# Contextual Bandits with Non-Linear / General Reward

Chen-Yu Wei

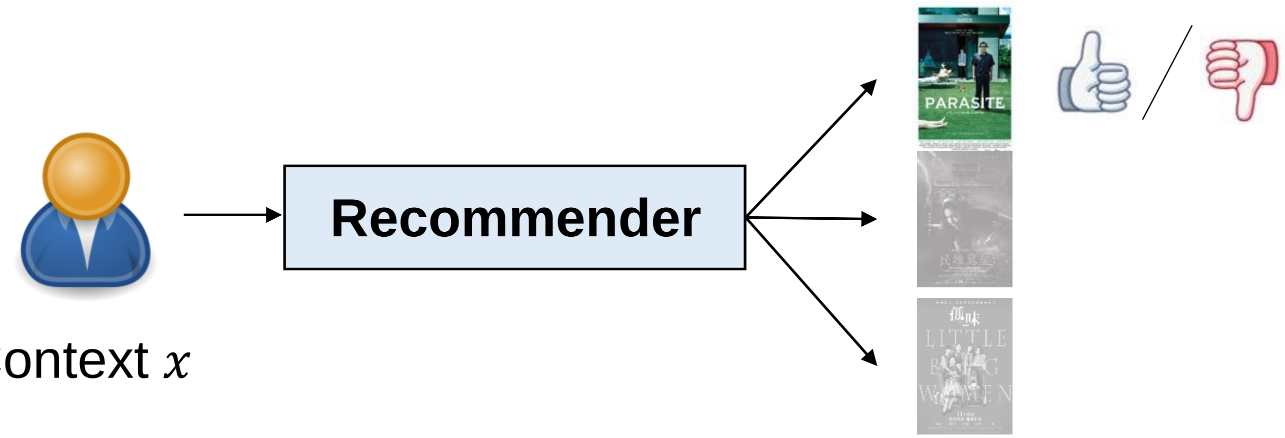
# Topics

- Generalized linear contextual bandits
- Reductions 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.

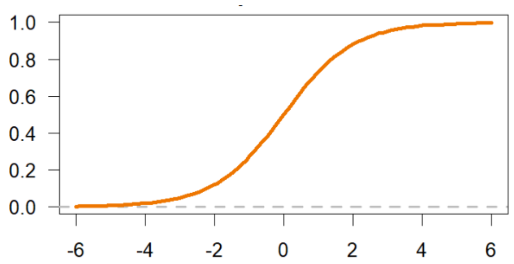



Context $x$

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

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 $\phi: \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$

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

# 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., $\epsilon$-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 $\Theta_t$)
  - 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**: $\theta_t \sim \text{dist. over } \Theta_t, \quad a_t = \operatorname*{argmax}_a R_{\theta_t}(x_t, a)$

# 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 $$\hat{\theta}_t = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{t-1} (\phi_i^\top \theta - r_i)^2 + \lambda \|\theta\|^2$$

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

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:** $\tilde{O}(d\sqrt{T})$


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 $\mu$

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 $\widehat{R}_t$:** $\widehat{R}_t = \operatorname*{argmin}_{f\in\mathcal{F}} \sum_{i=1}^{t-1} (f(x_i, a_i) - r_i)^2$ (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:** $a_t = \operatorname*{argmax}_a \max_{f\in\mathcal{F}_t} f(x_t, a)$ (Constrained optimization over $\mathcal{F}$)

It's theoretically sub-optimal in some cases (unlike in MAB and LinearCB)

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.

# Realizing UCB for General Function Class

$$\mathcal{F}_t = \left\{ f \in \mathcal{F} : \quad \sum_{i=1}^{t-1} \left( f(x_i, a_i) - \hat{R}_t(x_i, a_i) \right)^2 \le \beta \right\}$$

$$a_t = \operatorname*{argmax}_a \max_{f \in \mathcal{F}_t} \; f(x_t, a)$$

$$\min_{f \in \mathcal{F}} \min_a \underbrace{\sum_{i=1}^{t-1} \left( f(x_i, a_i) - \hat{R}_t(x_i, a_i) \right)^2}_{1} - \boldsymbol{\lambda} \underbrace{f(x_t, a)}_{2}$$

$\lambda \uparrow \Rightarrow 1 \uparrow \; 2 \uparrow$

$\lambda \downarrow \Rightarrow 1 \downarrow \; 2 \downarrow$

Binary search for $\lambda$ such that $1 \approx \beta$

# RegCB

Foster et al. Practical contextual bandits with regression oracles. 2018.

$$\mathcal{F}_t = \left\{ f \in \mathcal{F}: \quad \sum_{i=1}^{t-1} (f(x_i, a_i) - r_i)^2 - \sum_{i=1}^{t-1} \left( \widehat{R}_t(x_i, a_i) - r_i \right)^2 \le \beta \right\}$$

Another theoretically feasible way to construct the confidence set.


$$a_t = \operatorname*{argmin}_a \min_{f \in \mathcal{F}_t} \; (f(x_t, a) - 2)^2$$

$$\min_{f \in \mathcal{F}} \min_a \underbrace{\sum_{i=1}^{t-1} (f(x_i, a_i) - r_i)^2}_{1} + \lambda \underbrace{(f(x_t, a) - 2)^2}_{2}$$

Exactly a "regression problem" (with one artificial sample)

Binary search for $\lambda$ such that 1 $\approx \sum_{i=1}^{t-1} (\hat{R}_t(x_i, a_i) - r_i)^2 + \beta$

# Other Solutions?

- Can we avoid solving the constrained optimization?
  - Yes. $\epsilon$-greedy and Boltzmann exploration only needs $\hat{R}_t$
- However…
  - $\epsilon$-greedy is non-adaptive and sub-optimal
  - Boltzmann exploration (original form) does not have good 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

Boltzmann

$$p_t(a) = \frac{\exp(-\lambda \, \text{Gap}_t(a))}{\sum_{a'} \exp(-\lambda \, \text{Gap}_t(a'))}$$

**SquareCB** (Parameter: $\gamma$)

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

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

$b_t$

$$p_t(a) = \frac{1}{\lambda + \gamma \text{Gap}_t(a)}, \qquad \text{(Inverse Gap Weighting)}$$

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

$$\sum_t \left( \frac{A}{\gamma} + \gamma \left( \quad \right)^2 \right) = \boxed{\frac{AT}{\gamma} + \gamma \sum_t \mathbb{E}\left[\left(\hat{R}_t(x_t,a_t) - R(x_t,a_t)\right)^2\right]}$$

## Regret Bound of SquareCB

SquareCB ensures

optimal $\gamma$

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

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

$$\frac{a}{\gamma} + \gamma b \geq 2\sqrt{ab}$$

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

# Regret Analysis for SquareCB

$b_t = \arg\max_a \hat{R}_t(x_t, a)$

$a_t^* \triangleq \arg\max_a R(x_t, a)$

$$\mathbb{E}\,\text{Regret} = \sum_{t=1}^{T} \left( R(x_t, a_t^*) - \sum_{a=1}^{A} p_t(a) R(x_t, a) \right)$$

$$= \underbrace{\sum_{t=1}^{T} \left( \hat{R}_t(x_t, a_t^*) - \sum_{a=1}^{A} p_t(a) \hat{R}_t(x_t, a) \right)}_{①} + \underbrace{\sum_t \left( R(x_t, a_t^*) - \hat{R}_t(x_t, a_t^*) \right)}_{②} + \underbrace{\sum_t \sum_a p_t(a) \left( \hat{R}_t(x_t, a) - R(x_t, a) \right)}_{③}$$

$$① = \sum_{a=1}^{A} p_t(a) \left( \hat{R}_t(x_t, b_t) - \hat{R}_t(x_t, a) \right) - \left[ \hat{R}_t(x_t, b_t) - \hat{R}_t(x_t, a_t^*) \right]$$

$$= \sum_a p_t(a)\, \text{Gap}_t(a) - \text{Gap}(a_t^*) = \sum_a \frac{\text{Gap}_t(a)}{\lambda + \gamma \text{Gap}_t(a)} - \text{Gap}_t(a_t^*) \leq \underline{\frac{A}{\gamma} - \text{Gap}_t(a_t^*)}$$

$$\text{IA} \quad \sum_a \frac{\text{Gap}_t(a)}{\gamma \text{Gap}_t(a)}$$

$$③ = \sum_a p_t(a)\left(\hat{R}_t(x_t,a) - R(x_t,a)\right)$$

$$\le \frac{1}{2}\sum_a p_t(a)\left(\frac{1}{\gamma} + \gamma\left(\hat{R}_t(x_t,a) - R(x_t,a)\right)^2\right)$$

$$\le \frac{1}{2\gamma} + \frac{1}{2}\gamma \sum_a p_t(a)\left(\hat{R}_t(x_t,a) - R(x_t,a)\right)^2$$

$$= \frac{1}{2\gamma} + \frac{1}{2}\gamma\, \mathbb{E}\left[\left(\hat{R}_t(x_t,a_t) - R(x_t,a_t)\right)^2\right]$$

AM-GM: $\cancel{ab \le \frac{1}{2}(a^2+b^2)}$

$$\sqrt{ab} = \frac{1}{2}(a+b)$$

$$\boxed{\phantom{x}} \le \frac{1}{2}\left(\frac{1}{\gamma} + \gamma \boxed{\phantom{x}}^2\right)$$

$$\boxed{\phantom{x}} \le \frac{1}{2}\left(\frac{1}{\gamma p_t(a_t^*)} + \gamma p_t(a_t^*) \boxed{\phantom{x}}^2\right)$$

$$② = R(x_t,a_t^*) - \hat{R}_t(x_t,a_t^*)$$

$$\le \underbrace{\frac{1}{2\gamma p_t(a_t^*)}} + \underbrace{\frac{\gamma p_t(a_t^*)}{2}\left(R(x_t,a_t^*) - \hat{R}_t(x_t,a_t^*)\right)^2}$$

AM-GM

$$\le \frac{\gamma}{2}\mathbb{E}\left[\left(\hat{R}_t(x_t,a_t) - R(x_t,a_t)\right)^2\right]$$

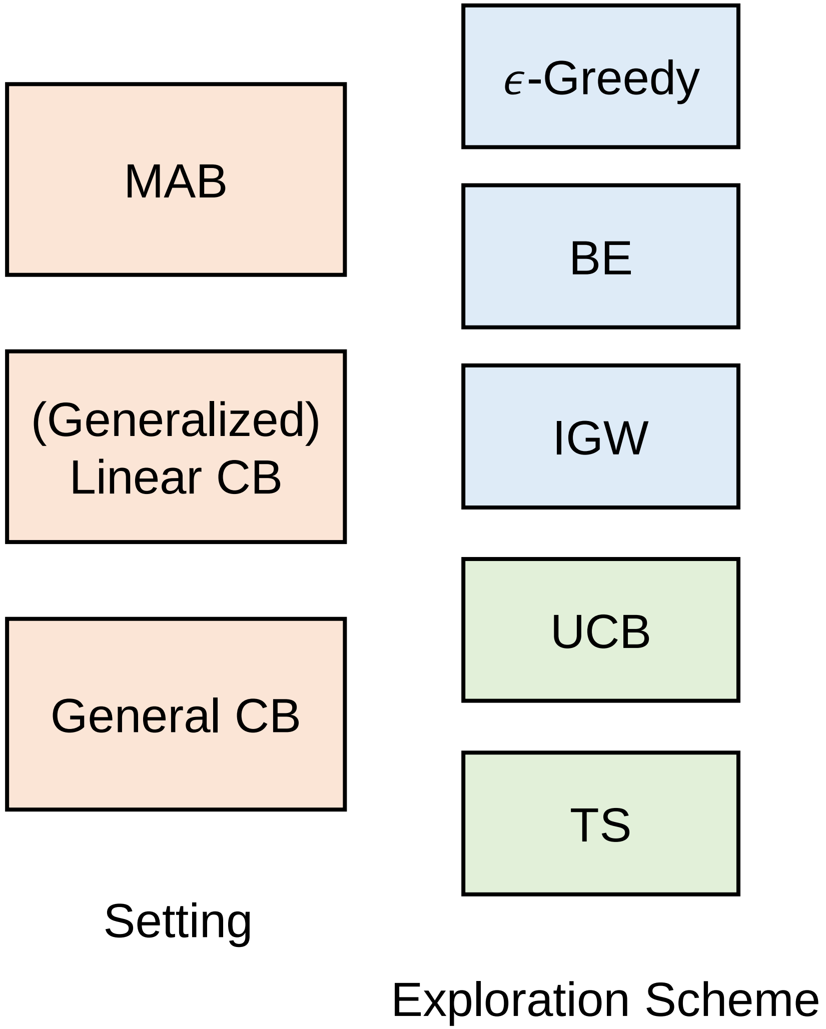
$$\frac{1}{2\gamma p_t(a_t^*)} = \frac{\lambda + \gamma \mathrm{Gap}_t(a_t^*)}{2\gamma}$$

$$\le \frac{A}{2\gamma} + \frac{1}{2}\mathrm{Gap}_t(a_t^*)$$

$$①+②+③ \le O\left(\frac{A}{\gamma} + \gamma\, \mathbb{E}\left[\left(\hat{R}_t(x_t,a_t) - R(x_t,a_t)\right)^2\right]\right)$$

# Summary for Bandits/Contextual Bandits

# What we have discussed so far:  Exploration




**Common Idea:** Regression (SL) + Exploration

$$\min_{\hat{R}_t} \sum_{i=1}^{t-1} \left(\hat{R}_t(x_i, a_i) - r_i\right)^2$$

# What we have discussed so far: Exploration

MAB

(Generalized) Linear CB

General CB

Setting

$\epsilon$-Greedy

BE

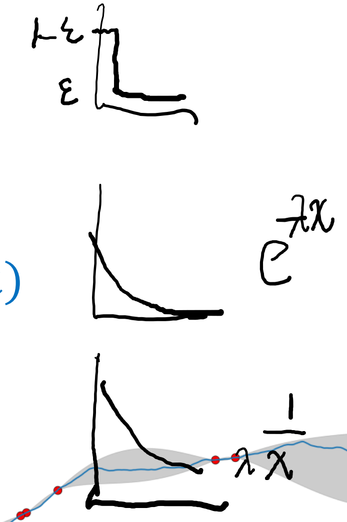
IGW

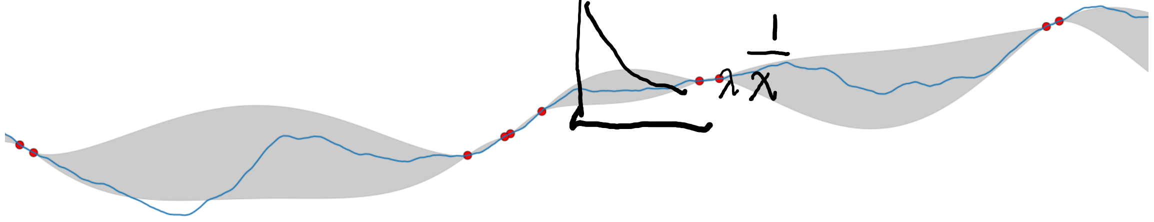
UCB

TS

Exploration Scheme

Only based on $\hat{R}_t(x, a)$



Based on $\hat{R}_t(x, a)$ + measure on the uncertainty of $\hat{R}_t(x, a)$

# Course Content

(Focusing on exploration-exploitation tradeoff)

## Part I. Learning in Bandits

- Multi-armed bandits
- Linear bandits
- Contextual bandits
- Adversarial multi-armed bandits
- Adversarial linear bandits

## Part II. Basics of MDPs

- Bellman (optimality) equations
- Value iteration
- Policy iteration

(Focusing on credit assignment and distribution mismatch)

## Part III. Learning in MDPs

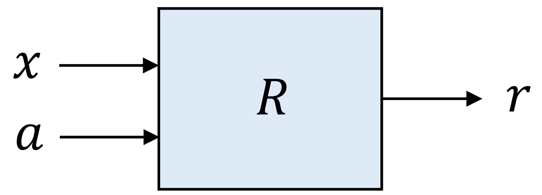
- Approximate value iteration and variants
  - Least-square value iteration
  - Q-Learning
  - DQN
- Policy evaluation
  - Temporal difference
  - Monte Carlo
- Approximate policy iteration and variants
  - Least-square policy iteration
  - (Natural) policy gradient and actor-critic
  - REINFORCE, A2C, PPO, SAC
  - DDPG

## Part IV. Offline RL

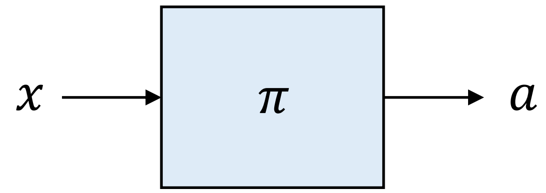
## Student Project Presentation

# Another Class of Bandit Algorithms

- So far, we have focused on **value-centric** approaches
  - Policies are *derived* from the value estimations




- **Policy-centric** approaches perform direct updates on the polices




- Policy-centric approaches have stronger theoretical guarantees for **non-stationary environments**

- As a warmup, we will start from studying **full-information** feedback