

Linear Contextual Bandits

Chen-Yu Wei

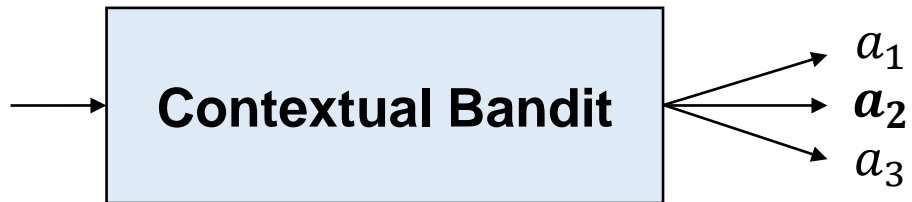
Contextual Bandits



all-user recommendation system



Context



personalized recommendation system

e.g. the user's historical
purchase record, location,
social network activity, ...

Contextual Bandits

For time $t = 1, 2, \dots, T$:

Environment generates a context $x_t \in \mathcal{X}$

Learner chooses an action $a_t \in \mathcal{A}$

Learner observes $r_t = R(x_t, a_t) + w_t$

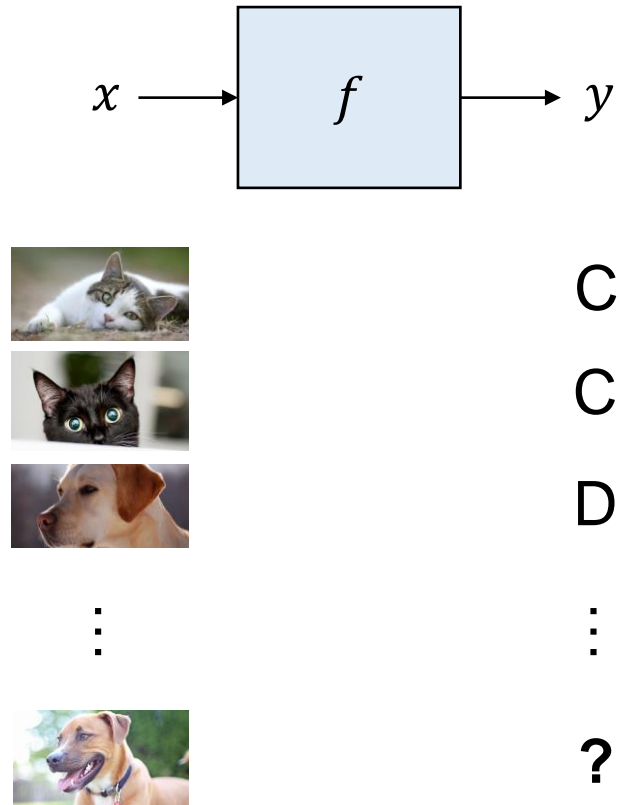
$$\begin{aligned} \text{Regret} &= \max_{\pi} \sum_{t=1}^T R(x_t, \pi(x_t)) - \sum_{t=1}^T R(x_t, a_t) & \text{Optimal policy: } \pi(x) &= \operatorname{argmax}_{a \in \mathcal{A}} R(x, a) \\ &= \sum_{t=1}^T \max_{a \in \mathcal{A}} R(x_t, a) - \sum_{t=1}^T R(x_t, a_t) \end{aligned}$$

View Each Context as a Separate MAB

$$\begin{aligned}\text{Regret} &= \sum_{t=1}^T \max_{a \in \mathcal{A}} R(x_t, a) - \sum_{t=1}^T R(x_t, a_t) \\ &= \sum_{x \in \mathcal{X}} \left(\sum_{t: x_t = x} \max_{a \in \mathcal{A}} R(x, a) - \sum_{t: x_t = x} R(x, a_t) \right)\end{aligned}$$

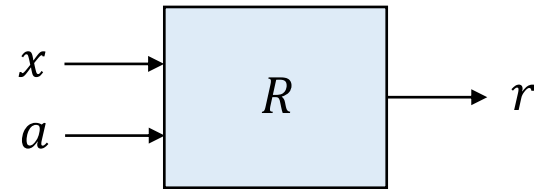
Not scalable and not generalizable

Function Approximation in Contextual Bandits

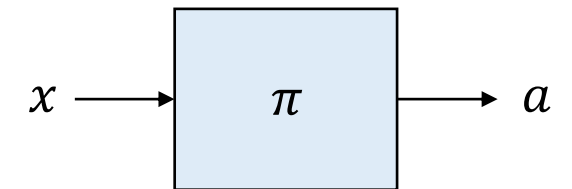


Find an f so that $f(x) \approx y$ for **seen** (x, y) pairs
Hoping that $f(x') \approx y'$ also holds for **unseen** x'

x : context, a : action, r : reward



value-based approach



policy-based approach

If a good approximation \hat{R} is found, a good policy can be derived as

$$\pi(x) = \operatorname{argmax}_a \hat{R}(x, a)$$

Linear Contextual Bandits

This is a linear **assumption**, not just linear **function approximation**. The former is stronger.

Linear Reward Assumption: $R(x, a) = \phi(x, a)^\top \theta^*$

$\phi(x, a) \in \mathbb{R}^d$ is a **feature vector** for the context-action pair (known to learner)

$\theta^* \in \mathbb{R}^d$ is the ground-truth **weight vector** (hidden from learner)

Given: feature mapping $\phi: \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}^d$

For time $t = 1, 2, \dots, T$:

Environment generates a context $x_t \in \mathcal{X}$

Learner chooses an action $a_t \in \mathcal{A}$

Learner observes $r_t = \phi(x_t, a_t)^\top \theta^* + w_t$ (w_t is zero-mean)

$$\text{Regret} = \sum_{t=1}^T \max_{a \in \mathcal{A}} R(x_t, a) - \sum_{t=1}^T R(x_t, a_t) = \sum_{t=1}^T \max_{a \in \mathcal{A}} \phi(x_t, a)^\top \theta^* - \sum_{t=1}^T \phi(x_t, a_t)^\top \theta^*$$

Linear CB is a Generalization of MAB

$$\phi(x,a) = e_a$$

$$\theta^* = \begin{bmatrix} R(1) \\ \vdots \\ R(A) \end{bmatrix}$$

$$e_a: (0, 0, \dots, 1, 0, \dots, 0)$$

↑
a-th entry

Key Questions in Linear Contextual Bandits

- How to obtain an estimated reward function $\hat{R}(x, a)$?
 - Was easy in multi-armed bandits – today we'll see how to do this in linear CB
- How to explore?
 - ϵ -greedy

$$a_t = \begin{cases} \text{uniform}(\mathcal{A}) & \text{with prob. } \epsilon \\ \operatorname{argmax}_a \hat{R}_t(x_t, a) & \text{with prob. } 1 - \epsilon \end{cases}$$

- Boltzmann exploration

$$p_t(a) \propto \exp(\lambda_t \hat{R}_t(x_t, a))$$

- Optimism in the face of uncertainty (LinUCB)
 - Thompson Sampling

How to Estimate the Reward Function $R(x, a)$?

- Recall $R(x, a) = \phi(s, a)^\top \theta^*$. We only need to estimate θ^* .
- At time t , we already gathered

$$r_1 = \phi(x_1, a_1)^\top \theta^* + w_1$$

$$r_2 = \phi(x_2, a_2)^\top \theta^* + w_2$$

\vdots

$$r_{t-1} = \phi(x_{t-1}, a_{t-1})^\top \theta^* + w_{t-1}$$

How to estimate θ^* ?

Linear Regression

Linear Regression

At time t , we have collected $(x_1, a_1, r_1), (x_2, a_2, r_2), \dots, (x_{t-1}, a_{t-1}, r_{t-1})$.

We want to generate an estimation $\hat{\theta}_t$ such that $\phi(x_i, a_i)^\top \hat{\theta}_t \approx r_i$

Linear Regression / Ridge Regression (define $\phi_i = \phi(x_i, a_i)$)

$$\hat{\theta}_t = \min_{\theta} \sum_{i=1}^{t-1} (\phi_i^\top \theta - r_i)^2 + \lambda \|\theta\|^2 \Leftrightarrow \hat{\theta}_t = \left(\lambda I + \sum_{i=1}^{t-1} \phi_i \phi_i^\top \right)^{-1} \left(\sum_{i=1}^{t-1} \phi_i r_i \right)$$

$\Rightarrow \hat{R}_t(x, a) = \phi(x, a)^\top \hat{\theta}_t$ (Use this directly in ϵ -greedy or Boltzmann exploration!)

To design a UCB algorithm, we have to quantify the estimation error $\hat{\theta}_t - \theta^*$

What can we say about $\hat{\theta}_t - \theta^*$?

Let's develop some intuition first.. (This intuition comes from Haipeng Luo's [lecture](#))

Let $r_i = \phi_i^\top \theta^* + w_i$ for $i = 1, \dots, N$

Assume $w_i \sim \mathcal{N}(0, \sigma^2)$, and

Assume $\{\phi_1, \dots, \phi_N\}$ are fixed vectors independent from $\{w_1, \dots, w_N\}$

Let

$$\hat{\theta} = \left(\sum_{i=1}^N \phi_i \phi_i^\top \right)^{-1} \left(\sum_{i=1}^N \phi_i r_i \right)$$

Question: What can we say about $\hat{\theta} - \theta^*$?

$$\begin{aligned} \hat{\theta} - \theta^* &= \left(\sum_{i=1}^N \phi_i \phi_i^\top \right)^{-1} \left(\sum_{i=1}^N \phi_i r_i \right) - \left(\sum_{i=1}^N \phi_i \phi_i^\top \right)^{-1} \left(\sum_{i=1}^N \phi_i \phi_i^\top \theta^* \right) \\ &= \left(\sum_{i=1}^N \phi_i \phi_i^\top \right)^{-1} \left(\sum_{i=1}^N \phi_i \underbrace{(r_i - \phi_i^\top \theta^*)}_{w_i} \right) \\ &= \underbrace{\left(\sum_{i=1}^N \phi_i \phi_i^\top \right)^{-1}}_{\Sigma} \left(\sum_{i=1}^N \phi_i w_i \right) \end{aligned}$$

$$\downarrow$$

$$a^\top M a = \|a\|_M$$

$$\begin{aligned} \Lambda &= \sum_{i=1}^N \phi_i \phi_i^\top \\ \mathbb{E}[\underbrace{z z^\top}_{\Lambda}] &= \mathbb{E} \left[\left(\sum_{i=1}^N \phi_i w_i \right) \left(\sum_{i=1}^N \phi_i w_i \right)^\top \right] \\ &= \mathbb{E} \left[\sum_{i=1}^N \underbrace{w_i^2}_{\sigma^2} \phi_i \phi_i^\top \right] \\ &= \Lambda \end{aligned}$$

let's check the covariance matrix of

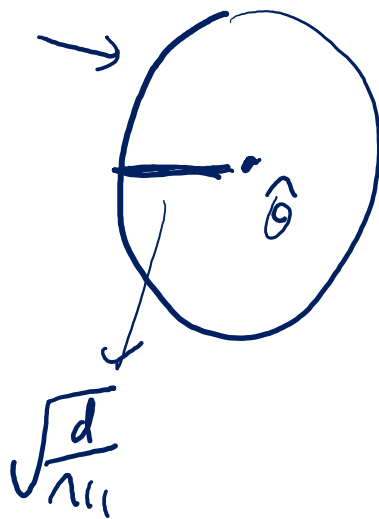
$$\mathbb{E} \left[\underbrace{\Lambda^{\frac{1}{2}} (\hat{\theta} - \theta^*) (\hat{\theta} - \theta^*)^\top \Lambda^{\frac{1}{2}}}_{\Lambda^{\frac{1}{2}} z z^\top \Lambda^{\frac{1}{2}}} \right] = I$$

$$\mathbb{E} \left[(\hat{\theta} - \theta^*)^\top \Lambda^{\frac{1}{2}} \Lambda^{\frac{1}{2}} (\hat{\theta} - \theta^*) \right] = d$$

$$\Rightarrow \mathbb{E} \left[\|\hat{\theta} - \theta^*\|_M \right] = d$$

Geometric Intuition

$$\|\hat{\theta} - \theta^*\|_{\Lambda}^2 = (\hat{\theta} - \theta^*) \begin{bmatrix} \Lambda_{11} & \Lambda_{12} & 0 \\ 0 & \ddots & \Lambda_{dd} \end{bmatrix} (\hat{\theta} - \theta^*) \leq d$$



$$\sum_i (\hat{\theta}_i - \theta_i^*)^2 \Lambda_{ii} \leq d$$

$$\boxed{(\hat{\theta}_1 - \theta_1^*)^2 \Lambda_{11} = d} \Rightarrow \text{radius}_1 = \sqrt{\frac{d}{\Lambda_{11}}}$$

$$\Lambda = \sum_{i=1}^{t-1} \phi_i \phi_i^T + I$$

Concentration Inequality for Linear Regression

Theorem.

In linear contextual bandits, assume w_t is zero-mean and 1-sub-Gaussian.
 $\|\phi(x, a)\|_2 \leq 1$, $\|\theta^*\|_2 \leq 1$.

Let

$$\hat{\theta}_t = \Lambda_t^{-1} \left(\sum_{i=1}^{t-1} \phi_i r_i \right), \quad \text{where } \Lambda_t = I + \sum_{i=1}^{t-1} \phi_i \phi_i^\top.$$

Then with probability at least $1 - \delta$, for all $t = 1, \dots, T$,

$$\|\theta^* - \hat{\theta}_t\|_{\Lambda_t}^2 \leq \beta \triangleq d \log \left(1 + \frac{T}{d} \right) + 3 \log \frac{1}{\delta}$$

Another Viewpoint on the Concentration Inequality

$$\begin{aligned}\|\theta^* - \hat{\theta}_t\|_{\Lambda_t}^2 &= (\theta^* - \hat{\theta}_t)^\top \left(I + \sum_{i=1}^{t-1} \phi_i \phi_i^\top \right) (\theta^* - \hat{\theta}_t) \\ &= \underbrace{\sum_{i=1}^{t-1} (\phi_i^\top \theta^* - \phi_i^\top \hat{\theta}_t)^2}_{\text{The difference between the predictions of } \theta^* \text{ and } \hat{\theta}_t \text{ over the past samples}} + \|\theta^* - \hat{\theta}_t\|^2 = O(d \log(T/\delta))\end{aligned}$$

The difference between the predictions of θ^* and $\hat{\theta}_t$ over the past samples

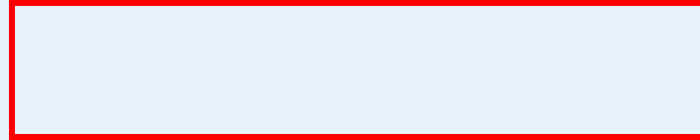
LinUCB

Most “optimistic” estimation for the reward of a

LinUCB

In round t , receive x_t , draw

$$a_t = \operatorname{argmax}_{a \in \mathcal{A}}$$



Observe $r_t = \phi(x_t, a_t)^\top \theta^* + w_t$.

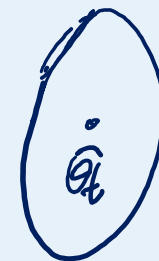
LinUCB

$$\max_{\theta} \phi(x_t, a)^T \theta = \underbrace{\phi(x_t, a)^T \hat{\theta}_t}_{\text{"}} + \underbrace{\phi(x_t, a)^T (\theta - \hat{\theta}_t)}_{\| \phi(x_t, a) \|_{\Lambda_t}^{-1} \underbrace{\| \theta - \hat{\theta}_t \|_{\Lambda_t}}_{\sqrt{\beta}}}$$

LinUCB

In round t , receive x_t , draw

$$a_t = \operatorname{argmax}_{a \in \mathcal{A}} \max_{\theta: \|\theta - \hat{\theta}_t\|_{\Lambda_t} \leq \beta} \phi(x_t, a)^T \theta$$



where

$$\hat{\theta}_t = \Lambda_t^{-1} \left(\sum_{i=1}^{t-1} \phi_i r_i \right), \quad \Lambda_t = I + \sum_{i=1}^{t-1} \phi_i \phi_i^T.$$

Observe $r_t = \phi(x_t, a_t)^T \theta^* + w_t$.

LinUCB

LinUCB

In round t , receive x_t , draw

$$a_t = \operatorname{argmax}_{a \in \mathcal{A}} \quad \phi(x_t, a)^\top \hat{\theta}_t + \sqrt{\beta} \|\phi(x_t, a)\|_{\Lambda_t^{-1}}$$

where

$$\hat{\theta}_t = \Lambda_t^{-1} \left(\sum_{i=1}^{t-1} \phi_i r_i \right), \quad \Lambda_t = I + \sum_{i=1}^{t-1} \phi_i \phi_i^\top.$$

Observe $r_t = \phi(x_t, a_t)^\top \theta^* + w_t$.

$$R(x, a) = \phi(x, a)^T \theta^*$$

Regret Analysis for LinUCB

Regret Bound of LinUCB

With probability at least $1 - \delta$,

$$\text{Regret} \leq O(d\sqrt{T} \log(T/\delta)) = \tilde{O}(d\sqrt{T}) .$$

$$\begin{aligned} \text{Regret} &= \sum_{t=1}^T \left(\max_a R(x_t, a) - R(x_t, a_t) \right) \\ &= \sum_{t=1}^T \left[\max_a \phi(x_t, a)^T \theta^* \right] - \phi(x_t, a_t)^T \theta^* \end{aligned}$$

$$\begin{aligned} \square &\leq \phi(x_t, a_t)^T \hat{\theta}_t + \sqrt{\beta} \|\phi(x_t, a_t)\|_{\Lambda_t^{-1}} \\ \text{Regret} &\leq \sum_t \underbrace{\phi(x_t, a_t)^T (\hat{\theta}_t - \theta^*)}_{\leq \sqrt{\beta} \|\phi(x_t, a_t)\|_{\Lambda_t^{-1}} \|\hat{\theta}_t - \theta^*\|_{\Lambda_t}} + \sqrt{\beta} \|\phi(x_t, a_t)\|_{\Lambda_t^{-1}} \\ &\leq \sum_t \underbrace{\|\phi(x_t, a_t)\|_{\Lambda_t^{-1}} \|\hat{\theta}_t - \theta^*\|_{\Lambda_t}}_{\leq \sqrt{\beta}} + \sqrt{\beta} \|\phi(x_t, a_t)\|_{\Lambda_t^{-1}} \\ &\leq 2 \sum_t \sqrt{\beta} \|\phi(x_t, a_t)\|_{\Lambda_t^{-1}} \sqrt{\beta} \end{aligned}$$

Elliptical Potential Lemma

Let $\phi_i \in \mathbb{R}^d$ and $\|\phi_i\|_2 \leq 1$. Define $\Lambda_t = I + \sum_{i=1}^{t-1} \phi_i \phi_i^\top$.

Then

$$\sum_{t=1}^T \|\phi_t\|_{\Lambda_t^{-1}}^2 \leq d \log \left(1 + \frac{T}{d} \right).$$

$$\sum_{t=1}^T \frac{a_t}{\left(\sum_{s=1}^{t-1} a_s \right) + 1} \leq \log T$$

when $0 \leq a_s \leq 1$

$$2 \sum_{t=1}^T \|\phi_t\|_{\Lambda_t^{-1}} \sqrt{\beta} \leq 2 \sqrt{\left(\sum_{t=1}^T \|\phi_t\|_{\Lambda_t^{-1}}^2 \right) \left(\sum_{t=1}^T 1 \right) \beta} \leq 2 \sqrt{d \log(\dots) \cdot T \beta}$$

$$\approx O(d \sqrt{T} \log(\dots))$$

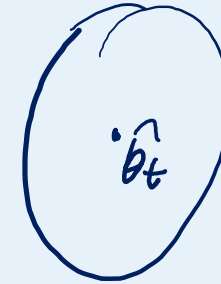
Thompson Sampling

Thompson Sampling for Linear Contextual Bandits

In round t , receive x_t , draw

$$\theta_t \sim \mathcal{N}(\hat{\theta}_t, \Lambda_t^{-1})$$

$$a_t = \operatorname{argmax}_{a \in \mathcal{A}} \phi(x_t, a)^\top \theta_t$$



where

$$\hat{\theta}_t = \Lambda_t^{-1} \left(\sum_{i=1}^{t-1} \phi_i r_i \right), \quad \Lambda_t = I + \sum_{i=1}^{t-1} \phi_i \phi_i^\top.$$

Observe $r_t = \phi(x_t, a_t)^\top \theta^* + w_t$.

There is no assumption on the distribution of x_t

- How is this possible?

