Linear (and contextual) Bandits: Rich decision sets (and side information)

Sham M. Kakade

Machine Learning for Big Data
CSE547/STAT548

University of Washington
Poster session: June 1, 9-11:30a

- **Request:** CSE grad students, could you please help others with poster printing?
- Aravind: Ask by 2p on Weds for help printing.
- Prepare, **at most**, a 2 minute verbal summary.
- Come earlier to setup.
- Submit your poster on Canvas.

**Due Dates:** Please be on time.

---

Today:

- review: Linear bandits
- today: contextual bandits, game trees?
Review
Bandits in practice: two major issues

The decision space is very large.
  - Drug cocktails
  - Ad design

We often have “side information” when making a decision
  - history of a user
More real motivations...

**Clinical trials:**

\[ B(\mu_1) \quad B(\mu_2) \quad B(\mu_3) \quad B(\mu_4) \quad B(\mu_5) \]

- Choose a treatment \( A_t \) for patient \( t \)
- Observe a response \( X_t \in \{0, 1\} : \mathbb{P}(X_t = 1) = \mu_{A_t} \)
- **Goal:** maximize the number of patient healed

**Recommendation tasks:**

\[ \nu_1 \quad \nu_2 \quad \nu_3 \quad \nu_4 \quad \nu_5 \]

- Recommend a movie \( A_t \) for visitor \( t \)
- Observe a rating \( X_t \sim \nu_{A_t} \) (e.g. \( X_t \in \{1, \ldots, 5\} \))
Linear bandits

- An additive effects model.
- Suppose each round we take a decision $x \in \mathcal{D} \subset \mathbb{R}^d$.
  - $x$ is paths on a graph.
  - $x$ is a feature vector of properties of an ad
  - $x$ is a which drugs are being taken
- Upon taking action $x$, we get reward $r$, with expectation:
  $$\mathbb{E}[r|x] = \mu^\top x$$
- only $d$ unknown parameters (and “effectively” $2^d$ actions)
- We desire an algorithm $\mathcal{A}$ (mapping histories to decisions), which has low regret.
  $$T \mu^\top x_* - \sum_{t=1}^{T} \mathbb{E}[\mu^\top x_t|\mathcal{A}] \leq ??$$
  (where $x_*$ is the best decision)
Example: Shortest paths...

\[ x \in \{0, 1\}^E \]

\[ x \text{ is a path in } G \]
again, let’s think of optimism in the face of uncertainty
we observed some \( r_1, \ldots, r_{t-1} \), and have taken \( x_1, \ldots, x_{t-1} \).

Questions:
- what is an estimate of the reward of \( \mathbb{E}[r|x] \) and what is our uncertainty?
- what is an estimate of \( \mu \) and what is our uncertainty?
Regression!

Define:

\[ A_t := \sum_{\tau < t} x_{\tau} x_{\tau}^\top + \lambda I, \quad b_t := \sum_{\tau < t} x_{\tau} r_{\tau} \]

Our estimate of \( \mu \)

\[ \hat{\mu}_t = A_t^{-1} b_t \]

Confidence of our estimate:

\[ \| \mu - \hat{\mu}_t \|_{A_t}^2 \leq \mathcal{O}(d \log t) \]

\[ ((\mu - \hat{\mu}_t)^\top A_t^{-1} \mu - \lambda_2 (\mu - \hat{\mu}_t)^\top A_t^{-1} (\mu - \hat{\mu}_t)) \]
Again, optimism in the face of uncertainty.

Define:

\[ B_t := \{ \nu \mid \| \nu - \hat{\mu}_t \|_{A_t}^2 \leq O d \log t \} \]

(Lin UCB) take action:

\[ x_t = \arg\max_{x \in \mathcal{D}} \max_{\nu \in B_t} \nu^T x \]

then update \( A_t, B_t, b_t, \) and \( \hat{\mu}_t. \)

Equivalently, take action:

\[ x_t = \arg\max_{x \in \mathcal{D}} \hat{\mu}_t^T x + (d \log t) \sqrt{x A_t^{-1} x} \]
LinUCB: Confidence intervals

\[ \sqrt{\frac{1}{t} \sum_{i=1}^{t} V_i} \]
Today
Regret bound of LinUCB

\[ T \mu^\top x^* - \sum_{t=1}^{T} \mathbb{E}[\mu^\top x_t] \leq^* (d \sqrt{T}) \]

(this is the best possible, up to log factors).

Compare to \( O(\sqrt{KT}) \)

- Independent of number of actions.
- \( k \)-arm case is a special case.

Thompson sampling: This is a good algorithm in practice.
Proof Idea...

Stats: need to show that $B_t$ is a valid confidence region.

Geometric lemma: The regret is upper bounded by the:

$$\log \frac{\text{volume of posterior cov}}{\text{volume of prior cov}}$$

Then just bound the worst case log volume change.
What about context?

Clinical trials:

\[ B(\mu_1) \quad B(\mu_2) \quad B(\mu_3) \quad B(\mu_4) \quad B(\mu_5) \]

- choose a treatment \( A_t \) for patient \( t \)
- observe a response \( X_t \in \{0, 1\} : \mathbb{P}(X_t = 1) = \mu_{A_t} \)

Goal: maximize the number of patient healed

Recommendation tasks:

\[ \nu_1 \quad \nu_2 \quad \nu_3 \quad \nu_4 \quad \nu_5 \]

- recommend a movie \( A_t \) for visitor \( t \)
- observe a rating \( X_t \sim \nu_{A_t} \) (e.g. \( X_t \in \{1, \ldots, 5\} \))
The Contextual Bandit Game

- **Game:** for $t = 1, 2, \ldots$
  - At each time $t$, we obtain context (e.g. side information, user information) $c_t$
  - Our feasible action set is $A_t$.
  - We choose arm $a_t \in A_t$ and receive reward $r_{t,a_t}$.
    (what assumptions on the reward process?)

- **Goal:** Algorithm $\mathcal{A}$ to have low regret:

$$
\mathbb{E}[\sum_t (r_{t,a_t^*} - r_t) | \mathcal{A}] \leq ??
$$

where $\mathbb{E}[r_{t,a_t^*}]$ is the optimal expected reward at time $t$. 
How should we model outcomes?

Example: ad (or movie, song, etc) prediction. What is prob. that a user $u$ clicks on an ad $a$?

How should we model the click probability of $a$ for user $u$?

Featurizations: suppose we have $\phi_{ad}(a) \in \mathcal{R}^{d_{ad}}$ and $\phi_{user}(u) \in \mathcal{R}^{d_{user}}$.

We could make an “outer product” feature vector $x$ as:

$$x(a, u) = \text{Vector}(\phi_{ad}(a)\phi_{user}(u)^\top) \in \mathcal{R}^{d_{ad}\cdot d_{user}}$$

We could model the probabilities as:

$$\mathbb{E}[\text{click} = 1 | a, u] = \mu^\top x(a, u)$$

(or log linear)

How do we estimate $\mu$?
Contextual Linear bandits

- Suppose each round $t$, we take a decision $x \in D_t \subseteq \mathcal{R}^d$ ($D_t$ may be time varying).
- map each ad/user $a$ to $x(a, u)$.
- $D_t = \{x(a, u_t)| a \text{ is a feasible ad at time } t\}$
- Our decision is a feature vector in $x \in D_t$.

- Upon taking action $x_t \in D_t$, we get reward $r_t$, with expectation:

$$
E[r_t|x_t \in D_t] = \mu^\top x_t
$$

(here $\mu$ is assumed constant over time).

- Our regret:

$$
E[\sum_t (\mu^\top x_t, a_t^* - \mu^\top x_t)|A] \leq ??
$$

(where $x_t, a_t^*$ is the best decision at time $t$)
let’s just run linUCB (or Thompson sampling)

Nothing really changes:
- $A_t$ and $b_t$ are the same updating rules
- now our decision is:

$$x_t = \arg\max_{x \in \mathcal{D}_t} \max_{\nu \in B_t} \nu^\top x$$

i.e.

$$x_t = \arg\max_{x \in \mathcal{D}_t} \hat{\mu}_t^\top x + (d \log t) \sqrt{x A_t^{-1} x}$$

Regret bound is still $O(d \sqrt{T})$. 
Acknowledgements

- https://sites.google.com/site/banditstutorial/
- http://www.yisongyue.com/courses/cs159/lectures/LinUCB.pdf