Some Recent Advances in Bandit Theory

Chen-Yu Wei 魏振宇

Postdoc @ UC Berkeley

Learning to Make Decisions

Machine Learning ≈ Looking for a Function

$$x \longrightarrow \boxed{f} \longrightarrow y$$

feature　　　　　　　　label

$f($  $) =$ Cat　　(classification)

$f($ 氣溫、濕度、風速… $) =$ 1000mm 降雨量　　(regression)

from 李宏毅老師 “機器學習”

Decision Making Problem

Problems where the **learner's behavior** affects the **feedback.**

Learning to Make Decisions

$$x \longrightarrow \boxed{f} \longrightarrow a$$

feature | action

f (

movie watching history

) =

recommendation

Reduction from **decision-making** to **classification**

Learning to Make Decisions

feature x $\rightarrow$ $\boxed{f}$ $\rightarrow$ r reward

action a $\rightarrow$ $\boxed{f}$

f (

, 

) $= 7.5$

movie watching history

recommendation

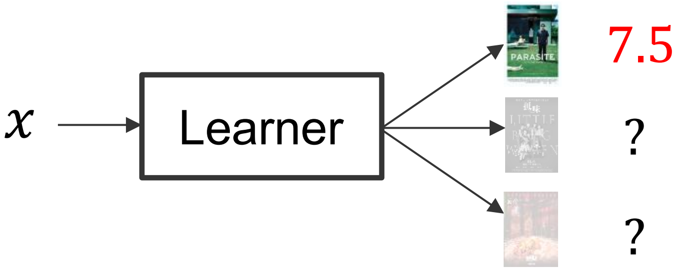
score

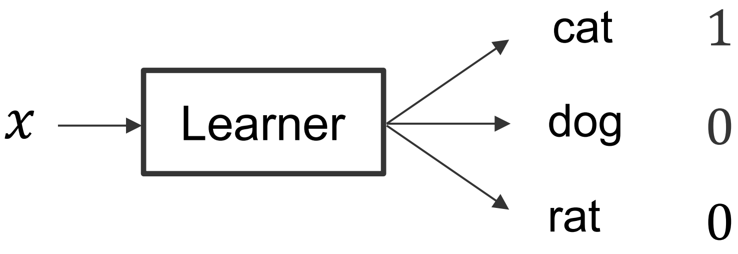
Reduction from decision-making to regression

Reduction from **decision-making** to **regression**

Challenges in Decision Problems (1/2)

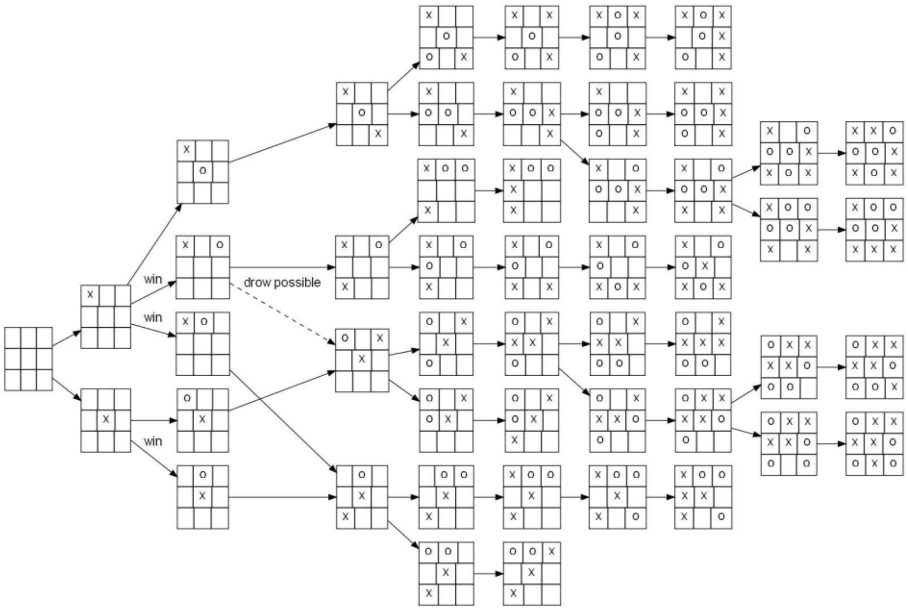
1. Partial feedback → Need to **explore**

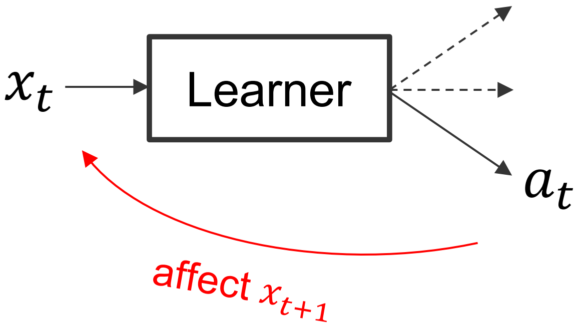




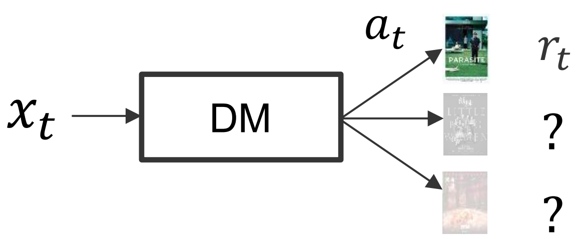
Challenges in Decision Problems (2/2)

2. Delayed feedback / long-term dependency → Need to do **credit assignment**





Today, we will consider problems with <u>partial feedback</u> but no long-term dependency.



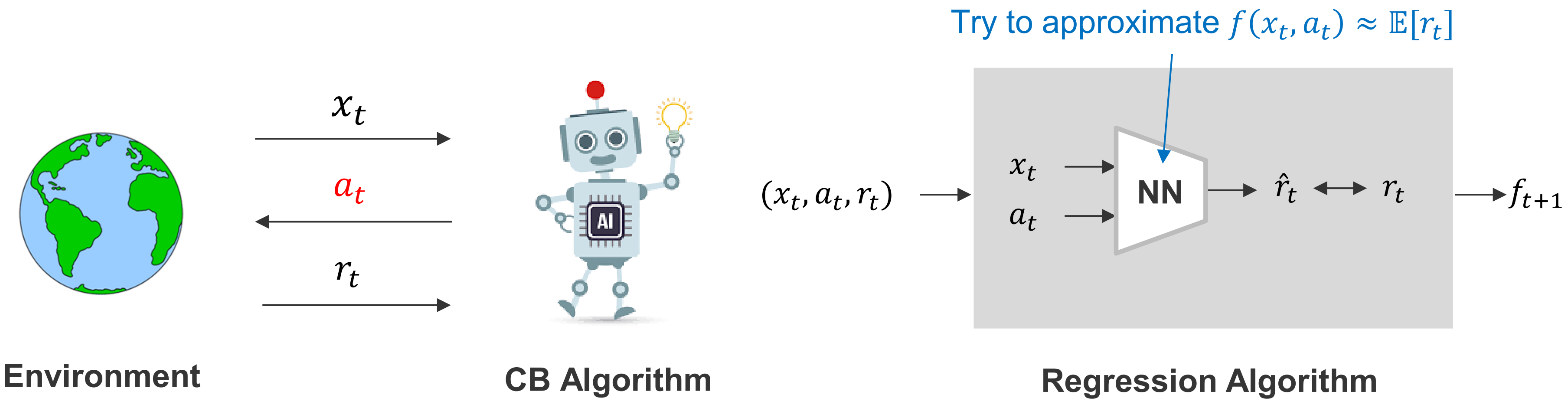
A simple protocol **(Contextual Bandit)**

For $t = 1, 2, \ldots$

- Environment generates a **context** x_t (user profile, watching list)
- Decision-maker takes an **action** a_t (a movie)
- Environment reveals a **reward** r_t (user scoring)
  (r_t depends on x_t and a_t)
  (a_t does NOT affect x_{t+1})

Supervised Learning	$<$	**Contextual Bandits**	$\ll$	Reinforcement Learning
(full feedback)		(partial feedback)		(partial + delayed feedback)

Reduction to Regression



Greedy: $a_t = \text{argmax}_a \; f_t(x_t, a)$

Failure of the Greedy Strategy

Adding Exploration

- ϵ-greedy

$$a_t = \begin{cases} \text{argmax}_a \ f_t(x_t, a) & \text{w. p.} \ \ 1 - \epsilon \\ \text{uniformly random} & \text{w. p.} \ \ \epsilon \end{cases}$$

- Boltzmann exploration

$$a_t \sim p_t(a) = \frac{\exp\left(\gamma \cdot f_t(x_t, a)\right)}{\sum_{a'} \exp\left(\gamma \cdot f_t(x_t, a')\right)}$$

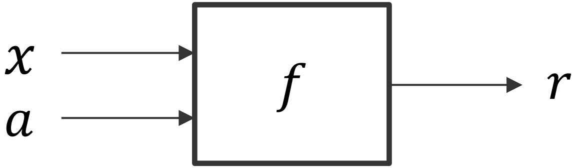
李宏毅老師 “DRL: Q-learning”

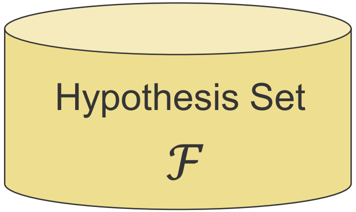
How good are they?

Today’s mission: show an *optimal* exploration strategy for CB.

Some Theory

Formalization





Assumption: $\exists\, f^{\star} \in \mathcal{F}$, s.t. $\mathbb{E}[r] = f^{\star}(x, a)$

Optimal policy: choose $\mathrm{argmax}_a\, f^{\star}(x_t, a)$

∴ Performance Gap (a.k.a. **Regret**)

$$= \underbrace{\sum_{t=1}^{T} \max_a f^{\star}(x_t, a)}_{\text{Optimal policy's expected total reward}} - \underbrace{\sum_{t=1}^{T} f^{\star}(x_t, a_t)}_{\text{Learner's expected total reward}}$$

Optimal policy's
expected total reward

Learner's
expected total reward

The goal of the learner:
Minimize Regret

Strategy and Its Interpretation

output from the regression algorithm

$$p_t = \min_{p} \max_{f \in \mathcal{F}} \left\{ \underbrace{\max_{a'} f(x_t, a) - \mathbb{E}_{a \sim p}[f(x_t, a)]}_{1} - \gamma \underbrace{\mathbb{E}_{a \sim p}\left[\left(f(x_t, a) - f_t(x_t, a)\right)^2\right]}_{2} \right\}$$

3 4

1 = **regret** in round t (supposed that $f^\star = f$)

2 = **information gain** in round t (supposed that $f^\star = f$)

3 Minimize regret and maximize information gain simultaneously

4 Consider the worst-case $f^\star$

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

Simplification

It suffices to use the following:
(the exact solution of the min-max program assuming $\mathcal{F}(x_t,\cdot) = \mathbb{R}^A$)

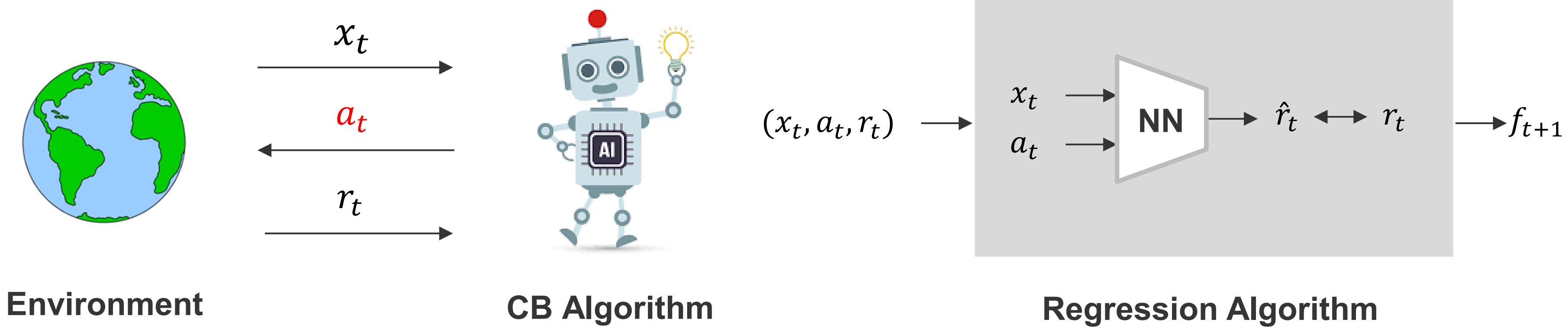
$$\boxed{p_t(a) = \frac{1}{\lambda + \gamma\underbrace{\left(\max_{a'} f_t(x_t,a') - f_t(x_t,a)\right)}_{\text{sub-optimality gap of action } a \text{ predicted by } f_t}}}$$

(Inverse-Gap Weighting)

normalization factor making $\sum_a p_t(a) = 1$

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

Reduction to Regression



IGW: $$a_t \sim p_t(a) = \frac{1}{\lambda + \gamma \left(\max_{a'} f_t(x_t, a') - f_t(x_t, a)\right)}$$

Theorem by Foster and Rakhlin (2020)

With Inverse-Gap Weighting,

$$\mathrm{Regret}_{\mathrm{CB}} \le \frac{AT}{\gamma} + \gamma \underbrace{\sum_{t=1}^{T} \left(f_t(x_t, a_t) - f^\star(x_t, a_t)\right)^2}_{\substack{\text{Performance of the regression algorithm} \\ \text{Can be } O(\log|\mathcal{F}|)}} \le O\left(\sqrt{AT \log|\mathcal{F}|}\right)$$

Choose a suitable γ

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

Exploration Strategies

- ϵ-greedy

$$a_t = \begin{cases} \text{argmax}_a \ f_t(x_t, a) & \text{w.p.} \ \ 1-\epsilon \\ \text{uniformly random} & \text{w.p.} \ \ \epsilon \end{cases}$$

- Boltzmann exploration

Cesa-Bianchi, Gentile, Lugosi, Neu.
Boltzmann exploration done right. NeurIPS 2017.

$$a_t \sim p_t(a) \propto \exp\left(\gamma \cdot f_t(x_t, a)\right) \propto \exp\left(-\gamma \cdot \text{Gap}(a)\right)$$

- Inverse-gap weighting (optimal reduction from Contextual Bandit to Regression)

$$a_t \sim p_t(a) = \frac{1}{\lambda + \gamma \cdot \ \text{Gap}(a)}$$

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

Milestones in the Theory of Contextual Bandits

Abe and Long. Associative reinforcement learning using linear probabilistic concepts. 1999.

Langford and Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. 2007.
(coin down the name *contextual bandit*)

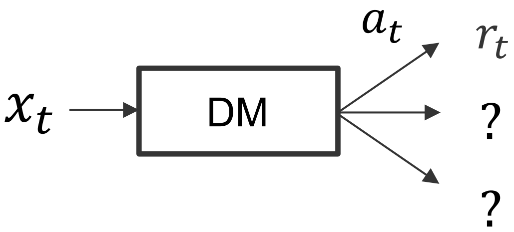
Reduction to classification:

- Dudik, et al. Efficient optimal learning for contextual bandits. 2011.
  (optimal regret)
- Agarwal et al. Taming the monster: a fast and simple algorithm for contextual bandits. 2014.
  (computationally efficient + optimal regret)

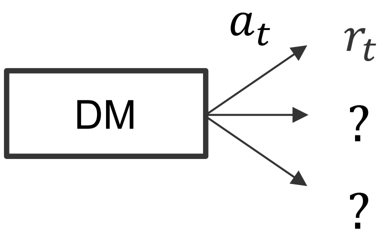
Reduction to regression:

- Agarwal et al. Contextual bandit learning with predictable rewards. 2012.
  (optimal regret)
- Foster and Rakhlin. Beyond UCB: optimal and efficient contextual bandits with regression oracles. 2020.
  (computationally efficient + optimal regret)

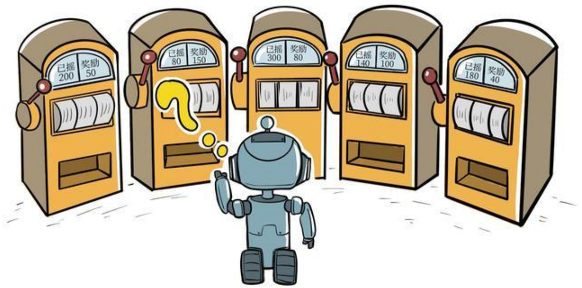
Robust and Adaptive Bandit Algorithms

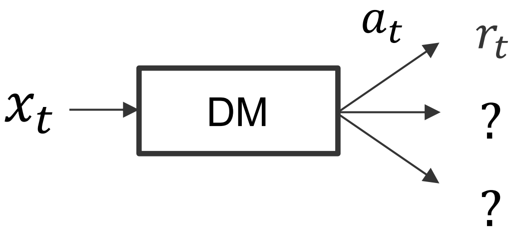


Contextual Bandits

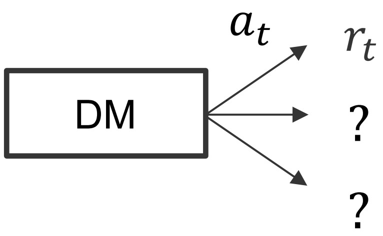


Multi-Armed Bandits (多臂吃角子老虎機)





Contextual Bandits



Multi-Armed Bandits (多臂吃角子老虎機)

For $t = 1, 2, \ldots$

Decision-maker takes an **action** a_t

Environment reveals a **reward** r_t

Two Reward Generation Processes (Figures from Wouter Koolen)

For $t = 1, 2, \ldots$

Environment decides **rewards** $r_t(a)$ for all actions a

Decision-maker takes an **action** a_t

Environment reveals a **reward** $r_t(a_t)$

fixed over time

$\mathbb{E}[r_t(a)] = f^\star(a)$

(stochastic world)

$r_t(a)$ are arbitrarily decided

(adversarial world)

$$\text{Regret} = \max_a \sum_{t=1}^{T} r_t(a) - \sum_{t=1}^{T} r_t(a_t)$$

Why consider adversarial worlds?

(Robustness)

李彥寰老師 “預測、學習與賽局”

Development of Stochastic and Adversarial MAB

Thompson 1933

Lai and Robbins 1985

Agrawal 1995

Auer et al. 2002 $O\left(\frac{A \log T}{\Delta}\right)$ **optimal**

Auer et al. 2002

Audibert and Bubeck, 2009 $O\left(\sqrt{AT}\right)$ **optimal**

$\Delta =$ difference between the expected reward of the best and the second-best action

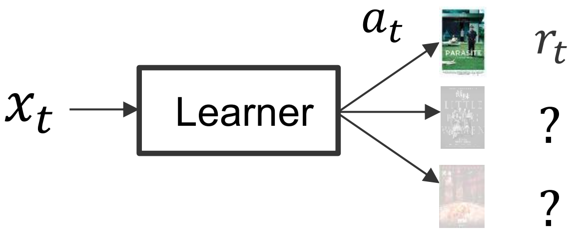
Bubeck and Slivkins 2012: Is there a **single algorithm** with optimal regret in both worlds?

The **best-of-both-world** problem, **“robust and adaptive”** learning

Development of Stochastic and Adversarial MAB

	Regret		Strategy
			$\mathrm{Gap}(a)$: current estimation for the sub-optimality gap of action a
Auer et al., 2002 **(classic algorithm for adversarial MAB)**	$\sqrt{AT\log A}$	$\sqrt{AT\log A}$	$p(a) = \mathrm{softmax}(a) \propto e^{-\gamma\cdot\mathrm{Gap}(a)}$
Seldin and Slivkins, 2014	$\frac{A\log^2 T}{\Delta} + \frac{A^2}{\Delta^3}$	$\sqrt{AT\log A}$	$p(a) \propto \mathrm{softmax}(a) + \min\left\{\frac{1}{\sqrt{At}}, \frac{1}{t\,\mathrm{Gap}(a)^2}\right\}$
W and Luo, 2018	$\boldsymbol{\frac{A\log T}{\Delta}}$	$\sqrt{AT\log T}$	$p(a) = \frac{1}{\lambda + \gamma_t \mathrm{Gap}(a)}$
Zimmert and Seldin, 2019	$\boldsymbol{\frac{A\log T}{\Delta}}$	$\boldsymbol{\sqrt{AT}}$	$p(a) = \frac{1}{\left(\lambda + \sqrt{t}\,\mathrm{Gap}(a)\right)^2}$

Summary



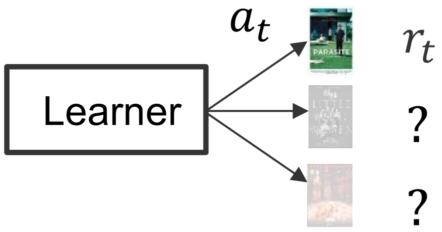
Contextual Bandit

Assumption: $\mathbb{E}[r_t] = f^\star(x_t, a_t)$

Can solve it by regression + exploration strategy

(Inverse Gap Weighting)

$$\text{Regret} \leq \frac{AT}{\gamma} + \gamma \cdot \text{regression\_error}$$



Multi-Armed Bandit

Either adversarial or stochastic ($\mathbb{E}[r_t] = f^\star(a_t)$)

Can get robustness and adaptivity (best-of-both-world) by **Inverse Squared Gap Weighting**

$\text{Regret} \leq \sqrt{AT}$ (adversarial), or $\frac{A \log T}{\Delta}$ (stochastic)