Lecture 20: Linear Contextual Bandits

1) Setting

Our simplified MDP setting consists of:
- contexts \( x \in \mathcal{X} \subseteq \mathbb{R}^d \)
  - drawn from distribution \( \mathcal{D} \in \Delta(\mathcal{X}) \) \( x_t \sim \mathcal{D} \)
- actions “arms” \( a \in \mathcal{A} = \{1, \ldots, K\} \)
- rewards \( r_t = r(x_t, a_t) \) with
  \( \mathbb{E}[r(x_t, a)] = y_a(x) = \Theta^T_a x \) linear function
- Horizon \( T \)

Goal: find a policy \( a_t = \Pi(x_t) \) that achieves low regret.

\[
R(T) = \sum_{t=1}^{T} \mathbb{E} \left[ \max_{a} \Theta^T_a x_t - \Theta^T_{a^*_t} x_t \right]
\]

\( y_*(x), a_* \)

Example: music recommendation
- arms \( a \) are artists
- \( \Theta_a \in \mathbb{R}^d \) represents attributes
- \( x \in \mathbb{R}^d \) represents a user's affinity towards the attributes (observed from listening history)
Last lecture we considered an explore-then-commit algorithm for general function approximation/supervised learning.

\[ a_t = \arg \max_a \hat{y}_a(x_t) \quad \text{where} \]

\[ \hat{y}_a = \arg \min_{y \in M} \sum_{i=1}^{N} (y(x_i^a) - r_i^a)^2 \]

data collected during exploration phase.

**Linear Regression**

If we know that \( y_a(x) = \Theta^T x \) we can instantiate the general supervised learning framework with

\[ M = \{ y(x) = \Theta^T x \mid \Theta \in \mathbb{R}^d \} \]

In this case the learning problem is equivalent to

\[ \hat{\Theta}_a = \arg \min_{\Theta} \sum_{i=1}^{N} (\Theta^T x_i^a - r_i^a)^2 \]

We will sometimes drop the subscript in these notes.

**Lemma:** As long as \( (x_i)_{i=1}^{N} \) span \( \mathbb{R}^d \),

\[ \hat{\Theta} = \left( \sum_{i=1}^{N} x_i x_i^T \right)^{-1} \sum_{i=1}^{N} x_i r_i = A^{-1} b \]

**Proof:**

\[ \nabla_{\Theta} \sum_{i=1}^{N} (\Theta^T x_i - r_i)^2 = 2 \sum_{i=1}^{N} x_i (x_i^T \Theta - r_i) \]

setting the gradient equal to zero,

\[ \left( \sum_{i=1}^{N} x_i x_i^T \right) \Theta = \sum_{i=1}^{N} x_i r_i \]

\[ A \quad b \]
The matrix on the left hand side is invertible if \((x_i)_{i=1}^N \text{ span } \mathbb{R}^d\).

(Why? Let \(X = \begin{bmatrix} x_1^T \\ \vdots \\ x_N^T \end{bmatrix} \in \mathbb{R}^{N \times d}\). Then if \(x_i\) span \(\mathbb{R}^d\), \(X\) has full row rank, \(\text{rank}(X) = d\).

\[
\sum_{i=1}^N x_i x_i^T = X^T X \in \mathbb{R}^{d \times d}
\]

is full rank because \(\text{rank}(X^T X) = \text{rank}(X) = d\). Therefore, it is invertible.

The matrix \(A\) is related to the empirical covariance

\[
\Sigma = \mathbb{E}_{x \sim \mathcal{D}}[xx^T]
\]

\[
\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^N x_i x_i^T
\]

We can relate \(A = N \hat{\Sigma}\).

2) Interactive Demo - dyputer Notebook
3) LinUCB Algorithm

Recall that last lecture we wanted to estimate conditional errors $\mathbb{E}[(y_a(x) - y_a(x))^2 | X]$. Using the structure of the linear regression problem, we can do this.

We keep track of

$$A_a^t = \frac{t}{t} \sum_{k=1}^t x_k x_k^T \mathbb{I} \sum_{} a_k = a^2, \quad b_a^t = \sum_{k=1}^t x_k r_k \mathbb{I} \sum_{} a_k = a^2$$

$$\hat{\Theta}_a^t = (A_a^t)^{-1} b_a^t$$

Alg: LinUCB

1. Initialize $\Theta$ mean & infinite confidence intervals
2. For $t=1, \ldots, T$:
   $$a_t = \arg\max_a \Theta_a^T x_t + \alpha \sqrt{x_t (A_a^t)^{-1} x_t}$$
   update $\Theta_a^t$, $b_a^t$, $A_a^t$

Geometric Intuition:

$$\hat{\Theta}_a^T x + \alpha \sqrt{x^T A_a^{-1} x}$$

large if $x$ and $\hat{\Theta}$ are aligned

large if $x$ is not aligned with much historical data

$X^T A_a^{-1} X = X^T (N \hat{\Sigma})^{-1} X$

$= \frac{1}{\sqrt{2}} \sum X^T \hat{\Sigma}^{-1} X$

amount of data w/ data
ex- if previous data is $(X_1, -X_1, X_1, -X_1, -X_2)$
then $A^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1/2 \end{bmatrix}$

$X_1^TA^{-1}X_1 < X_2^TA^{-1}X_2$

more sure on speed less sure about pop

less sure on speed pretty sure about pop

Statistical Explanation:

Claim: With high probability (over noisy rewards)

$\Theta_a^T X \leq \hat{\Theta}_a^T X + \alpha \sqrt{x^T A_a^{-1} x}$

where $\alpha$ depends on probability & variance of rewards

Lemma (Chebychev's inequality)

for a random variable $u$ with $\mathbb{E}(u) = 0$,

$|u| \leq \beta \sqrt{\mathbb{E}(u^2)}$ with probability $1 - \frac{1}{\beta^2}$
Proof: we will use chebychevs to show that w.h.p  
\[
\left| \hat{\Theta}_a^T x - \Theta_a^T x \right| \leq \alpha \sqrt{\frac{\sum x_i^T A^{-1} x_i}{\text{Eu}^2}}
\]

1) compute expectation. Define \( W_i = r_i - \mathbb{E}[r_i] \) so \( r_i = \Theta_a^T x_i + w_i \).

\[
\Theta = \left( \sum_{i=1}^{N} x_i x_i^T \right)^{-1} \sum_{i=1}^{N} x_i (\Theta_a^T x_i + w_i)
\]

\[
= \left( \sum_{i=1}^{N} x_i x_i^T \right)^{-1} \sum_{i=1}^{N} x_i \Theta_a + \left( \sum_{i=1}^{N} x_i x_i^T \right)^{-1} \sum_{i=1}^{N} x_i w_i
\]

\[
\Theta - \Theta_a = A^{-1} \sum_{i=1}^{N} x_i w_i
\]

Therefore \( \mathbb{E}((\Theta - \Theta_a)^T x) = A^{-1} \sum_{i=1}^{N} x_i \mathbb{E}[w_i] = 0 \)

2) compute variance

\[
\mathbb{E}_\Theta \left[ (\Theta - \Theta_a)^T x \right]^2 = \mathbb{E}_w \left[ x^T A^{-1} \sum_{i=1}^{N} x_i w_i \cdot \sum_{i=1}^{N} x_i^T w_i A^{-1} x \right]
\]

\[
= x^T A^{-1} \mathbb{E} \left[ \sum_{i=1}^{N} x_i x_i^T w_i w_j \right] A^{-1} x
\]

The noise in rewards is iid so the expectation is 0 if \( i \neq j \). Define \( \sigma^2 \) as variance of rewards.

\[
= x^T A^{-1} \sum_{i=1}^{N} x_i x_i^T \sigma^2 A^{-1} x
\]

\[
= \sigma^2 x^T A^{-1} x
\]
Therefore, using Chebychev's, we have that w.p. $1 - \frac{1}{\beta^2}$,

$$|\hat{\theta}_a^T x - \theta_0^T x| \leq \beta \sigma \sqrt{x^T A^{-1} x}$$

Thus the upper bound of this confidence interval is

$$\hat{\theta}_a^T x + \alpha \sqrt{x^T A^{-1} x}$$