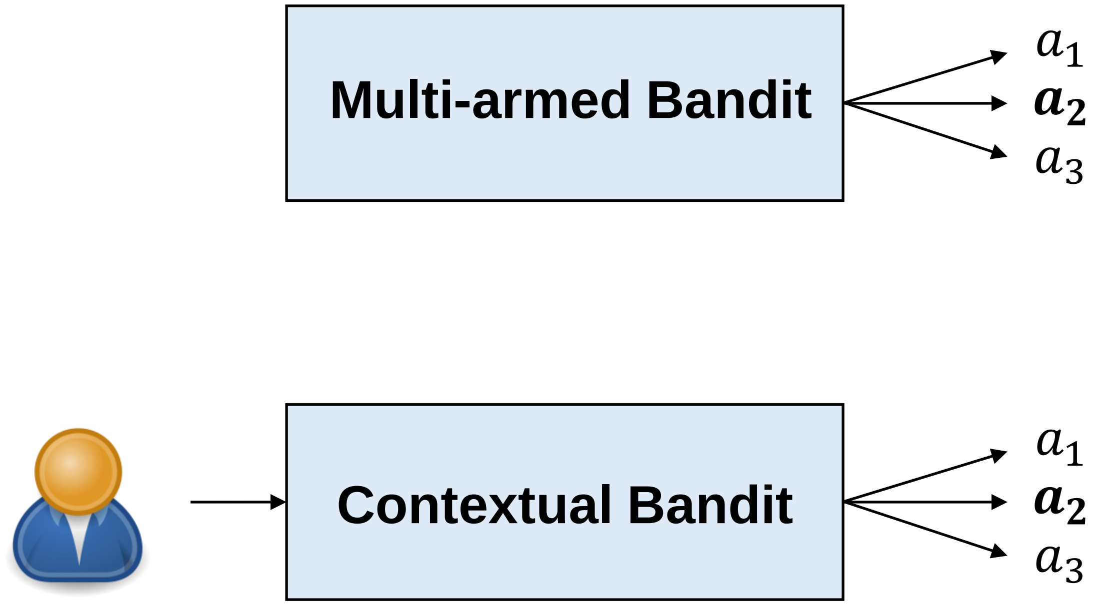
Linear Contextual Bandits

Contextual Bandits



all-user recommendation system

personalized recommendation system

Context

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

This example is from Chicheng Zhang's lecture notes

Contextual Bandits

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

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

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

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

$$\text{Regret} = \max_{\pi} \sum_{t=1}^{T} R(x_t, \pi(x_t)) - \sum_{t=1}^{T} R(x_t, a_t)$$

$$= \sum_{t=1}^{T} \max_{a \in \mathcal{A}} R(x_t, a) - \sum_{t=1}^{T} R(x_t, a_t)$$

Optimal policy: $\pi(x) = \operatorname*{argmax}_{a \in \mathcal{A}} R(x, a)$

View Each Context as a Separate MAB

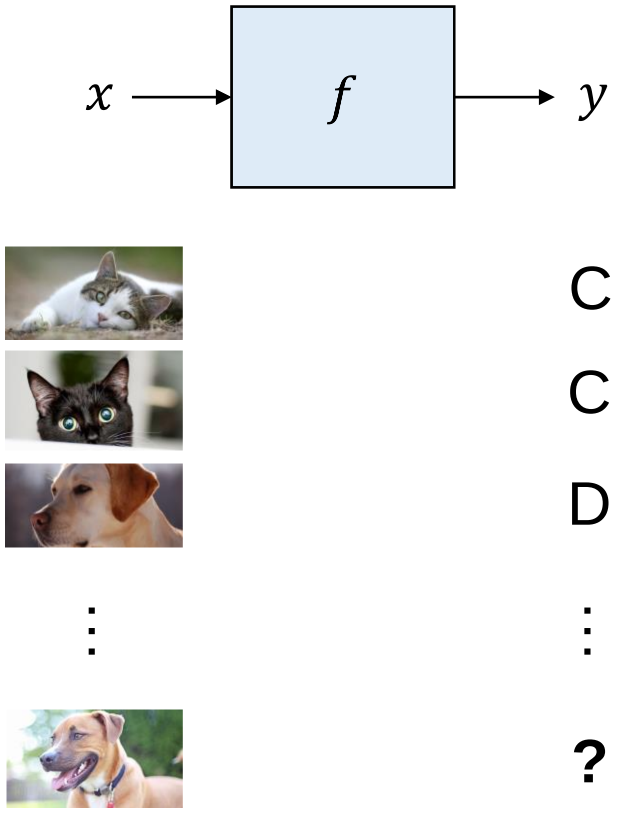
$$\text{Regret} = \sum_{t=1}^{T} \max_{a \in \mathcal{A}} R(x_t, a) \; - \sum_{t=1}^{T} R(x_t, a_t)$$

$$= \sum_{x \in \mathcal{X}} \left(\sum_{t: x_t = x} \max_{a \in \mathcal{A}} R(x, a) \; - \sum_{t: x_t = x} R(x, a_t) \right)$$

Not scalable and not generalizable

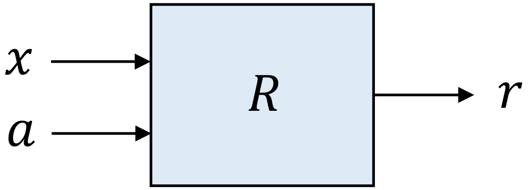
Function Approximation in Contextual Bandits

x: context, a: action, r: reward

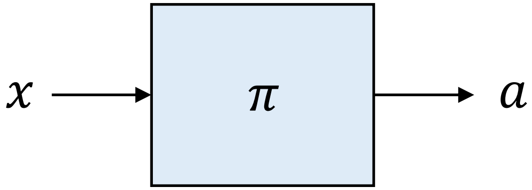


Find an f so that $f(x) \approx y$ for **seen** (x, y) pairs

Hoping that $f(x') \approx y'$ also holds for **unseen** x'



value-based approach



policy-based approach

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

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

Linear Contextual Bandits

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

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

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

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

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

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

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

Linear CB is a Generalization of MAB

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

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

  - Boltzmann exploration

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

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

How to Estimate the Reward Function $\boldsymbol{R(x, a)}$?

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

$$r_1 = \phi(x_1, a_1)^\top \theta^\star + w_1$$
$$r_2 = \phi(x_2, a_2)^\top \theta^\star + w_2$$
$$\vdots$$
$$r_{t-1} = \phi(x_{t-1}, a_{t-1})^\top \theta^\star + w_{t-1}$$

How to estimate $\theta^\star$?

Linear Regression

Linear Regression

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

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 $\phi_i = \phi(x_i, a_i)$)

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

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

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

$\boxed{\text{What can we say about } \hat{\theta}_t - \theta^\star\text{?}}$

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

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

Assume $w_i \sim \mathcal{N}(0, \sigma^2)$, and
Assume $\{\phi_1, \ldots, \phi_N\}$ are fixed vectors independent from $\{w_1, \ldots, w_N\}$

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 $\hat{\theta} - \theta^\star$?

Geometric Intuition

Concentration Inequality for Linear Regression

Theorem.

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

Let

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

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

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

Abbasi-Yadkori, Pal, Szepesvari. **Improved algorithms for linear stochastic bandits.** 2011.

LinUCB

Most “optimistic” estimation for the reward of a

LinUCB

In round t, receive x_t, draw

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

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

LinUCB

Most "optimistic" estimation for the reward of a

LinUCB

In round t, receive x_t, draw

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

where

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

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

Regret Analysis for LinUCB

Regret Bound of LinUCB

With probability at least $1 - T\delta$,

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

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

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

There is no assumption on the distribution of x_t

- How is this possible?