Contextual Bandits with Non-Linear / General Reward

Chen-Yu Wei

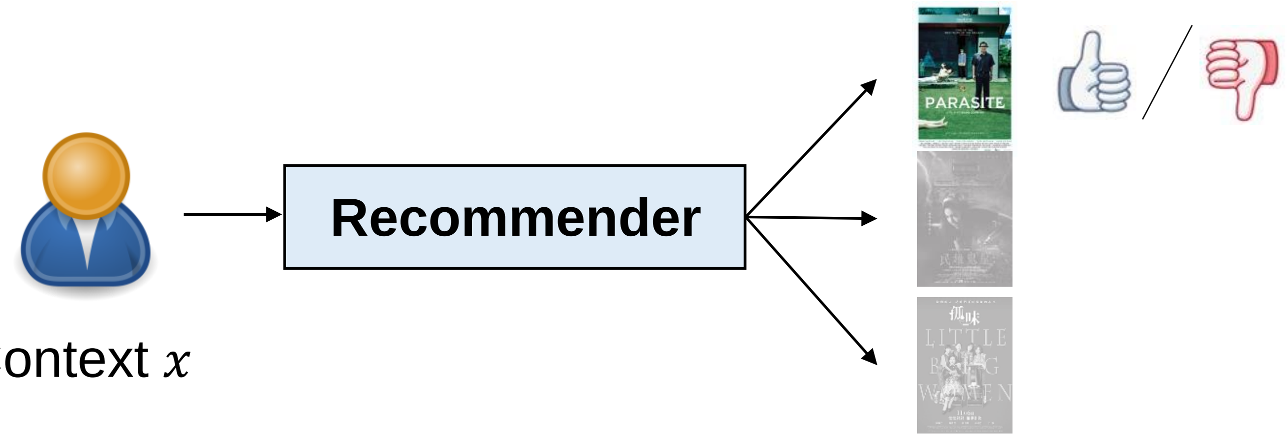
Topics

- Generalized linear contextual bandits
- A (optimal) reduction from contextual bandits to regression

Generalized Linear Contextual Bandits

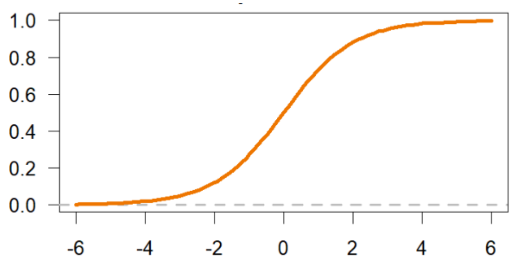
Contextual Bandits with Non-Linear Reward

Oftentimes, the reward may not be “approximately linear” in the feature vector.



$\approx \phi(x,a)^\top \theta^\star$?

Another option: Reward $\approx \mu(\, \phi(x,a)^\top \theta^\star \,)$

$\mu =$ 

Logistic Contextual Bandits

Logistic function: $\mu(z) = \frac{1}{1+e^{-z}}$

Logistic Reward Assumption: $R(x, a) = \frac{1}{1+\mathrm{e}^{-\phi(x,a)^\top \theta^\star}}$

$\phi(x, a) \in \mathbb{R}^d$ 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 $t = 1, 2, \ldots, T$:

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

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

Learner observes $r_t \sim \mathrm{Bernoulli}\left(\frac{1}{1+\mathrm{e}^{-\phi(x_t,a_t)^\top \theta^\star}}\right)$

Designing a CB algorithm involves

- Estimate $\theta^\star$ using data from time $1, 2, \ldots, t-1$.
  - MAB: calculate empirical mean for each arm
  - Linear CB: linear regression

  (The estimated $\hat{\theta}_t$ can be readily combined with naïve exploration methods e.g., ϵ-greedy, Boltzmann exploration)

- For more strategic exploration methods: identify the **confidence set** of $\theta^\star$ by quantifying the error between $\hat{\theta}_t$ and $\theta^\star$ (call this set Θ_t)
  - MAB: Hoeffding's inequality
  - Linear CB: some advanced concentration inequality

- **UCB**: $a_t = \operatorname*{argmax}_a \max_{\theta \in \Theta_t} R_\theta(x_t, a)$

  TS: $\theta_t \sim$ dist. over Θ_t, $\quad a_t = \operatorname*{argmax}_a R_{\theta_t}(x_t, a)$

UCB for Logistic Contextual Bandits

Estimation of $\boldsymbol{\theta}^\star$: $$\hat{\theta}_t = \underset{\theta}{\operatorname{argmin}} \underbrace{\sum_{i=1}^{t-1} \left(r_i \log\left(\frac{1}{\mu(\phi_i^\top \theta)}\right) + (1 - r_i) \log\left(\frac{1}{1 - \mu(\phi_i^\top \theta)}\right)\right)}_{\text{Logistic Loss}} + \lambda \|\theta\|^2$$

Cf. in Linear CB we use $$\hat{\theta}_t = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{t-1} \left(\phi_i^\top \theta - r_i\right)^2 + \lambda \|\theta\|^2$$

Confidence set: $$\|g_t(\theta_t) - g_t(\theta^\star)\|^2_{H_t(\theta^\star)^{-1}} \leq \beta \approx d$$

where $$g_t(\theta) \coloneqq \sum_{i=1}^{t-1} \mu(\phi_i^\top \theta)\phi_i + \lambda\theta, \quad H_t(\theta) := \sum_{i=1}^{t-1} \mu(\phi_i^\top \theta)\left(1 - \mu(\phi_i^\top \theta)\right)\phi_i \phi_i^\top + \lambda I$$

Regret bound: $\tilde{O}\left(d\sqrt{T}\right)$

Faury et al. Improved optimistic algorithms for logistic bandits. 2020.
Abeille et al. Instance-wise minimax-optimal algorithms for logistic bandits. 2021.
Faury et al. Jointly efficient and optimal algorithms for logistic bandits. 2022.

Generalized Linear Contextual Bandits

$R(x, a) = \mu(\phi(x, a)^\top \theta^\star)$ for any increasing function μ

Logistic CB $\subset$ Generalized Linear CB

UCB Algorithm:
Li et al. Provably optimal algorithms for generalized linear contextual bandits. 2017.

Even More General Case

General Function Class

- **Assumption:** the learner has access to a **function class** $\mathcal{F}$. It is guaranteed that the true reward function R is in $\mathcal{F}$.
- Linear CB is a special case where $\mathcal{F} = \left\{f\colon\ f(x,a) = \phi(s,a)^\top\theta \ \text{ for } \ \theta \in \mathbb{R}^d\right\}$
- Generalized linear CB is a special case where $\mathcal{F} = \left\{f\colon\ f(x,a) = \mu(\phi(s,a)^\top\theta) \ \text{ for } \ \theta \in \mathbb{R}^d \text{ and increasing } \mu\right\}$

UCB for General Function Class

- **Estimation of $\widehat{R}_t$:** $\widehat{R}_t = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \sum_{i=1}^{t-1} (f(x_i, a_i) - r_i)^2$ (Regression)
- **Confidence set:** $\mathcal{F}_t = \left\{ f \in \mathcal{F}: \ \sum_{i=1}^{t-1} \left(f(x_i, a_i) - \widehat{R}_t(x_i, a_i) \right)^2 \le \beta \right\}$
- **Decision:** $a_t = \underset{a}{\operatorname{argmax}} \max_{f \in \mathcal{F}_t} f(x_t, a)$ (Constrained optimization over $\mathcal{F}$)

This algorithm works in theory, but not implementable in practice.
(It's also highly sub-optimal in some cases)

Russo and Van Roy. Eluder Dimension and the Sample Complexity of Optimistic Exploration. 2013.
Lattimore and Szepesvari. The End of Optimism? An Asymptotic Analysis of Finite-Armed Linear Bandits. 2016.

Other Solutions?

- Can we avoid solving the constrained optimization?
  - Yes. ϵ-greedy and Boltzmann exploration only needs $\hat{R}_t$
- However…
  - ϵ-greedy is non-adaptive and sub-optimal
  - Boltzmann exploration (original form) does not have theoretical guarantee
- It turns out there is an adaptive exploration scheme that has near-optimal regret bound, without explicitly quantifying the uncertainty of $\hat{R}_t$

SquareCB

SquareCB (Parameter: γ)

At round t, receive x_t, and obtain $\hat{R}_t$ from *any regression procedure.*

Define $\mathrm{Gap}_t(a) = \max_{b \in \mathcal{A}} \hat{R}_t(x_t, b) - \hat{R}_t(x_t, a)$ and

$$p_t(a) = \frac{1}{\lambda + \gamma \mathrm{Gap}_t(a)},$$

(Inverse Gap Weighting)

where $\lambda \in (0, A]$ is a normalization factor that makes p_t a distribution.

Sample $a_t \sim p_t$ and receive $r_t = R(x_t, a_t) + w_t$.

Foster and Rakhlin. Beyond UCB: Optimal and Efficient Contextual Bandits with Regression Oracles. 2020.

SquareCB

Regret Bound of SquareCB

SquareCB ensures

$$\mathbb{E}[\text{Regret}] \le O\left(\sqrt{AT\mathbb{E}\left[\sum_{t=1}^{T}\left(\hat{R}_t(x_t, a_t) - R(x_t, a_t)\right)^2\right]}\right).$$

If the function class $\mathcal{F}$ is finite, it's possible to ensure

$$\sum_{t=1}^{T}\left(\hat{R}_t(x_t, a_t) - R(x_t, a_t)\right)^2 \le \log|\mathcal{F}|.$$