Bandit Algorithms

Tor Lattimore & Csaba Szepesvári

Outline

1. From Contextual to Linear Bandits
2. Stochastic Linear Bandits
3. Confidence Bounds for Least-Squares Estimators
4. Improved Regret for Fixed, Finite Action Sets

Outline

Stochastic Contextual Bandits

Set of contexts, $\mathcal{C}$, set of actions $[K]$; distributions $(P_{c,a})$.

Interaction

For rounds $t = 1, 2, 3, \ldots$:

1. Context $C_t \in \mathcal{C}$ is revealed to the learner.
2. Based on its past observations (including C_t), the learner chooses an action $A_t \in [K]$. The chosen action is sent to the environment.
3. The environment sends the reward $X_t \sim P_{C_t,A_t}$ to the learner.

Regret Definition

Definition: Expected reward for action a under context c:

$$r(c, a) = \int x \, P_{c,a}(dx) \, .$$

Regret:

$$R_n = \mathbb{E}\left[\sum_{t=1}^{n} \max_{a \in [K]} r(C_t, a) - \sum_{t=1}^{n} X_t\right] \, .$$

Poor Man's Contextual Bandit Algorithm

Assumption: C is finite.

Idea: Assign a bandit to each context.

Worst-case regret: $R_n = \Theta(\sqrt{nMK})$, where $M = |\mathrm{C}|$.

Problem: M (and K) can be very large.

How to save this? Assume structure.

Linear Models

Assumption:

$$r(c, a) = \langle \psi(c, a), \theta_* \rangle \,, \qquad \forall (c, a) \in \mathrm{C} \times [K] \,.$$

where $\psi : \mathrm{C} \times [K] \to \mathbb{R}^d$, $\theta_* \in \mathbb{R}^d$.

- ψ: **feature map**;
- $\mathcal{H}_\psi \doteq \mathrm{span}(\, \psi(c, k) \; : \; c \in \mathrm{C}, k \in [K] \,) \subset \mathbb{R}^d$: **feature space**;
- $\mathcal{S}_\psi \doteq \{r_\theta \; : \; r_\theta : \mathrm{C} \times [K] \to \mathbb{R}, \theta \in \mathbb{R}^d\}$, where
  $r_\theta(c, k) = \langle \psi(c, a), \theta \rangle$:
  space of linear reward functions.

Note: RKHS let us deal with $d = \infty$ (or d very large).

Outline

Choosing $\psi \Rightarrow$ betting on smoothness of r

Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$, $\|\cdot\|_*$ its dual (e.g., both 2-norms).
Hölder's inequality:

$$|r(c,a) - r(c',a')| \leq \|\theta_*\| \, \|\psi(c,a) - \psi(c',a')\|_* \, .$$

$r = \langle \psi, \theta_* \rangle \Rightarrow r$ is $(\rho, \|\theta_*\|)$-smooth:

$$|r(z) - r(z')| \leq \|\theta_*\| \, \rho(z,z')$$

where $z = (c,a), z' = (c',a')$, $\rho(z,z') = \|\psi(z) - \psi(z')\|_*$.

Choice of ψ + preference for small $\|\theta_*\| \Leftrightarrow$
preference for ρ-smooth r.

Influence of ψ is the largest when $d = +\infty$.

Sparsity

Concern: Is $r \in \mathcal{S}_\psi$ really true?

Thinking of free lunches: Can't we just add a lot of features to make sure $r \in \mathcal{S}_\psi$?

- Feature i unused: $\theta_{*,i} = 0$;
- Small $\|\theta_*\|_0 = \sum_{i=1}^{d} \mathbb{I}_{\{\theta_{*i} \neq 0\}}$;
- '0-norm'.

Sparsity

Outline

Features Are All What You Need

Given $C_1, A_1, \ldots, C_t, A_t$, the reward X_t in round t satisfies

$$\mathbb{E}\left[X_t | C_1, A_1, \ldots, C_t, A_t\right] = \langle \psi(C_t, A_t), \theta_* \rangle ,$$

for some known ψ and unknown θ_*.

$$\Leftrightarrow$$

At the beginning of any round t, observe action set $\mathcal{A}_t \subset \mathbb{R}^d$. If $A_t \in \mathcal{A}_t$,

$$\mathbb{E}\left[X_t | \mathcal{A}_1, A_1, \ldots, \mathcal{A}_t, A_t\right] = \langle A_t, \theta_* \rangle$$

with some unknown θ_*.

Why? Let $$\mathcal{A}_t = \{\psi(C_t, a) \,:\, a \in [K]\} .$$

Stochastic 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

$$\mathbb{E}\left[X_t \middle| \mathcal{A}_1, A_1, \dots, \mathcal{A}_t, A_t\right] = \langle A_t, \theta_* \rangle$$

with some unknown θ_*.

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.

Additional assumptions: *(i)* $\mathcal{A}_1, \dots, \mathcal{A}_n$ is any fixed sequence; *(ii)* $X_t - \langle A_t, \theta_* \rangle$ is light tailed, given $\mathcal{A}_1, A_1, \dots, \mathcal{A}_t, A_t$.

Finite-armed bandits

Case (a): $\mathcal{A}_t$ has always the same number of vectors in it:

"finite-armed stochastic contextual bandit".

Case (b): Also, $\mathcal{A}_t$ does not change, or $\mathcal{A}_t = \{a_1, \ldots, a_K\}$:

"finite-armed stochastic linear bandit".

Case (c): If the vectors in $\mathcal{A}_t$ are also orthogonal to each other:

"finite-armed stochastic bandit".

Difference between cases (c) and (b):

- Case (c): Learn about mean of arm $i \Leftrightarrow$ Choose action i;
- Case (b): Learn about mean of arm $i \Leftrightarrow$ Choose action j s.t. $\langle x_j, x_i \rangle \neq 0$.

Outline

Once an Optimist, Always an Optimist

Optimism in the Face of Uncertainty Principle:

"Choose the best action in the best environment amongst the plausible ones."

Environment $\Leftrightarrow \theta \in \mathbb{R}^d$.

Plausible environments:

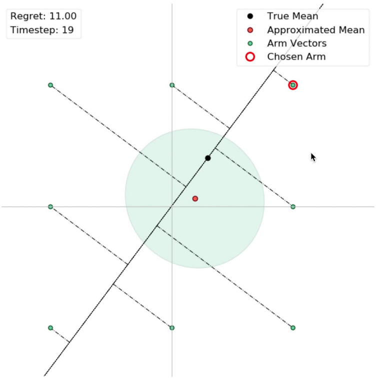
$\mathcal{C}_t \subset \mathbb{R}^d$ s.t. $\mathbb{P}\left(\theta_* \notin \mathcal{C}_t\right) \sim 1/t$.

Best environment:

$\tilde{\theta}_t = \mathrm{argmax}_{\theta \in \mathcal{C}_t} \max_{a \in \mathcal{A}} \langle a, \theta \rangle$.

Best action:

$\mathrm{argmax}_{a \in \mathcal{A}} \langle a, \hat{\theta}_t \rangle$.



Choosing the Confidence Set

Say, reward in round t is X_t, action in round t is $A_t \in \mathbb{R}^d$:

$$X_t = \langle A_t, \theta_* \rangle + \eta_t \,,$$

η_t is noise.
Regularized least-squares estimator: $\hat{\theta}_t = V_t^{-1} \sum_{s=1}^t A_s X_s$,

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

Choice of $\mathcal{C}_t$:

$$\mathcal{C}_t \subset \mathcal{E}_t \doteq \left\{ \theta \in \mathbb{R}^d \,:\, \|\theta - \hat{\theta}_{t-1}\|_{V_{t-1}}^2 \leq \beta_t \right\} \,.$$

Here, $(\beta_t)_t$ decreasing, $\beta_t \geq 1$, for A positive definite, $\|x\|_A^2 = x^\top A x$.

LinUCB

Choose $\mathcal{C}_t = \mathcal{E}_t$ with suitable $(\beta_t)_t$ and let

$$A_t = \operatorname*{argmax}_{a \in \mathcal{A}} \max_{\theta \in \mathcal{C}_t} \langle a, \theta \rangle \,.$$

Then,

$$A_t = \operatorname*{argmax}_{a} \langle a, \hat{\theta}_t \rangle + \sqrt{\beta_t}\, \|a\|_{V_{t-1}^{-1}} \,.$$

LinUCB (a.k.a. LinRel, OFUL, ConfEllips, ...)

Outline

Regret Analysis

Assumptions:

1. *Bounded scalar mean reward:* $|\langle a, \theta_* \rangle| \leq 1$ for any $a \in \cup_t \mathcal{A}_t$.
2. *Bounded actions:* for any $a \in \cup_t \mathcal{A}_t$, $\|a\|_2 \leq L$.
3. *Honest confidence intervals:* There exists a $\delta \in (0, 1)$ such that with probability $1 - \delta$, for all $t \in [n]$, $\theta_* \in \mathcal{C}_t$ where $\mathcal{C}_t$ satisfies $\mathcal{C}_t \subset \mathcal{E}_t$ with $(\beta_t)_t \geq 1$, decreasing as on the previous slide.

Theorem (LinUCB Regret)

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

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

Proof

Assume $\theta_* \in \mathcal{C}_t$, $t \in [n]$. Let $A_t^* \doteq \mathrm{argmax}_{a \in \mathcal{A}_t} \langle a, \theta_* \rangle$, $r_t = \langle A_t^* - A_t, \theta_* \rangle$ and let $\tilde{\theta}_t \in \mathcal{C}_t$ s.t. $\langle A_t, \tilde{\theta}_t \rangle = \mathrm{UCB}_t(A_t)$.
From $\theta_* \in \mathcal{C}_t$ and the definition of LinUCB,

$$\langle A_t^*, \theta_* \rangle \le \mathrm{UCB}_t(A_t^*) \le \mathrm{UCB}_t(A_t) = \langle A_t, \tilde{\theta}_t \rangle \,.$$

Then,

$$r_t \le \langle A_t, \tilde{\theta}_t - \theta_* \rangle \le \|A_t\|_{V_{t-1}^{-1}} \|\tilde{\theta}_t - \theta_*\|_{V_{t-1}} \le 2 \|A_t\|_{V_{t-1}^{-1}} \beta_t \,.$$

From $\langle a, \theta_* \rangle \le 1$, $r_t \le 2$. This combined with $\beta_n \ge \max\{1, \beta_t\}$ gives

$$r_t \le 2 \wedge 2\sqrt{\beta_t} \|A_t\|_{V_{t-1}^{-1}} \le 2\sqrt{\beta_n}(1 \wedge \|A_t\|_{V_{t-1}^{-1}}) \,.$$

Jensen's inequality shows that

$$\hat{R}_n^{\text{pseudo}} = \sum_{t=1}^{n} r_t \le \sqrt{n \sum_{t=1}^{n} r_t^2} \le 2\sqrt{n\beta_n \sum_{t=1}^{n} (1 \wedge \|A_t\|_{V_{t-1}^{-1}}^2)} \,.$$

Lemma

Lemma

Let $x_1, \ldots, x_n \in \mathbb{R}^d$, $V_t = V_0 + \sum_{s=1}^t x_s x_s^\top$, $t \in [n]$, $v_0 = \mathrm{trace}(V_0)$ *and* $L \geq \max_t \|x_t\|_2$. *Then,*

$$\sum_{t=1}^n \left(1 \wedge \|x_t\|^2_{V_{t-1}^{-1}}\right) \leq 2 \log\left(\frac{\det V_n}{\det V_0}\right) \leq d \log\left(\frac{v_0 + nL^2}{d \det^{1/d}(V_0)}\right) .$$

Outline

LinUCB and Finite-Armed Bandits

Recall that if $\mathcal{A}_t = \{e_1, \ldots, e_d\}$, we get back the finite armed bandits.
LinUCB:

$$A_t = \operatorname*{argmax}_a \langle a, \hat{\theta}_t \rangle + \sqrt{\beta_t} \, \|a\|_{V_{t-1}^{-1}} \, .$$

If we set $\lambda = 0$, $\langle e_i, \hat{\theta}_t \rangle = \hat{\mu}_{i,t}$: empirical mean, $V_{t-1} = \mathrm{diag}(T_1(t-1), \ldots, T_K(t-1))$. Hence,

$$\sqrt{\beta_t} \, \|e_i\|_{V_{t-1}^{-1}} = \sqrt{\frac{\beta_t}{T_i(t-1)}} \, ,$$

and

$$A_t = \operatorname*{argmax}_a \langle a, \hat{\theta}_t \rangle + \sqrt{\frac{\beta_t}{T_i(t-1)}} \, ,$$

We recover UCB when $\beta_t = 2 \log(\cdot)$.

History

- Abe and Long (1999) introduced stochastic linear bandits into machine learning literature.
- Auer (2002) was the first to consider optimism for linear bandits (LinRel, SupLinRel). Main restriction: $|\mathcal{A}_t| < +\infty$.
- Confidence ellipsoids: Dani et al. (2008) (ConfidenceBall$_2$), Rusmevichientong and Tsitsiklis (2010) (Uncertainty Ellipsoid Policy), Abbasi-Yadkori et al. (2011) (OFUL).
- The name LinUCB comes from Chu et al. (2011).
- Alternative routes:
  - Explore then commit for action sets with smooth boundary. Abbasi-Yadkori et al. (2009); Abbasi-Yadkori (2009); Rusmevichientong and Tsitsiklis (2010).
  - Phased elimination
  - Thompson sampling

Extensions of Linear Bandits

- **Generalized linear model** (Filippi et al., 2009):

$$X_t = g^{-1}(\langle A_t, \theta_* \rangle + \eta)\,, \tag{1}$$

  where $g : \mathbb{R} \to \mathbb{R}$ is called the **link function**.
  Common choice: $g(p) = \log(p/(1-p))$ when
  $g^{-1}(x) = 1/(1+\exp(-x))$ (sigmoid).
- Spectral bandits: Spectral Eliminator Valko et al. (2014).
- Kernelised UCB: Valko et al. (2013).
- (Nonlinear) Structured bandits: $r : \mathcal{C} \times [K] \to [0,1]$ belongs to some known set (Anantharam et al., 1987; Russo and Roy, 2013; Lattimore and Munos, 2014).

Outline

Setting

1. *Subgaussian rewards:* The reward is $X_t = \langle A_t, \theta_* \rangle + \eta_t$, where η_t is conditionally 1-subgaussian ($\eta_t | \mathcal{F}_{t-1} \sim \mathrm{subG}(1)$):

$$\mathbb{E}[\exp(\lambda \eta_t) | \mathcal{F}_{t-1}] \leq \exp(\lambda^2/2) \qquad \text{almost surely for all } \lambda \in \mathbb{R}\,,$$

   where $\mathcal{F}_t = \sigma(A_1, \eta_1, \ldots, A_{t-1}, \eta_{t-1}, A_t)$.

2. *Bounded parameter vector:* $\|\theta_*\|_2 \leq S$ with $S > 0$ known.

Least Squares: Recap

Linear model:

$$X_t = \langle A_t, \theta_* \rangle + \eta_t \,.$$

Regularized squared loss:

$$L_t(\theta) = \sum_{s=1}^{t} (X_s - \langle A_s, \theta \rangle)^2 + \lambda \, \|\theta\|_2^2 \,,$$

Least squares estimate: $\hat{\theta}_t = \operatorname{argmin}_\theta L_t(\theta)$:

$$\hat{\theta}_t = V_t(\lambda)^{-1} \sum_{s=1}^{t} X_s A_s \quad \text{with } V_t(\lambda) = \lambda I + \sum_{s=1}^{t} A_s A_s^\top \,. \tag{2}$$

Abbreviation: $V_t = V_t(0)$.

Main difficulty

The actions $(A_s)_{s<t}$ are neither fixed nor independent, but are intricately correlated via the rewards $(X_s)_{s<t}$.

Outline

Fixed Design

1. *Nonsingular Grammian:* $\lambda = 0$ and V_t is invertible.
2. *Independent subgaussian noise:* $(\eta_s)_s$ are independent and 1-subgaussian.
3. *Fixed design:* $A_1, \ldots, A_t$ are deterministically chosen without the knowledge of $X_1, \ldots, X_t$.

Notation: $A_1, \ldots, A_t$ replaced by $a_1, \ldots, a_t$, so

$$V_t = \sum_{s=1}^{t} a_s a_s^\top \qquad \text{and} \qquad \hat{\theta}_t = V_t^{-1} \sum_{s=1}^{t} a_s X_s \,.$$

Note: $(a_s)_{s=1}^t$ must span $\mathbb{R}^d$ for $\exists V_t^{-1} \Rightarrow t \geq d$.

First Steps

Fix $x \in \mathbb{R}^d$. From $X_s = a_s^\top \theta_* + \eta_s$,

$$\hat{\theta}_t = V_t^{-1} \sum_{s=1}^{t} a_s X_s = \cancel{V_t^{-1}} \cancel{V_t}\, \theta_* + V_t^{-1} \sum_{s=1}^{t} a_s \eta_s \,,$$

so

$$\langle x, \hat{\theta}_t - \theta_* \rangle = \sum_{s=1}^{t} \langle x, V_t^{-1} a_s \rangle \, \eta_s \,.$$

Since $(\eta_s)_s$ are independent and 1-subgaussian: With probability $1 - \delta$,

$$\langle x, \hat{\theta}_t - \theta_* \rangle < \sqrt{2 \sum_{s=1}^{t} \langle x, V_t^{-1} a_s \rangle^2 \log\left(\frac{1}{\delta}\right)} = \sqrt{2 \|x\|^2_{V_t^{-1}} \log\left(\frac{1}{\delta}\right)} \,.$$

Bounding $\|\hat{\theta}_t - \theta^*\|_{V_t}$

For $x \in \mathbb{R}^d$ fixed, with probability $1 - \delta$,

$$\langle x, \hat{\theta}_t - \theta_* \rangle < \sqrt{2\|x\|^2_{V_t^{-1}} \log\left(\tfrac{1}{\delta}\right)}\,. \qquad (*)$$

We have $\|u\|^2_{V_t} = (V_t u)^\top u$.

Idea 1: Apply (*) to $X = V_t(\hat{\theta}_t - \theta_*)$..
Problem: (*) holds for non-random x only!

Idea 2: Take finite $\mathcal{C}_\varepsilon$ s.t. $X \approx_\varepsilon x$ for some $x \in \mathcal{C}_\varepsilon$. Make (*) hold for each $x \in \mathcal{C}_\varepsilon$, combine.
Refinement: Let $X = V_t^{1/2}(\hat{\theta}_t - \theta_*)/\|\hat{\theta}_t - \theta_*\|_{V_t}$.
Then $X \in S^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ and we can choose a finite $\mathcal{C}_\varepsilon \subset S^{d-1}$ s.t. for any $u \in S^{d-1}$, $\exists x \in \mathcal{C}_\varepsilon$ with $\|u - x\| \le \varepsilon$.

Bounding $\|\hat{\theta}_t - \theta^*\|_{V_t}$: Part II.

Let

$$X = \frac{V_t^{1/2}(\hat{\theta}_t - \theta_*)}{\|\hat{\theta}_t - \theta_*\|_{V_t}}, \qquad X^* = \operatorname*{argmin}_{x \in \mathcal{C}_\varepsilon} \|x - X\| \ .$$

Then, with probability $1 - \delta$,

$$\begin{aligned}
\|\hat{\theta}_t - \theta_*\|_{V_t} &= \langle X, V_t^{1/2}(\hat{\theta}_t - \theta_*)\rangle \\
&= \langle X - X^*, V_t^{1/2}(\hat{\theta}_t - \theta_*)\rangle + \langle V_t^{1/2} X^*, \hat{\theta}_t - \theta_*\rangle \\
&\leq \varepsilon \|\hat{\theta}_t - \theta_*\|_{V_t} + \sqrt{2\cancel{\|V_t^{1/2} X^*\|^2_{V_t^{-1}}} \log\left(\frac{|\mathcal{C}_\varepsilon|}{\delta}\right)} \,.
\end{aligned}$$

or

$$\|\hat{\theta}_t - \theta_*\|_{V_t} \leq \frac{1}{1 - \varepsilon} \sqrt{2 \log\left(\frac{|\mathcal{C}_\varepsilon|}{\delta}\right)} \,.$$

We can choose $\mathcal{C}_\varepsilon$ so that $|\mathcal{C}_\varepsilon| \leq (5/\varepsilon)^d$:

$$\|\hat{\theta}_t - \theta_*\|_{V_t} < 2\sqrt{2\left(d \log(10) + \log\left(\frac{1}{\delta}\right)\right)} \,.$$

Outline

Bounding $\|\hat{\theta}_t - \theta^*\|_{V_t}$: Sequential Design

$X_s = \langle A_s, \theta_* \rangle + \eta_s$,
$\eta_s | \mathcal{F}_{s-1} \sim \mathrm{subG}(1)$ for $\mathcal{F}_s = \sigma(A_1, \eta_1, \ldots, A_s, \eta_s)$.

Previous bound exploited $A_1, \ldots, A_t$ fixed, non-random. Known as:

fixed design

When $A_1, \ldots, A_t$ is i.i.d., we have a

random design

Bandits: A_s is chosen based on $A_1, X_1, \ldots, A_{s-1}, X_{s-1}$!

sequential design

How to bound $\|\hat{\theta}_t - \theta^*\|_{V_t}$ in this case?

Bounding $\|\hat{\theta}_t - \theta^*\|_{V_t}$: Sequential Design

- Linearization trick
- Vector Chernoff
- Laplace method

A Start: Linearization & Vector Chernoff

Let $S_t = \sum_{s=1}^t \eta_s A_s$. "Linearization" of the quadratic:

$$\frac{1}{2}\|\hat{\theta}_t - \theta_*\|_{V_t}^2 = \frac{1}{2}\|S_t\|_{V_t^{-1}}^2 = \max_{x \in \mathbb{R}^d} \langle x, S_t \rangle - \frac{1}{2}\|x\|_{V_t}^2 \ .$$

Let

$$M_t(x) = \exp\left(\langle x, S_t \rangle - \frac{1}{2}\|x\|_{V_t}^2\right) \ .$$

One can show that $\mathbb{E}\left[M_t(x)\right] \leq 1$ for any $x \in \mathbb{R}^d$. Chernoff's method:

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{2}\|\hat{\theta}_t - \theta_*\|_{V_t}^2 \geq u\right) &= \mathbb{P}\left(\exp(\max_x \log M_t(x)) \geq \exp(u)\right) \\
&\leq \mathbb{E}\left[\exp(\max_x \log M_t(x))\right] \exp(-u) = \mathbb{E}\left[\max_x M_t(x))\right] \exp(-u) \ .
\end{aligned}$$

Can we control $\mathbb{E}\left[\max_x M_t(x)\right]$?

Controlling $\mathbb{E}\left[\max_x M_t(x)\right]$: Covering Argument

Recall:

$$M_t(x) = \exp\left(\langle x, S_t\rangle - \frac{1}{2}\|x\|^2_{V_t}\right) .$$

Let $\mathcal{C}_\varepsilon \subset \mathbb{R}^d$ be finite, to be chosen later,

$$X = \operatorname*{argmax}_{x\in\mathbb{R}^d} M_t(x), \quad Y = \operatorname*{argmin}_{y\in\mathcal{C}_\varepsilon} \|X - y\| .$$

Then,

$$\max_{x\in\mathbb{R}^d} M_t(x) = M_t(X) = M_t(X) - M_t(Y) + M_t(Y) \leq \varepsilon + \sum_{y\in\mathcal{C}_\varepsilon} M_t(y) .$$

Challenge: ensure $M_t(X) - M_t(Y) \leq \varepsilon$!

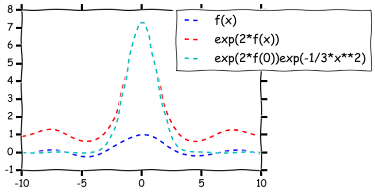
Laplace: One Step Back, Two Forward

The need to control $\mathbb{E}\left[\max_x M_t(x)\right]$ comes from the identity $\exp(\frac{1}{2}\|\hat{\theta}_t - \theta_*\|^2_{V_t}) = \max_x M_t(x)$.

Laplace: Integral of $\exp(sf(x))$ is dominated by $\exp(s \max_x f(x))$:

$$\int_a^b e^{sf(x)} dx \sim e^{sf(x_0)} \sqrt{\frac{2\pi}{s|f''(x_0)|}},$$

$$x_0 = \mathrm{argmax}_{x \in [a,b]} f(x) \in (a,b).$$



Idea: Replace $\max_x M_t(x)$ with $\int M_t(x)h(x)dx$ with h appropriate:

1. $\int M_t(x)h(x)dx \approx \max_x M_t(x)$ (in a way);
2. $\mathbb{E}\left[\int M_t(x)h(x)dx\right] = \int \mathbb{E}\left[M_t(x)\right] h(x)dx \leq 1$.

Choose $h(x)$ as density of $\mathcal{N}(0, H^{-1})$ for $H \succ 0$.

Step 2: Finishing

$$\int M_t(x)h(x)dx = \left(\frac{\det(H)}{\det(H+V)}\right)^{1/2} \exp\left(\frac{1}{2}\|S_t\|^2_{(H+V_t)^{-1}}\right) .$$

Choose $H = \lambda I$. Then, with probability $1 - e^{-u}$,

$$\tfrac{1}{2}\|S_t\|^2_{V_t^{-1}(\lambda)} < u + \frac{1}{2}\log\left(\frac{\det(V_t(\lambda))}{\lambda^d}\right) \qquad (**)$$

and from $\hat{\theta}_t - \theta_* = V_t^{-1}(\lambda)S_t - \lambda V_t^{-1}(\lambda)\theta_*$,

$$\begin{aligned}\|\hat{\theta}_t - \theta_*\|_{V_t(\lambda)} &\le \|V_t^{-1}(\lambda)S_t\|_{V_t(\lambda)} + \lambda\left\|V_t^{-1}(\lambda)\theta_*\right\|_{V_t(\lambda)} \\ &\le \|S_t\|_{V_t^{-1}(\lambda)} + \lambda^{1/2}\|\theta_*\| .\end{aligned}$$

Confidence Ellipsoid for Sequential Design

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

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

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

and

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

Theorem

$\mathcal{C}_{t+1}$ is a confidence set for θ_* at level $1-\delta$:

$$\mathbb{P}\left(\theta_* \in \mathcal{C}_{t+1}\right) \geq 1 - \delta\,.$$

Note: β_{t+1} is a function of $(A_s)_{s \leq t}$.

Freedman's Stopping Trick

We want $\theta_* \in \mathcal{C}_t$ hold with probability $1 - \delta$, simultaneously for all $1 \leq t \leq n$.

Can we avoid the union bound over time?

Freedman's Stopping Trick: II

Let

$$\mathcal{E}_t = \left\{ \|\hat{\theta}_t - \theta_*\|_{V_{t-1}(\lambda)} \geq \sqrt{\lambda} S + \sqrt{2u + \log\left(\frac{\det V_{t-1}(\lambda)}{\lambda^d}\right)} \right\}, t \in [n] .$$

Define $\tau \in [n]$ as follows: τ to be the smallest round index $t \in [n]$ such that $\mathcal{E}_t$ holds, or n when none of $\mathcal{E}_1, \ldots, \mathcal{E}_n$ hold.

Note: Because $\hat{\theta}_t$ and $V_{t-1}(\lambda)$ is a function of $H_{t-1} = (A_1, \eta_1, \ldots, A_{t-1}, \eta_{t-1})$, whether $\mathcal{E}_t$ holds can be decided based on H_{t-1}.

$\Rightarrow \tau$ is a $(H_t)_t$ stopping time $\Rightarrow \mathbb{E}\left[M_\tau(x)\right] \leq 1$ and also $\int \mathbb{E}\left[M_\tau(x)\right] h(x) dx \leq 1$ and thus $\mathbb{P}\left(\cup_{t \in [n]} \mathcal{E}_t\right) \leq \mathbb{P}\left(\mathcal{E}_\tau\right) \leq e^{-u}$. $n \to \infty$.

Corollary

$\mathbb{P}\left(\exists t \geq 0 \text{ such that } \theta_* \notin C_{t+1}\right) \leq \delta.$

Historical Remarks

- Presentation mostly follows Abbasi-Yadkori et al. (2011).
- Auer (2002); Chu et al. (2011) avoided the need to construct ellipsoidal confidence sets
- Previous ellipsoidal constructions by Dani et al. (2008) and Rusmevichientong and Tsitsiklis (2010) used covering arguments.
- The improvement that results from using Laplace's method as compared to the previous ellipsoidal constructions that are based on covering arguments is quite enormous.
- Laplace's method is also called the "Method of Mixtures" (Peña et al., 2008); its use goes back to the work of Robbins and Siegmund in the 1970s (Robbins and Siegmund, 1970, 1971).
- Freedman's Stopping is by Freedman (1975).

Outline

Regret for LinUCB: Final Steps

Previously we have seen, for $\beta_t \geq 1$, nondecreasing, using LinUCB with $V_0 = \lambda I$, w.p. $1-\delta$,

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

Now,

$$\beta_n = \sqrt{\lambda}S + \sqrt{2\log\left(\tfrac{1}{\delta}\right) + \log\left(\frac{\det V_{n-1}(\lambda)}{\lambda^d}\right)}$$
$$\leq \sqrt{\lambda}S + \sqrt{2\log\left(\tfrac{1}{\delta}\right) + d\log\left(\frac{\lambda d + nL^2}{\lambda d}\right)},$$

from $\|A_s\| \leq L$ and $\log\det V \leq d\log(\mathrm{trace}(V)/d)$.

$$\Rightarrow \hat{R}_n^{\text{pseudo}} \leq C_1 d\sqrt{n\log(n)} + C_2\sqrt{nd}\log(1/\delta) + C_3.$$

Summary

$$\hat{R}_n^{\text{pseudo}} \le C_1 d \sqrt{n \log(n)} + C_2 \sqrt{nd} \log(1/\delta) + C_3 .$$

- Optimism, confidence ellipsoid
- Getting the ellipsoid is tricky because the bandit algorithm makes $(A_s)_s$ and $(\eta_s)_s$ interdependent: “sequential design”.
- Hsu et al. (2012): Random design, fixed design
- Kernelization: One can directly kernelize the proof presented here. See Abbasi-Yadkori (2012). Gaussian Process Bandits effectively do the same (Srinivas et al., 2010).
- This presentation followed mostly Abbasi-Yadkori et al. (2011); Abbasi-Yadkori (2012).

Outline

Challenge!

The previous bound is $\tilde{O}(d\sqrt{n})$ even for $\mathcal{A} = \{e_1, \dots, e_d\}$ – finite-armed stochastic bandits.

Can we do better?

Setting

1. *Fixed finite action set:* The set of actions available in round t is $\mathcal{A} \subset \mathbb{R}^d$ and $|\mathcal{A}| = K$ for some natural number K.

2. *Subgaussian rewards:* The reward is $X_t = \langle \theta_*, A_t \rangle + \eta_t$ where η_t is conditionally 1-subgaussian: $\eta_t | \mathcal{F}_{t-1} \sim \mathrm{subG}(1)$, where $\mathcal{F}_t = \sigma(A_1, \eta_1, \ldots, A_{t-1}, \eta_t, A_t)$.

3. *Bounded mean rewards:* $\Delta_a = \max_{b \in \mathcal{A}} \langle \theta_*, b - a \rangle \leq 1$ for all $a \in \mathcal{A}$.

Key difference to previous setting:

Finite, fixed action set.

Avoiding Sequential Designs

Recall result for fixed design:

For $x \in \mathbb{R}^d$ fixed, with probability $1-\delta$,

$$\langle x, \hat{\theta}_t - \theta_* \rangle < \sqrt{2\|x\|^2_{V_t^{-1}} \log\left(\frac{1}{\delta}\right)} \,. \qquad (*)$$

Goal: Use this result! How?

Idea: Use a phased elimination algorithm!

- $\mathcal{A} = \mathcal{A}_1 \supset \mathcal{A}_2 \supset \mathcal{A}_3 \supset \ldots .$
- In phase ℓ, use actions in $\mathcal{A}_\ell$ to collect enough data to ensure that by the end of the phase, the data collected in the phase is sufficient to rule out all $\varepsilon_\ell \doteq 2^{-\ell}$-suboptimal actions.

Which action to use & how many times in phase ℓ?

How to Collect Data?

Recall: If $V_t = \sum_{s=1}^{t} a_s a_s^\top$, for any $x \in \mathbb{R}^d$.

$$\langle x, \hat{\theta}_t - \theta_* \rangle < \sqrt{2\|x\|^2_{V_t^{-1}} \log\left(\frac{1}{\delta}\right)}\,. \qquad (*)$$

If we need to know whether $\langle x, \hat{\theta}_t - \theta_* \rangle \leq 2^{-\ell}$, $x \in \mathcal{A}$, we better choose $a_1, a_2, \ldots, a_t$ so that

$$\max_{x \in \mathcal{A}} \|x\|^2_{V_t^{-1}} \qquad (*)$$

is minimized (and make t big enough).

$\Rightarrow$ Experimental design

Minimizing (*) is known as the G-optimal design problem.

Outline

G-optimal Design

Let $\pi : \mathcal{A} \to [0,1]$ be a distribution on $\mathcal{A}$: $\sum_{a\in\mathcal{A}} \pi(a) = 1$. Define

$$V(\pi) = \sum_{a\in\mathcal{A}} \pi(a) a a^\top \,, \qquad\qquad g(\pi) = \max_{a\in\mathcal{A}} \|a\|^2_{V(\pi)^{-1}} \,.$$

G-optimal design π^*:

$$g(\pi^*) = \min_{\pi} g(\pi) \,.$$

How to use this?

Using a Design π

Given a design π, for $a \in \mathrm{Supp}(\pi)$, set

$$n_a = \left\lceil \pi(a) \frac{g(\pi)}{\varepsilon^2} \log\left(\frac{1}{\delta}\right) \right\rceil .$$

Choose each action $a \in \mathrm{Supp}(\pi)$ exactly n_a times. Then:

$$V = \sum_{a \in \mathrm{Supp}(\pi)} n_a \, a a^\top \geq \frac{g(\pi)}{\varepsilon^2} \log\left(\frac{1}{\delta}\right) V(\pi) ,$$

and so for any $a \in \mathcal{A}$, w.p. $1 - \delta$,

$$\langle \hat{\theta} - \theta_*, a \rangle \leq \sqrt{\|a\|^2_{V^{-1}} \log\left(\frac{1}{\delta}\right)} \leq \varepsilon .$$

How big is n?

$$n = \sum_{a \in \mathrm{Supp}(\pi)} n_a = \sum_{a \in \mathrm{Supp}(\pi)} \left\lceil \pi(a) \frac{g(\pi)}{\varepsilon^2} \log\left(\frac{1}{\delta}\right) \right\rceil$$

$$\leq |\,\mathrm{Supp}(\pi)| + \frac{g(\pi)}{\varepsilon^2} \log\left(\frac{1}{\delta}\right) .$$

Bounding $g(\pi)$ and $|\operatorname{Supp}(\pi)|$

Theorem (Kiefer–Wolfowitz)

The following are equivalent:

1. π^* *is a minimizer of* g.
2. π^* *is a minimizer of* $f(\pi) = -\log\det V(\pi)$.
3. $g(\pi^*) = d$.

Note: Designs, minimizing f are known as D-optimal designs.

KW says that G-optimality is the same as D-optimality.

Combining this with John's Theorem for minimum-volume enclosing ellipsoids (John, 1948), we get $|\operatorname{Supp}(\pi)| \le d(d+3)/2$.

Outline

PEGOE Algorithm[1]

Input: $\mathcal{A} \subset \mathbb{R}^d$ and δ. Set $\mathcal{A}_1 = \mathcal{A}$, $\ell = 1$, $t = 1$.

1. Let $t_\ell = t$: current round. Find G-optimal design $\pi_\ell : \mathcal{A}_\ell \to [0, 1]$ that maximizes
$$\log \det V(\pi_\ell) \text{ subject to } \sum_{a \in \mathcal{A}_\ell} \pi_\ell(a) = 1$$
2. Let $\varepsilon_\ell = 2^{-\ell}$ and
$$N_\ell(a) = \left\lceil \frac{2\pi(a)}{\varepsilon_\ell^2} \log\left(\frac{K\ell(\ell+1)}{\delta}\right) \right\rceil \text{ and } N_\ell = \sum_{a \in \mathcal{A}_\ell} N_\ell(a)$$
3. Choose each action $a \in \mathcal{A}_\ell$ exactly $N_\ell(a)$ times
4. Calculate estimate: $\hat{\theta} = V_\ell^{-1} \sum_{t=t_\ell}^{t_\ell + N_\ell} A_t X_t$.
5. Eliminate poor arms:
$$\mathcal{A}_{\ell+1} = \left\{ a \in \mathcal{A}_\ell : \max_{b \in \mathcal{A}_\ell} \langle \hat{\theta}_\ell, b - a \rangle \geq 2\varepsilon_\ell \right\}.$$

[1] Phased Elimination with G-Optimal Exploration

The Regret of PEGOE

Theorem

With probability at least $1 - \delta$ *the pseudo-regret of PEGOE is at most:*

$$\hat{R}_n^{pseudo} \leq C\sqrt{nd\log\left(\frac{K\log(n)}{\delta}\right)},$$

where $C > 0$ *is a universal constant. If* $\delta = O(1/n)$, *then*

$$\mathbb{E}[R_n] \leq C\sqrt{nd\log(Kn)}$$

for appropriately chosen universal constant $C > 0$.

Summary and Historical Remarks

- Phased exploration allows one to use methods developed for fixed-design
- PEGOE: Exploration tuned to maximize information gain
- Finding an $(1+\varepsilon)$-optimal design is sufficient; convex problem
- This algorithm and analysis in this form is new.
- “Phased Elimination” is well known: Even-Dar et al. (2006) (pure exploration), Auer and Ortner (2010) (finite-armed bandits), Valko et al. (2014) (linear bandits, spanners instead of G-optimality).
- Finite, but changing action set: PEGOE cannot be applied! SupLinRel and SupLinUCB get the same bound (Auer, 2002; Chu et al., 2011). Sadly, these algorithms are very conservative..

Outline

General Setting

1. *(Sparse parameter)* There exist known constants M_0 and M_2 such that $\|\theta_*\|_0 \leq M_0$ and $\|\theta_*\|_2 \leq M_2$.
2. *(Bounded mean rewards):* $\langle a, \theta_* \rangle \leq 1$ for all $a \in \mathcal{A}_t$ and all rounds t.
3. *(Subgaussian noise):* The reward is $X_t = \langle A_t, \theta_* \rangle + \eta_t$ where $\eta_t | \mathcal{F}_{t-1} \sim \mathrm{subG}(1)$ for $\mathcal{F}_t = \sigma(A_1, \eta_1, \dots, A_t, \eta_t)$.

The Case of the Hypercube

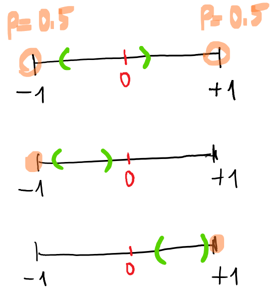
$$\mathcal{A} = [-1, 1]^d\,, \qquad \theta \doteq \theta_*\,, \qquad X_t = \langle A_t, \theta \rangle + \eta_t.$$

Assumptions:

1. *(Bounded mean rewards):* $\|\theta\|_1 \leq 1$, which ensures that $\langle a, \theta \rangle| \leq 1$ for all $a \in \mathcal{A}$.
2. *(Subgaussian noise):* The reward is $X_t = \langle A_t, \theta_* \rangle + \eta_t$ where $\eta_t | \mathcal{F}_{t-1} \sim \mathrm{subG}(1)$ for $\mathcal{F}_t = \sigma(A_1, \eta_1, \ldots, A_t, \eta_t)$.

Selective Explore-Then-Commit (SETC)

Recall: $\theta = \theta_*$.



For any $i \in [d]$ such that A_{ti} is randomized:

$$A_{ti}(A_t^\top \theta + \eta_t) = \theta_i + \underbrace{A_{ti} \sum_{j \neq i} A_{tj}\theta_j + A_{ti}\eta_t}_{\text{``noise''}} \; .$$

Regret of SETC

Theorem

There exists a universal constant $C > 0$ such that the regret of SETC satisfies:

$$R_n \leq 2 \|\theta\|_1 + C \sum_{i:\theta_i \neq 0} \frac{\log(n)}{|\theta_i|} \,.$$

Furthermore $R_n \leq C \|\theta\|_0 \sqrt{n \log(n)}$.

SETC adapts to $\|\theta\|_0$!

Outline

(General) LinUCB: Recap

GLinUCB

Choose $\mathcal{C}_t \subset \mathbb{R}^d$ and let

$$A_t = \operatorname*{argmax}_{a \in \mathcal{A}} \max_{\theta \in \mathcal{C}_t} \langle a, \theta \rangle \, .$$

Previous choice leads to regret $\tilde{O}(d\sqrt{n})$.

How to choose $\mathcal{C}_t$, knowing that $\|\theta_*\|_0 \leq p$,
so that the regret gets smaller?

Outline

Online Linear Regression (OLR)

Learner-environment interaction:

1. The environment chooses $X_t \in \mathbb{R}$ and $A_t \in \mathbb{R}^d$ in an arbitrary fashion.
2. The value of A_t is revealed to the learner (but not X_t).
3. The learner produces a real-valued prediction $\hat{X}_t$ in some way.
4. The environment reveals X_t to the learner and the loss is $(X_t - \hat{X}_t)^2$.

Goal: Compete with the total loss of the best linear predictors in some set $\Theta \subset \mathbb{R}^d$.

Regret against $\theta \in \Theta$:

$$\rho_n(\theta) = \sum_{t=1}^{n} (X_t - \hat{X}_t)^2 - \sum_{t=1}^{n} (X_t - \langle A_t, \theta \rangle)^2 .$$

From OLR to Confidence Sets

Let $\mathcal{L}$ be a learner that enjoys a regret guarantee $B_n = B_n(A_1, X_1, \ldots, A_n, X_n)$ relative to Θ: For any strategy of the environment,

$$\sup_{\theta \in \Theta} \rho_n(\theta) \leq B_n \,.$$

Combine

$$\rho_n(\theta) = \sum_{t=1}^{n} (X_t - \hat{X}_t)^2 - \sum_{t=1}^{n} (X_t - \langle A_t, \theta \rangle)^2 \,.$$

and $X_t = \langle A_t, \theta_* \rangle + \eta_t$ to get

$$\begin{aligned} Q_t \doteq \sum_{s=1}^{t} (\hat{X}_s - \langle A_s, \theta_* \rangle)^2 &= \rho_t(\theta_*) + 2 \sum_{s=1}^{t} \eta_s (\hat{X}_s - \langle A_s, \theta_* \rangle) \\ &\leq B_t + 2 \sum_{s=1}^{t} \eta_s (\hat{X}_s - \langle A_s, \theta_* \rangle) \,. \end{aligned}$$

From OLR to Confidence Sets: II.

$$Q_t \le B_t + 2Z_t\,, \qquad Z_t = \sum_{s=1}^{t} \eta_s(\hat{X}_s - \langle A_s, \theta_* \rangle)\,. \qquad (*)$$

Goal: Bound Z_t for $t \ge 0$.

$\hat{X}_t$, chosen by OLR learner $\mathcal{L}$, is $\mathcal{F}_{t-1}$-measurable,

$$(Z_t - Z_{t-1})|\mathcal{F}_{t-1} \sim \mathrm{subG}(\sigma_t)\,, \qquad \text{where } \sigma_t^2 = (\hat{X}_t - \langle A_t, \theta_* \rangle)^2\,.$$

Previous self-normalized bound (**): With probability $1-\delta$,

$$|Z_t| < \sqrt{(1+Q_t)\log\left(\frac{1+Q_t}{\delta^2}\right)}, t = 0, 1, \ldots\,.$$

Combining with (*), solve for Q_t:

$$Q_t \le \beta_t(\delta)\,, \quad \beta_t(\delta) = 1 + 2B_t + 32\log\left(\frac{\sqrt{8}+\sqrt{1+B_t}}{\delta}\right)\,.$$

OLR to Confidence Sets: III.

Theorem

Let $\delta \in (0,1)$ *and assume that* $\theta_* \in \Theta$ *and* $\sup_{\theta \in \Theta} \rho_t(\theta) \leq B_t$*. If*

$$\mathcal{C}_{t+1} = \left\{ \theta \in \mathbb{R}^d \,:\, \|\theta\|_2^2 + \sum_{s=1}^{t} (\hat{X}_s - \langle A_s, \theta \rangle)^2 \leq M_2^2 + \beta_t(\delta) \right\},$$

then $\mathbb{P}\left(\textit{exists } t \in \mathbb{N} \textit{ such that } \theta_* \notin \mathcal{C}_{t+1}\right) \leq \delta\,.$

Sparse LinUCB

Input OLR Learner $\mathcal{L}$, regret bound B_t, confidence parameter $\delta \in (0,1)$

for $t = 1, \dots, n$

  Receive action set $\mathcal{A}_t$

  Computer confidence set:

$$\mathcal{C}_t = \left\{ \theta \in \mathbb{R}^d \,:\, \|\theta\|_2^2 + \sum_{s=1}^{t-1} (\hat{X}_s - \langle A_s, \theta \rangle)^2 \leq M_2^2 + \beta_t(\delta) \right\}$$

  Calculate optimistic action

$$A_t = \operatorname*{argmax}_{a \in \mathcal{A}_t} \max_{\theta \in \mathcal{C}_t} \langle a, \theta \rangle$$

  Feed A_t to $\mathcal{L}$ and obtain prediction $\hat{X}_t$

  Play A_t and receive reward X_t

  Feed X_t to $\mathcal{L}$ as feedback

Regret of OLR-UCB

Theorem

With probability at least $1 - \delta$ the pseudo-regret of OLR-UCB satisfies

$$\hat{R}_n^{pseudo} \leq \sqrt{8dn\left(M_2^2 + \beta_{n-1}(\delta)\right)\log\left(1 + \frac{n}{d}\right)}.$$

The Regret of OLR-UCB(π)

Theorem (Sparse OLR Algorithm)

$\exists\pi$ for the learner such that for any $\theta \in \mathbb{R}^d$, the regret $\rho_n(\theta)$ of π against any strategic environment such that $\max_{t\in[n]} \|A_t\|_2 \le L$ and $\max_{t\in[n]} |X_t| \le X$ satisfies

$$\rho_n(\theta) \le cX^2 \|\theta\|_0 \left\{\log(e + n^{1/2}L) + C_n \log\left(1 + \frac{\|\theta\|_1}{\|\theta\|_0}\right)\right\} + (1 + X^2)C_n \,,$$

where $c > 0$ is some universal constant and $C_n = 2 + \log_2 \log(e + n^{1/2}L)$.

Corollary

The expected regret of OLR-UCB when using the strategy π from above satisfies

$$R_n = \tilde{O}(\sqrt{dpn})\,.$$

Outline

Summary

- OLR algorithm used inside OLR-UCB to construct center
- Regret guarantee of the OLR controls "width" of confidence ellipsoid
- Regret: $\tilde{O}(\sqrt{dpn})$, and p is known.
- Hypercube: $p\sqrt{n}$, p is unknown!
- In general, the regret can be as high as $\Omega(\sqrt{pdn})$ ($p = 1$: think of $\mathcal{A} = \{e_1, \ldots, e_d\}$)
- Under parameter noise ($X_t = \langle A_t, \theta_* + \eta_t \rangle$), for "rounded" action sets, $\tilde{O}(p\sqrt{n})$ is possible!
- Very much unlike in the "passive" case:

  Major conflict between exploration and exploitation!

Historical Notes

- Selective Explore-Then-Commit algorithm is due to (Lattimore et al., 2015).
- OLR-UCB is from Abbasi-Yadkori et al. (2012).
- The Sparse OLR algorithm is due to Gerchinovitz (2013).
- Rakhlin and Sridharan (2015) also discusses relationship between online learning regret bounds and self-normalized tail bounds of the type given here.

Outline

Minimax Lower Bound

Theorem

Let the action set be $\mathcal{A} = \{-1, 1\}^d$ *and* $\Theta = \{-n^{-1/2}, n^{-1/2}\}^d$. *Then for any policy* π *there exists a* $\theta \in \Theta$ *such that*

$$R_n^{\pi}(\mathcal{A}, \theta) \geq C\, d\sqrt{n}$$

for some universal constant $C > 0$.

Some Thoughts

- LinUCB with our confidence set construction is “nearly” worst-case optimal.
- The theorem is “new”, but the proof is standard; see (Shamir, 2015).
- Similar results for some other action sets: Rusmevichientong and Tsitsiklis (2010) (ℓ^2-ball), Dani et al. (2008) (products of 2D balls).
- Some action sets will have smaller minimax regret! Can you think of one?

Outline

Lower Bound

Setting:

1. Actions: $\mathcal{A} \subset \mathbb{R}^d$ finite, $K = |\mathcal{A}|$.
2. Reward is $X_t = \langle A_t, \theta \rangle + \eta_t$, where $\theta \in \mathbb{R}^d$ and η_t is a sequence of independent standard Gaussian variables.

Regret of policy π:

$$R_n^\pi(\mathcal{A}, \theta) = \mathbb{E}_{\theta,\pi}\left[\sum_{t=1}^n \Delta_{A_t}\right] , \qquad \Delta_a = \max_{a' \in \mathcal{A}} \langle a' - a, \theta \rangle ,$$

Recall: a policy π is consistent in some class of bandits $\mathcal{E}$ if the regret is subpolynomial for any bandit in that class:

$$R_n^\pi(\mathcal{A}, \theta) = o(n^p) \qquad \text{for all } p > 0 \text{ and } \theta \in \mathbb{R}^d .$$

Lower Bound: II

Theorem

Assume that $\mathcal{A} \subset \mathbb{R}^d$ *is finite and spans* $\mathbb{R}^d$ *and suppose* π *is consistent. Let* $\theta \in \mathbb{R}^d$ *be any parameter such that there is a unique optimal action and let* $\bar{G}_n = \mathbb{E}_{\theta,\pi}\left[\sum_{t=1}^n A_t A_t^\top\right]$ *be the expected Gram matrix . Then* $\liminf_{n\to\infty} \lambda_{\min}(\bar{G}_n)/\log(n) > 0$*. Furthermore, for any* $a \in \mathcal{A}$ *it holds that:*

$$\limsup_{n\to\infty} \log(n) \|a\|^2_{\bar{G}_n^{-1}} \leq \frac{\Delta_a^2}{2} .$$

Lower Bound: III

Corollary

Let $\mathcal{A} \subset \mathbb{R}^d$ be a finite set that spans $\mathbb{R}^d$ and $\theta \in \mathbb{R}^d$ be such that there is a unique optimal action. Then for any consistent policy π,

$$\liminf_{n \to \infty} \frac{R_n^\pi(\mathcal{A}, \theta)}{\log(n)} \geq c(\mathcal{A}, \theta)\,,$$

where $c(\mathcal{A}, \theta)$ is defined as

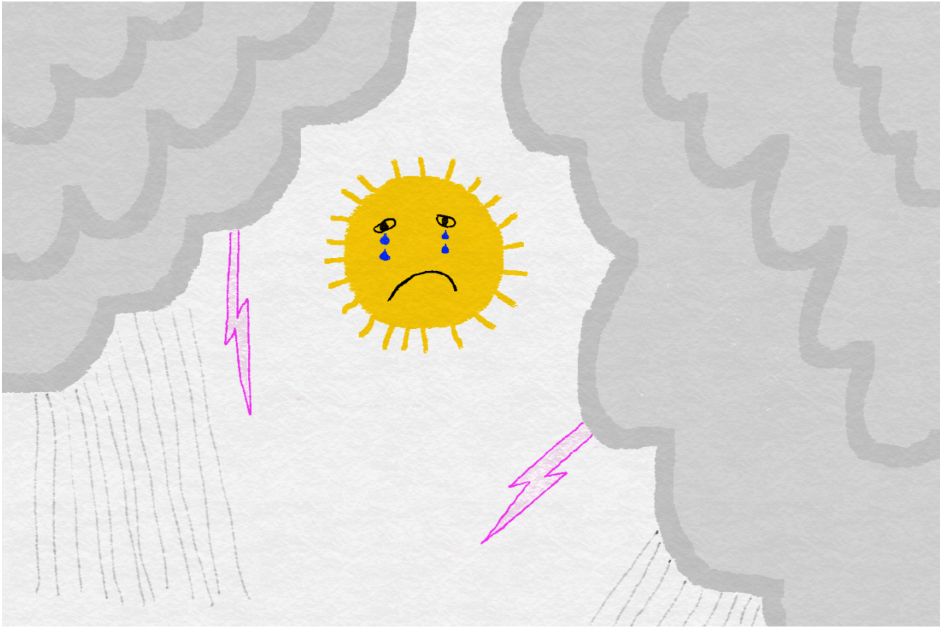
$$c(\mathcal{A}, \theta) = \inf_{\alpha \in [0,\infty)^{\mathcal{A}}} \sum_{a \in \mathcal{A}} \alpha(a) \Delta_a$$

$$\text{subject to } \|a\|^2_{H_\alpha^{-1}} \leq \frac{\Delta_a^2}{2} \text{ for all } a \in \mathcal{A} \text{ with } \Delta_a > 0\,,$$

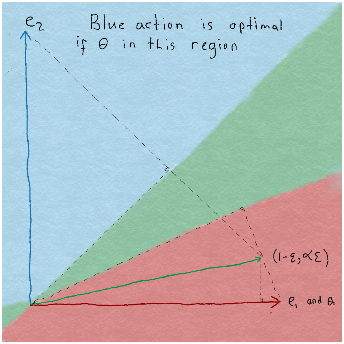
where $H = \sum_{a \in \mathcal{A}} \alpha(a) a a^\top$.

Outline

Poor Outlook for Optimism



Poor Outlook for Optimism



Actions:
$\mathcal{A} = \{a_1, a_2, a_3\}$, $a_1 = e_1$, $a_2 = e_2$, $a_3 = (1-\varepsilon, \gamma\varepsilon)$. $\varepsilon > 0$ small, $\gamma \geq 1$.

Let $\theta = (1, 0)$, so $a^* = a_1$.

Solving for the lower bound,
$\alpha(a_2) = 2\gamma^2$
and $\alpha(a_3) = 0$, $c(\mathcal{A}, \theta) = 2\gamma^2$ and

$$\liminf_{n\to\infty} \frac{R_n^\pi(\mathcal{A}, \theta)}{\log(n)} = 2\gamma^2 \,.$$

Moreover, for γ large, ε sufficiently small, π “optimistic”,

$$\limsup_{n\to\infty} \frac{R_n^\pi(\mathcal{A}, \theta)}{\log(n)} = \Omega(1/\varepsilon) \,,$$

Instance-Optimal Asymptotic Regret

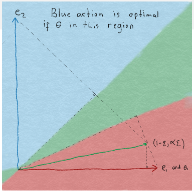
Theorem

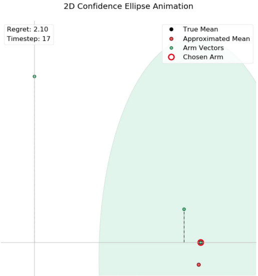
There exists a policy π that is consistent and satisfies

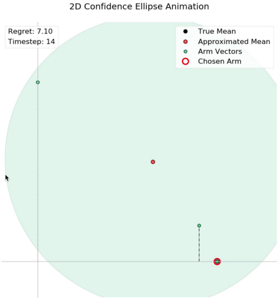
$$\limsup_{n\to\infty} \frac{R_n^\pi(\mathcal{A},\theta)}{\log(n)} = c(\mathcal{A},\theta)\,,$$

where $c(\mathcal{A},\theta)$ was defined in the lower bound.

Illustration: LinUCB







Summary

The instance-optimal regret of consistent algorithms is asymptotically $c(\mathcal{A}, \theta) \log(n)$.

Optimistic algorithms fail to achieve this: Their regret can be worse by an arbitrarily large constant factor.

Remember:

Finite-armed bandits

Case (a): $\mathcal{A}_t$ has always the same number of vectors in it:

"finite-armed stochastic contextual bandit".

Case (b): Also, $\mathcal{A}_t$ does not change, or $\mathcal{A}_t = \{a_1, \ldots, a_K\}$:

"finite-armed stochastic linear bandit".

Case (c): If the vectors in $\mathcal{A}_t$ are also orthogonal to each other:

"finite-armed stochastic bandit".

Difference between cases (c) and (b):

- Case (c): Learn about mean of arm $i \Leftrightarrow$ Choose action i;
- Case (b): Learn about mean of arm $i \Leftrightarrow$ Choose action j s.t. $\langle x_j, x_i \rangle \neq 0$.

Departing Thoughts

- These results are from Lattimore and Szepesvári (2016)
- The asymptotically optimal algorithm is given there (the algorithm solves for the optimal allocation, while monitoring whether things went wrong)
- Combes et al. (2017) refine the algorithm and generalize it to other settings.
- Soare et al. (2014), in best arm identification with linear payoff functions, gave essentially the same example that we use to argue for the large regret of optimistic algorithms.
- Open questions:
  - Simultaneously finite-time near-optimal and asymptotically optimal algorithm
  - Changing, or infinite action sets?

Summary

Summary of This Talk

- Contextual vs. linear bandits:

  Changing action sets can model contextual bandits

- Optimistic algorithms:
  - Optimism can achieve minimax optimality
  - Optimism can be expensive
  - Optimistic algorithms require a careful design of the underlying confidence sets

- Sparsity:

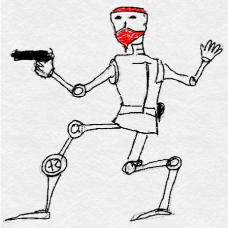
  Exploiting sparsity is sometimes at odds with the requirement to collect rewards

What's Next for Bandits?

- Today: Finite-armed and linear stochastic bandits.
- But bandits come in all forms and shapes!
  - Adversarial (finite, linear, . . .)
  - Combinatorial action sets: From shortest path to ranking
  - Continuous action sets, continuous time, delays
  - Resourceful, nonstationary, various structures (low-rank), . . .
- Nearby problems:
  - Reinforcement learning/Markov decision processes
  - Partial monitoring

Learning Material

- Bandit Visualizer:
  `https://github.com/alexrutar/banditvis`
- Online bandit simulator:
  `http://downloads.tor-lattimore.com/bandits/`
- Most of this tutorial (and more): `http://banditalgs.com`
  - Book to be published by early next year: Looking for reviewers!
  - Tor's lightweight C++ bandit library
- Sebastien Bubeck's tutorial
  - Blog post 1
  - Blog post 2
- Bubeck and Cesa-Bianchi's book;
  (Bubeck and Cesa-Bianchi, 2012)



`banditalgs.com`

References I

Abbasi-Yadkori, Y. (2009). *Forced-exploration based algorithms for playing in bandits with large action sets*. PhD thesis, University of Alberta.

Abbasi-Yadkori, Y. (2012). *Online Learning for Linearly Parametrized Control Problems*. PhD thesis, University of Alberta.

Abbasi-Yadkori, Y., Antos, A., and Szepesvári, C. (2009). Forced-exploration based algorithms for playing in stochastic linear bandits. In *COLT Workshop on On-line Learning with Limited Feedback*.

Abbasi-Yadkori, Y., Pal, D., and Szepesvari, C. (2012). Online-to-confidence-set conversions and application to sparse stochastic bandits. In *Artificial Intelligence and Statistics*, pages 1–9.

Abbasi-Yadkori, Y., Szepesvári, C., and Tax, D. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2312–2320.

Abe, N. and Long, P. M. (1999). Associative reinforcement learning using linear probabilistic concepts. In *ICML*, pages 3–11.

Anantharam, V., Varaiya, P., and Walrand, J. (1987). Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: Iid rewards. *IEEE Transactions on Automatic Control*, 32(11):968–976.

Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422.

Auer, P. and Ortner, R. (2010). UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65.

References II

Bubeck, S. and Cesa-Bianchi, N. (2012). *Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems*. Foundations and Trends in Machine Learning. Now Publishers Incorporated.

Chu, W., Li, L., Reyzin, L., and Schapire, R. E. (2011). Contextual bandits with linear payoff functions. In *AISTATS*, volume 15, pages 208–214.

Combes, R., Magureanu, S., and Proutiere, A. (2017). Minimal exploration in structured stochastic bandits. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 1761–1769. Curran Associates, Inc.

Dani, V., Hayes, T. P., and Kakade, S. M. (2008). Stochastic linear optimization under bandit feedback. In *Proceedings of Conference on Learning Theory (COLT)*, pages 355–366.

Even-Dar, E., Mannor, S., and Mansour, Y. (2006). Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(Jun):1079–1105.

Filippi, S., Cappé, O., Garivier, A., and Szepesvári, Cs. (2009). Parametric bandits: The generalized linear case. In *NIPS-22*, pages 586–594.

Freedman, D. (1975). On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118.

Gerchinovitz, S. (2013). Sparsity regret bounds for individual sequences in online linear regression. *Journal of Machine Learning Research*, 14(Mar):729–769.

Hsu, D., Kakade, S. M., and Zhang, T. (2012). Random design analysis of ridge regression. In *Conference on Learning Theory*, pages 9–1.

References III

John, F. (1948). Extremum problems with inequalities as subsidiary conditions. *Courant Anniversary Volume, Interscience*.

Lattimore, T., Crammer, K., and Szepesvári, C. (2015). Linear multi-resource allocation with semi-bandit feedback. In *Advances in Neural Information Processing Systems*, pages 964–972.

Lattimore, T. and Munos, R. (2014). Bounded regret for finite-armed structured bandits. In *Advances in Neural Information Processing Systems*, pages 550–558.

Lattimore, T. and Szepesvári, C. (2016). The end of optimism? an asymptotic analysis of finite-armed linear bandits. *arXiv preprint arXiv:1610.04491*.

Peña, V. H., Lai, T. L., and Shao, Q.-M. (2008). *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media.

Rakhlin, A. and Sridharan, K. (2015). On equivalence of martingale tail bounds and deterministic regret inequalities. *arXiv preprint arXiv:1510.03925*.

Robbins, H. and Siegmund, D. (1970). Boundary crossing probabilities for the Wiener process and sample sums. *Annals of Math. Statistics*, 41:1410–1429.

Robbins, H. and Siegmund, D. (1971). A convergence theorem for non-negative almost supermartingales and some applications. In Rustagi, J., editor, *Optimizing Methods in Statistics*, pages 235–257. Academic Press, New York.

Rusmevichientong, P. and Tsitsiklis, J. N. (2010). Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411.

Russo, D. and Roy, B. V. (2013). Eluder dimension and the sample complexity of optimistic exploration. In *NIPS*, pages 2256–2264.

References IV

Shamir, O. (2015). On the complexity of bandit linear optimization. In *Conference on Learning Theory*, pages 1523–1551.

Soare, M., Lazaric, A., and Munos, R. (2014). Best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems*, pages 828–836.

Srinivas, N., Krause, A., Kakade, S., and Seeger, M. W. (2010). Gaussian process optimization in the bandit setting: No regret and experimental design. In *ICML*, pages 1015–1022.

Valko, M., Korda, N., Munos, R., Flaounas, I., and Cristianini, N. (2013). Finite-time analysis of kernelised contextual bandits. *arXiv preprint arXiv:1309.6869*.

Valko, M., Munos, R., Kveton, B., and Kocák, T. (2014). Spectral bandits for smooth graph functions. In *International Conference on Machine Learning*, pages 46–54.