The End of Optimism?†

(and posterior sampling)

Tor Lattimore and **Csaba Szepesvári**

AISTATS 2017 (and arXiv)

† for linear bandits, not life

Purpose of talk (besides...)

To show that the two standard design principles for stochastic multi-armed bandits have **serious drawbacks** beyond the **simplest** case

(and offer some alternatives)

Formal model

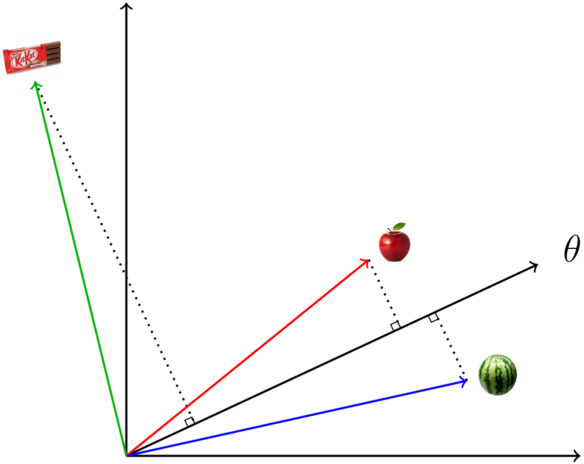
- ⋆ A set of actions $\mathcal{A}$ (given in advance)
- ⋆ For each $x \in \mathcal{A}$ there is a reward distribution P_x (unknown)
- ⋆ In each round you choose $A_t \in \mathcal{A}$
- ⋆ Observe reward $Y_t \sim P_{A_t}$
- ⋆ Want to collect as much reward as possible
- ⋆ The total number of rounds (interactions) is n

Example 1 $\mathcal{A} = \{1, \ldots, k\}$ and $Y_t \sim \mathcal{B}(\mu_{A_t})$ with $\mu \in [0, 1]^k$

Example 2 $\mathcal{A} \subset \mathbb{R}^d$ and $Y_t = \langle A_t, \theta \rangle + \eta_t$ with $\theta \in \mathbb{R}^d$ and η_t noise

$\mu_x = \langle x, \theta \rangle$ and $\mu^* = \max_{x \in \mathcal{A}} \mu_x$ and $\Delta_x = \mu^* - \mu_x$

$d = 2$ and $k = 3$ and $\mathcal{A} = \{$, , $\}$



Watermelon is optimal because $\langle x, \theta \rangle \propto \|x\| \cos \mathrm{angle}(x, \theta)$

Regret

Regret is the difference between the rewards you expect with the optimal strategy and what you expect

$$
\begin{aligned}
R_n &= n \max_{x \in \mathcal{A}} \mu_x - \mathbb{E}\left[\sum_{t=1}^{n} \mu_{A_t}\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n} \Delta_{A_t}\right]
\end{aligned}
$$

Maximise reward $\Leftrightarrow$ minimising regret

Strategy is **consistent** if $R_n = o(n^p) \; \forall p > 0$

How small can we make the regret?

Optimism for linear bandits

In each round, construct confidence set $\mathcal{C}_t \subseteq \mathbb{R}^d$ such that

$$\theta \in \mathcal{C}_t \qquad \text{with high probability}$$

Then choose action

$$A_t = \operatorname*{arg\,max}_{x \in \mathcal{A}} \max_{\tilde{\theta} \in \mathcal{C}_t} \langle x, \tilde{\theta} \rangle$$

Optimism for linear bandits

In each round, construct confidence set $\mathcal{C}_t \subseteq \mathbb{R}^d$ such that

$$\theta \in \mathcal{C}_t \qquad \text{with high probability}$$

Then choose action

$$A_t = \arg\max_{x \in \mathcal{A}} \max_{\tilde{\theta} \in \mathcal{C}_t} \langle x, \tilde{\theta} \rangle$$

Why it works: with high probability

$$\begin{aligned} \Delta_{A_t} &= \langle x^* - A_t, \theta \rangle = \langle x^*, \theta \rangle - \langle A_t, \theta \rangle \\ &\leq \langle A_t, \tilde{\theta} \rangle - \langle A_t, \theta \rangle = \underbrace{\langle A_t, \tilde{\theta} - \theta \rangle}_{\text{width of confidence set in direction } A_t} \end{aligned}$$

Confidence set construction

$$G_t = \sum_{s=1}^{t-1} A_s A_s^\top$$ (Gram matrix)

$$\hat{\theta}_t = G_t^{-1} \sum_{s=1}^{t-1} A_s Y_s$$ (Least squares estimator)

$$w.p{>}1{-}\delta, \forall x, t \leq n \quad \left|\langle x, \hat{\theta}_t - \theta \rangle\right| \leq c\sqrt{d \, \|x\|^2_{G_t^{-1}} \log\left(\frac{n}{\delta}\right)}$$ (Confidence)

$c > 0$ is a universal constant and $\|x\|^2_{G_t^{-1}} = x^\top G_t^{-1} x$

OFUL Algorithm (Abbasi-Yadkori, Pál, Szepesvári) chooses:

$$A_t = \arg\max_{a \in \mathcal{A}} \langle x, \hat{\theta}_t \rangle + c\sqrt{d \, \|x\|^2_{G_t^{-1}} \log\left(\frac{n}{\delta}\right)}$$

Regret bounds for optimistic algorithm

Theorem: (Abbasi-Yadkori, Pál, Szepesvári)

The regret of OFUL is bounded by

$$R_n = O(d\sqrt{n \, \mathrm{polylog}(n)})$$

Almost matches lower bound by Rusmevichientong and Tsitsiklis

(which is $\Omega(d\sqrt{n})$)

Worst-case bound **obscures instance-dependent** structure

Lower bound (Lattimore & Sz, '16)

For any consistent strategy:

$$\limsup_{n\to\infty} \log(n)\, \|x\|^2_{\bar{G}_n^{-1}} \le \frac{\Delta_x^2}{2} \text{ for all } x \in \mathcal{A}\,,$$

where $\bar{G}_n = \mathbb{E}\left[\sum_{t=1}^n A_t A_t^\top\right]$. Furthermore,

Lower bound (Lattimore & Sz, '16)

For any consistent strategy:

$$\limsup_{n\to\infty} \log(n)\, \|x\|^2_{\bar{G}_n^{-1}} \le \frac{\Delta_x^2}{2} \text{ for all } x \in \mathcal{A}\,,$$

where $\bar{G}_n = \mathbb{E}\left[\sum_{t=1}^n A_t A_t^\top\right]$. Furthermore,

$$\limsup_{n\to\infty} \frac{R_n}{\log(n)} \ge c(\theta, \mathcal{A})$$

where

$$c(\theta, \mathcal{A}) = \inf_{\alpha \in [0,\infty)^k} \sum_{x\in\mathcal{A}} \alpha(x)\Delta_x \qquad \text{subject to}$$

$$\|x\|^2_{H_\alpha^{-1}} \le \frac{\Delta_x^2}{2}\,, \qquad H_\alpha = \sum_{x\in\mathcal{A}} \alpha(x) x x^\top$$

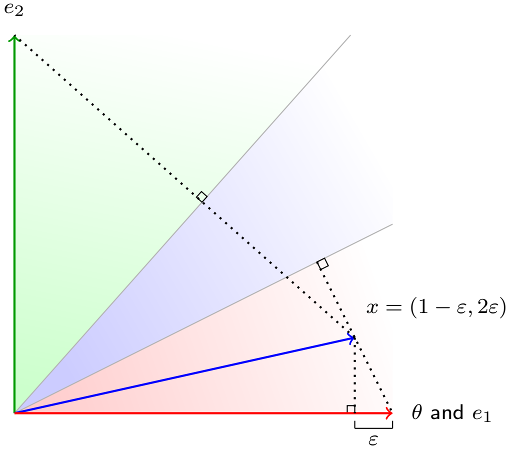
Upper bound (Lattimore & Sz, '16)

There exists a strategy such that

$$\limsup_{n\to\infty} \frac{R_n}{\log(n)} \leq c(\theta, \mathcal{A})$$

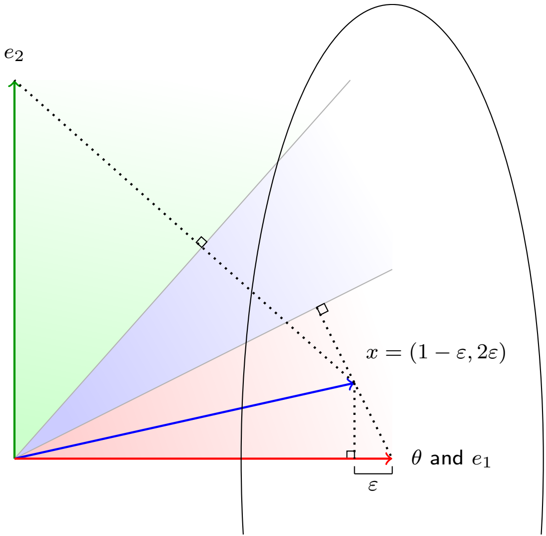
Failure of optimism

$d = 2$ and $k = 3$ and $\mathcal{A} = \{e_1, e_2, x\}$ and $\theta = e_1$



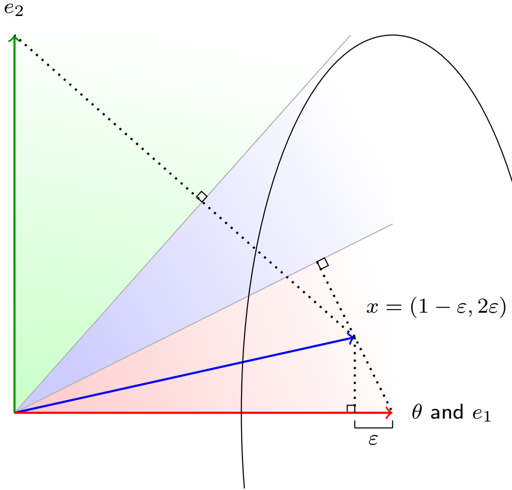
Failure of optimism

$d = 2$ and $k = 3$ and $\mathcal{A} = \{e_1, e_2, x\}$ and $\theta = e_1$



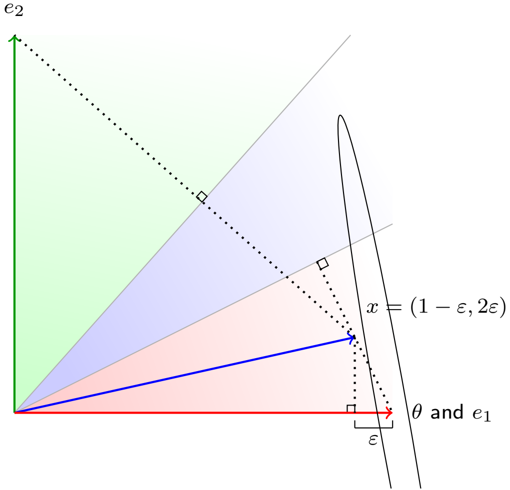
Failure of optimism

$d = 2$ and $k = 3$ and $\mathcal{A} = \{e_1, e_2, x\}$ and $\theta = e_1$



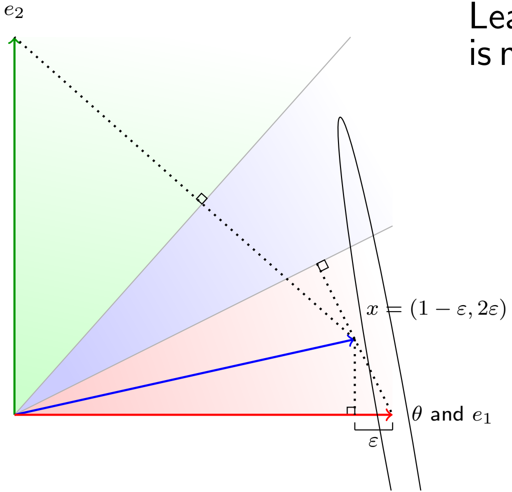
Failure of optimism

$d = 2$ and $k = 3$ and $\mathcal{A} = \{e_1, e_2, x\}$ and $\theta = e_1$



Failure of optimism

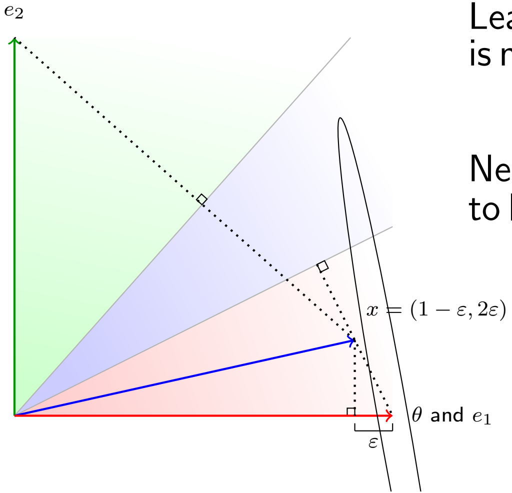
$d = 2$ and $k = 3$ and $\mathcal{A} = \{e_1, e_2, x\}$ and $\theta = e_1$



Learning is very slow once e_2 is not played

Failure of optimism

$d = 2$ and $k = 3$ and $\mathcal{A} = \{e_1, e_2, x\}$ and $\theta = e_1$

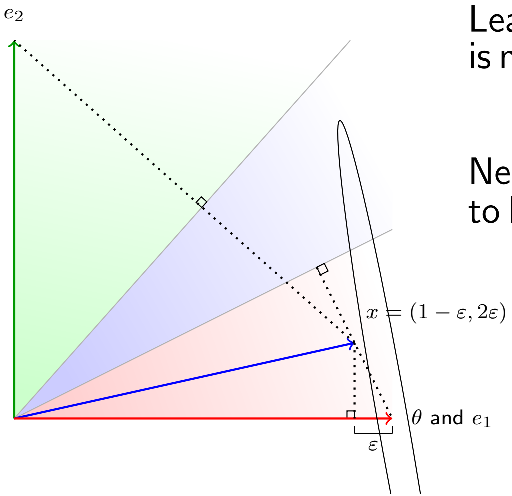


Learning is very slow once e_2 is not played

Need $\Omega(\log(n)/\varepsilon^2)$ plays of x to learn it is sub-optimal

Failure of optimism

$d = 2$ and $k = 3$ and $\mathcal{A} = \{e_1, e_2, x\}$ and $\theta = e_1$



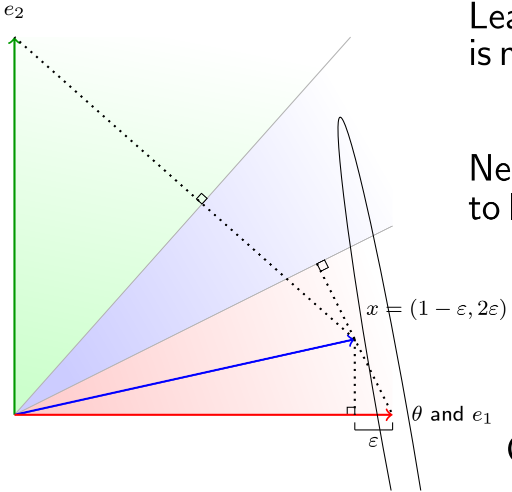
Learning is very slow once e_2 is not played

Need $\Omega(\log(n)/\varepsilon^2)$ plays of x to learn it is sub-optimal

Regret is $\Omega(\log(n)/\varepsilon)$

Failure of optimism

$d = 2$ and $k = 3$ and $\mathcal{A} = \{e_1, e_2, x\}$ and $\theta = e_1$



Learning is very slow once e_2 is not played

Need $\Omega(\log(n)/\varepsilon^2)$ plays of x to learn it is sub-optimal

Regret is $\Omega(\log(n)/\varepsilon)$

Optimal regret is $O(\log(n))$

Failure of optimism

Optimism fails because it never chooses actions that it has shown (statistically significantly) to be sub-optimal

But these actions should still be taken if the information gain about **other actions** is large relative to the regret

Phenomena not observed in the orthogonal case because there is no generalisation

Failure of Thompson sampling

Define prior P on $\theta \in \mathbb{R}^d$

In each round t:

1. Calculate posterior $P_t = P(\theta | A_1, Y_1, \ldots, A_{t-1}, Y_{t-1})$
2. Sample $\tilde{\theta}_t \sim P_t$
3. Choose $A_t = \arg\max_{x \in \mathcal{A}} \langle x, \tilde{\theta}_t \rangle$

Theorem (Agrawal & Goyal) $R_n = O\left(d^{3/2}\sqrt{n \operatorname{polylog}(n)}\right)$

Failure of Thompson sampling

Define prior P on $\theta \in \mathbb{R}^d$

In each round t:

1. Calculate posterior $P_t = P(\theta | A_1, Y_1, \dots, A_{t-1}, Y_{t-1})$
2. Sample $\tilde{\theta}_t \sim P_t$
3. Choose $A_t = \arg\max_{x \in \mathcal{A}} \langle x, \tilde{\theta}_t \rangle$

Theorem (Agrawal & Goyal) $R_n = O\left(d^{3/2}\sqrt{n \operatorname{polylog}(n)}\right)$

Suffers from exactly the same problem as optimism!

Chooses statistically sub-optimal actions with vanishingly small probability

Brace yourselves for the **optimal** algorithm

A three-phase algorithm

Phase 1 (exploration)

Find a barycentric spanner $B \subseteq \mathcal{A}$

Choose each $x \in B$ exactly $\lceil \log^{1/2}(n) \rceil$ times

Phase 2 (anomaly detection)

Compute $\hat{\theta} = G_t^{-1} \sum_{s=1}^t A_s Y_s$ and $\hat{\Delta}_x = \max_{y \in \mathcal{A}} \langle y - x, \hat{\theta} \rangle$

Solve $S = \underset{S \in [0,\infty]^k}{\arg\min} \displaystyle\sum_{x \in \mathcal{A}} S_x \hat{\Delta}_x$ subject to

$$\text{For all } x, \quad \|x\|^2_{H_S^{-1}} \leq \frac{\hat{\Delta}_x^2}{2(1 + o(1)) \log(n)} \qquad H_S = \sum_{x \in \mathcal{A}} S_x x x^\top$$

Loop as long as new observations are not too inconsistent with $\hat{\Delta}$, choosing arms x played less than S_x times

Phase 3 (recovery) Switch to UCB

Key elements of proof

An optimisation approach to learning

Key elements of proof

An optimisation approach to learning

Improved concentration guarantees

$$G_t = \sum_{s=1}^{t-1} A_s A_s^\top$$

$$\hat{\theta}_t = G_t^{-1} \sum_{s=1}^{t-1} A_s Y_s$$

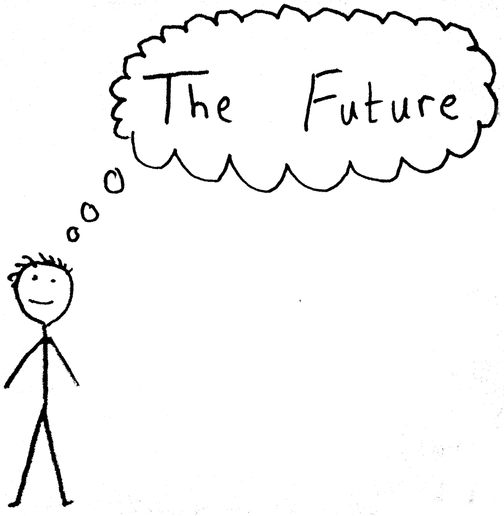
with probability at least $1 - \delta$ it holds for all $x \in \mathcal{A}$ and $t \leq n$

$$\left|\langle \hat{\theta}_t - \theta, x \rangle\right| \leq \sqrt{2 \left\|x\right\|^2_{G_t^{-1}} \left(c \cdot d \cdot \log\log(n) + \log\left(\frac{1}{\delta}\right)\right)}$$

correct constant

probably tight by
law of iterated logarithm

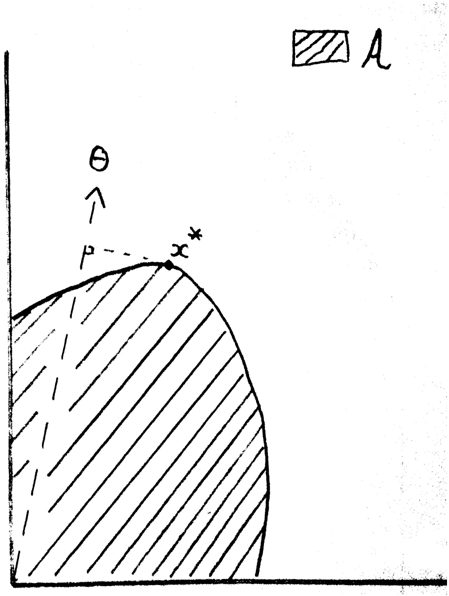
typically $\Theta(\log(n))$


The Future

Practical optimal algorithms

Finite-time guarantees

Infinite action sets



Shape-dependent regret in continuous case

A (very) few results known in adversarial setting

Curvature may play a role as it does in experts setting (Huang, Lattimore, György & Szepesvári, NIPS 2016)

Global information also important

Computation becomes interesting

The contextual case

What happens when the action-set is changing?

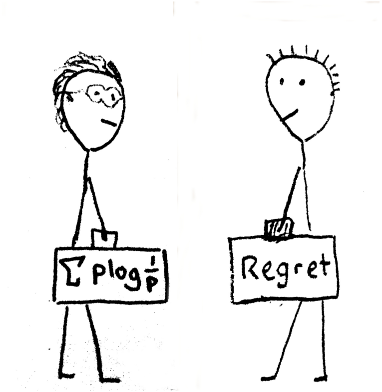
Optimisation problem should depend on future action sets. Seems complicated

Information/regret trade-off still present

Trading regret for information

We don't know how to do this in a generic way

(lots of interesting attempts though)



Summary

Optimism and Thompson sampling can fail badly when generalisation is possible

Concerning because both are widely used

(linear bandits, contextual bandits, reinforcement learning,...)

We need new tools (information-theoretic or optimisation approaches, perhaps)

References

Lattimore & Szepesvári. The End of Optimism? An Asymptotic Analysis of Finite-Armed Linear Bandits. 2016

Abbasi-Yadkori, Pál & Szepesvári. Improved Algorithms for Linear Stochastic Bandits. 2012

Agrawal & Goyal. Thompson Sampling for Contextual Bandits with Linear Payoffs. 2013

Rusmevichientong & Tsitsiklis. Linearly Parameterized Bandits. 2008

Dani, Hayes & Kakade. Stochastic Linear Optimization under Bandit Feedback. 2008

Huang, Lattimore, György & Szepesvári. Following the Leader and Fast Rates in Linear Prediction: Curved Constraint Sets and Other Regularities. 2016