theta}_t - \theta^*||_{\mathcal{V}_t} \leq \beta_t(\delta) + \sqrt{\lambda}||\theta^*|| \right) \geq 1 - \delta$$

For the bandit case:

$$||S_t||_{\mathcal{V}_{t-1}} = \sum_{j=1}^{k} \frac{R_t(j) - \mu_j N_t(j)}{\sqrt{\lambda + N_t(j)}}$$
A self-normalized inequality for a vector-valued martingale

• Bounding its tail

**Theorem**

Let $X_t$ and $\eta_t$ be two real-valued stochastic processes such that $X_t$ is $\mathcal{F}_{t-1}$ measurable, $\eta_t$ is $\mathcal{F}_t$ measurable and satisfies (1) Noting

$$V_t = V + \sum_{s=1}^{t} X_s X_s^T \quad \text{and} \quad S_t = \sum_{s=1}^{t} \eta_s X_s$$

Then, for any $\delta > 0$, with probability at least $1 - \delta$,

$$\forall t \geq 0 \quad \|S_t\|^2_{V_t^{-1}} \leq 2 \log \left( \frac{\det(V_t)^{1/2} \det(V)^{-1/2}}{\delta} \right)$$

Generalizes a technique by De La Peña (2004)
Sketch of the proof (1/2)
Sketch of the proof (1/2)

\[ M_t^\lambda = \exp \left( \langle \lambda | S_t \rangle - \frac{1}{2} ||\lambda||^2 (V_t - V) \right) \]

is a supermartingale such that for all stopping time \( \tau \),

\[ \mathbb{E}[M_{\tau}^\lambda] \leq 1 \]
Sketch of the proof (1/2)

- \( \forall \lambda \in \mathbb{R}^d, \)

\[
M^\lambda_t = \exp \left( \langle \lambda | S_t \rangle - \frac{1}{2} \| \lambda \|^2 \right) (V_t - V)
\]

is a supermartingale such that for all stopping time \( \tau, \)

\[ \mathbb{E}[M^\lambda_\tau] \leq 1 \]

- Average over \( \lambda \) by setting a ’prior’ \( \Lambda \sim \mathcal{N}(0, V^{-1}) \)
Sketch of the proof (1/2)

- \( \forall \lambda \in \mathbb{R}^d \),

\[
M^\lambda_t = \exp \left( \langle \lambda | S_t \rangle - \frac{1}{2} ||\lambda||^2 (V_t - V) \right)
\]

is a supermartingale such that for all stopping time \( \tau \),
\[
\mathbb{E}[M^\lambda_\tau] \leq 1
\]

- Average over \( \lambda \) by setting a 'prior' \( \Lambda \sim \mathcal{N}(0, V^{-1}) \)
  - on the one hand \( \mathbb{E}[M_t(\Lambda)] \leq 1 \)
Sketch of the proof (1/2)

- $\forall \lambda \in \mathbb{R}^d$, 

$$M^\lambda_t = \exp \left( \langle \lambda \mid S_t \rangle - \frac{1}{2} \|\lambda\|^2_{(V_t-V)} \right)$$

is a supermartingale such that for all stopping time $\tau$, 

$$\mathbb{E}[M^\lambda_{\tau}] \leq 1$$

- Average over $\lambda$ by setting a 'prior' $\Lambda \sim \mathcal{N}(0, V^{-1})$
  - on the one hand $\mathbb{E}[M_t(\Lambda)] \leq 1$
  - on the other hand $\mathbb{E}[M_t(\Lambda)] = \mathbb{E}[\mathbb{E}[M_t(\Lambda) \mid \mathcal{F}_\infty]]$ and

$$\mathbb{E}[M_t(\Lambda) \mid \mathcal{F}_\infty] = \int_{\mathbb{R}^d} \exp \left( \langle \lambda \mid S_t \rangle - \frac{1}{2} \|\lambda\|^2_{(V_t-V)} \right) f(\lambda) \, d\lambda$$

where $f$ denotes the pdf of $\mathcal{N}(0, V^{-1})$
Sketch of the proof (2/2)

\[
\begin{align*}
\mathbb{E}\left[ \det(V_t)^{1/2} \exp\left( \frac{1}{2} \|S_t\|^2 V_t^{-1} \right) \right] &= \mathbb{E}\left[ \det(V_t)^{1/2} \right] \leq 1 \\
\text{Using Markov inequality yields:} \\
P\left( \|S_t\|^2 V_t^{-1} > 2 \log \left( \det(V_t)^{1/2} \delta \det(V_t)^{1/2} \right) \right) &= P\left( \exp\left( \frac{1}{2} \|S_t\|^2 V_t^{-1} \right) > \det(V_t)^{1/2} \delta \det(V_t)^{1/2} \right) \leq \delta
\end{align*}
\]
Sketch of the proof (2/2)

Direct calculation shows that

\[ E[M_t(\Lambda) | \mathcal{F}_\infty] = \left( \frac{\det(V)}{\det(V_t)} \right)^{1/2} \exp \left( \frac{1}{2} \|S_t\|^2 V_t^{-1} \right) \]
Sketch of the proof (2/2)

- Direct calculation show that
  \[
  \mathbb{E}[M_t(\Lambda)|\mathcal{F}_\infty] = \left(\frac{\det(V)}{\det(V_t)}\right)^{1/2} \exp\left(\frac{1}{2} \|S_t\|_{V_t^{-1}}^2\right)
  \]

- Then \[ \mathbb{E}\left[\left(\frac{\det(V)}{\det(V_t)}\right)^{1/2} \exp\left(\frac{1}{2} \|S_t\|_{V_t^{-1}}^2\right)\right] \leq 1 \]
Sketch of the proof (2/2)

- Direct calculation show that

\[ \mathbb{E}[M_t(\Lambda) | \mathcal{F}_{\infty}] = \left( \frac{\det(V)}{\det(V_t)} \right)^{1/2} \exp \left( \frac{1}{2} \|S_t\|^2 V_t^{-1} \right) \]

- Then \( \mathbb{E} \left[ \left( \frac{\det(V)}{\det(V_t)} \right)^{1/2} \exp \left( \frac{1}{2} \|S_t\|^2 V_t^{-1} \right) \right] \leq 1 \)

- Using Markov inequality yields:

\[ \mathbb{P} \left( \|S_t\|^2 V_t^{-1} > 2 \log \left( \frac{\det(V_t)^{1/2}}{\delta \det(V)^{1/2}} \right) \right) \leq \delta \]

\[ = \mathbb{P} \left( \exp \left( \frac{1}{2} \|S_t\|^2 V_t^{-1} \right) > \frac{\det(V_t)^{1/2}}{\delta \det(V)^{1/2}} \right) \leq \delta \]
Sketch of the proof (2/2)

Direct calculation show that

\[ \mathbb{E}[M_t(\Lambda)|\mathcal{F}_\infty] = \left( \frac{\det(V)}{\det(V_t)} \right)^{1/2} \exp \left( \frac{1}{2} \|S_t\|_V^{-1} \right) \]

Then \( \mathbb{E} \left[ \left( \frac{\det(V)}{\det(V_t)} \right)^{1/2} \exp \left( \frac{1}{2} \|S_t\|_V^{-1} \right) \right] \leq 1 \)

Using Markov inequality yields:

\[ \mathbb{P} \left( \|S_t\|_V^{-1} > 2 \log \left( \frac{\det(V_t)^{1/2}}{\delta \det(V)^{1/2}} \right) \right) \]

\[ = \mathbb{P} \left( \exp \left( \frac{1}{2} \|S_t\|_V^{-1} \right) \leq \frac{\det(V_t)^{1/2}}{\delta \det(V)^{1/2}} \right) \leq \delta \]

For a bound uniform in \( t \): use stopping times
Final result for the regret

- **The OFUL algorithm**

Finally we use a confidence region based on the exploration rate

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

**Theorem**

Assume that $||\theta^*|| \leq S$ and for all $x \in D$, $||x|| \leq L$, $\langle x \mid \theta^* \rangle \in [-1, 1]$. Then with probability at least $1 - \delta$, the regret of the OFUL algorithm satisfies: $\forall n \geq 0$

$$R_n \leq 4 \sqrt{nd \log (\lambda + nL/d)} \left( \sqrt{\lambda} S + \sqrt{2 \log \frac{1}{\delta}} + d \log \left(1 + \frac{nL}{\lambda d}\right) \right)$$
Discussion

- optimality? Here $R_n = O(d\sqrt{n})$ w.h.p
  - lower bound for the worst case regret $O(\sqrt{dn})$ for multiarmed bandit
  - lower bound $O(d\sqrt{n})$ for the linear bandit?

- regularized versus non-regularized approaches

- empirical performance