Bandits 1

Chen-Yu Wei

Contextual Bandits and Non-Contextual Bandits

Context

Contextual Bandit

a_1 ?
$\boldsymbol{a_2}$ $\mathbf{0.5}$
a_3 ?

Policy: $\pi(a|\text{context})$

Non-Contextual Bandit

a_1 ?
$\boldsymbol{a_2}$ $\mathbf{0.5}$
a_3 ?

Policy: $\pi(a)$

Multi-Armed Bandits

Multi-Armed Bandits

A slot machine

One-armed bandit

A row of slot machines

Multi-armed bandit

Multi-Armed Bandits

Given: arm set $\mathcal{A} = \{1, \dots, A\}$

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

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

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

Arm = Action

Assumption: $R(a)$ is the (hidden) ground-truth reward function

w_t is a zero-mean noise

Goal: maximize the total reward $\sum_{t=1}^{T} R(a_t)$ (or $\sum_{t=1}^{T} r_t$)

How to Evaluate an Algorithm's Performance?

- "My algorithm obtains $0.3T$ total reward within T rounds"
  – Is my algorithm good or bad?

$\Rightarrow \max_a R(a) - \frac{1}{T}\sum_{t=1}^{T} R(a_t) \leq \frac{1}{\sqrt{T}}$

- Benchmarking the problem

$\leq \sqrt{T}$

$$\text{Regret} := \underbrace{\max_{\pi} \sum_{t=1}^{T} R(\pi)}_{\text{The total reward of the best policy}} - \sum_{t=1}^{T} R(a_t) \quad \underset{\text{In MAB}}{=} \quad \max_a TR(a) - \sum_{t=1}^{T} R(a_t)$$

- "My algorithm ensures $\text{Regret} \leq 5T^{\frac{3}{4}}$ "
- $\text{Regret} = o(T) \Rightarrow$ the algorithm is as good as the optimal policy asymptotically

Multi-Armed Bandits

- Key challenge: Exploration
- The other three challenges we will discuss for RL
  - Generalization (there is no input in MAB)
  - Temporal credit assignments (there is no delayed feedback)
  - Distribution mismatch (there is no pre-collected data)

- We will discuss about two categories of exploration strategies
  - Based on mean estimation
  - Based on mean and uncertainty estimation

Multi-Armed Bandits

Based on mean estimation

The Exploration and Exploitation Trade-off in MAB

- To perform as well as the best policy (i.e., best arm) asymptotically, the learner has to pull the best arm most of the time

  ⇒ need to **exploit**

- To identify the best arm, the learner has to try every arm sufficiently many times

  ⇒ need to **explore**

A Simple Strategy: Explore-then-Exploit

Explore-then-exploit (Parameter: T_0)

In the first T_0 rounds, sample each arm T_0/A times. **(Explore)**

Compute the **empirical mean** $\hat{R}(a)$ for each arm a

In the remaining $T - T_0$ rounds, draw $\hat{a} = \text{argmax}_a \hat{R}(a)$ **(Exploit)**

What is the *right* amount of exploration (T_0)?

Another Simple Strategy: ϵ-Greedy

Mixing exploration and exploitation in time

ϵ-Greedy (Parameter: ϵ)

In the first A rounds, draw each arm once.

In the remaining rounds $t > A$,

Take action

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

where $\hat{R}_t(a) = \frac{\sum_{s=1}^{t-1} \mathbb{I}\{a_s = a\}\, r_s}{\sum_{s=1}^{t-1} \mathbb{I}\{a_s = a\}}$ is the empirical mean of arm a using samples up to time $t - 1$.

Comparison

- ϵ-Greedy is more **robust to non-stationarity** than Explore-then-Exploit
- ϵ-Greedy has a better performance in the early phase of the learning process

Quantifying the Estimation Error

In the exploration phase, we obtain $N = T_0/A$ i.i.d. samples of each arm.

Key Question:

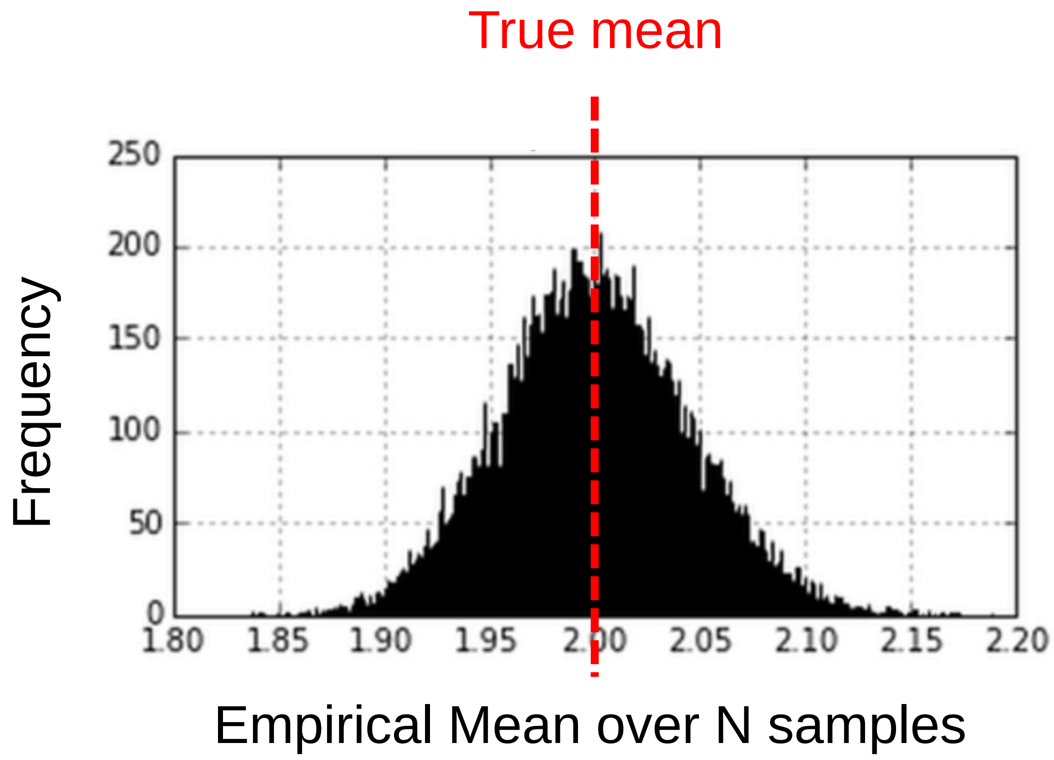
$$\left|\hat{R}(a) - R(a)\right| \leq ?\ f(N)$$

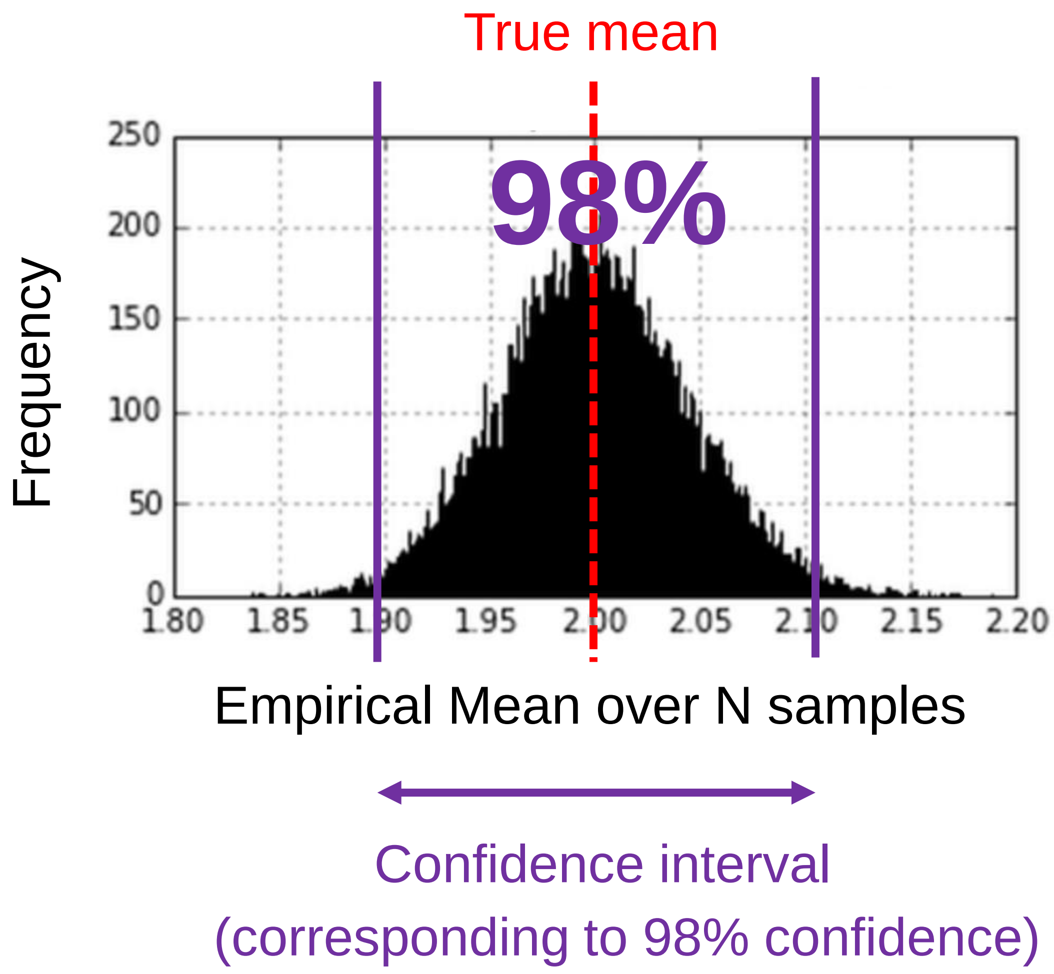
some decreasing function of N

Empirical mean of N i.i.d. samples

True mean

Quantifying the Estimation Error





Quantifying the Estimation Error

In the exploration phase, we obtain $N = T_0/A$ i.i.d. samples of each arm.

Key Question:

$$\left| \hat{R}(a) - R(a) \right| \leq ?\ f(N)$$

Empirical mean of N i.i.d. samples

True mean

some decreasing function of N

Quantifying the Estimation Error

In the exploration phase, we obtain $N = T_0/A$ i.i.d. samples of each arm.

Key Question:

With probability at least $1 - \underset{0.02}{\delta}$, $= 0.98$

$$\left| \hat{R}(a) - R(a) \right| \leq ? \ f(N, \delta)$$

Empirical mean of N i.i.d. samples

True mean

some decreasing function of N

Quantifying the Error: Concentration Inequality

Theorem. Hoeffding's Inequality

Let $X_1, \ldots, X_N$ be independent σ**-sub-Gaussian** random variables.

Then with probability at least $1 - \delta$,

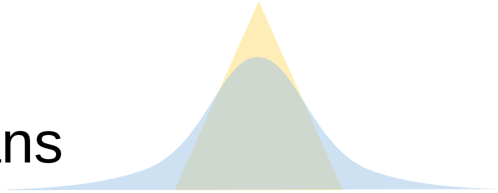
$$\left| \frac{1}{N} \sum_{i=1}^{N} X_i - \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[X_i] \right| \leq \sigma \sqrt{\frac{2 \log(2/\delta)}{N}} \quad .$$

A random variable is called σ-sub-Gaussian if $\mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X])}\right] \leq e^{\lambda^2 \sigma^2 / 2} \quad \forall \lambda \in \mathbb{R}$.

Fact 1. $\mathcal{N}(\mu, \sigma^2)$ is σ-sub-Gaussian.

Fact 2. A random variable $\in [a, b]$ is $(b - a)$-sub-Gaussian.

Intuition: tail probability $\Pr\{|X - \mathbb{E}[X]| \geq z\}$ bounded by that of Gaussians



Quantifying the Estimation Error

With probability at least $1-\delta$, $\left|\hat{R}(a)-R(a)\right| = O\left(\sqrt{\frac{\log(1/\delta)}{N}}\right)$

Omit constants

With high probability, $\left|\hat{R}(a)-R(a)\right| = \tilde{O}\left(\sqrt{\frac{1}{N}}\right)$

$\left|\hat{R}(a)-R(a)\right| \lesssim \sqrt{\frac{1}{N}}$

Omit constants and $\log(1/\delta)$ factors

Explore-then-Exploit Regret Bound Analysis

In the first T_0 rounds, sample each arm T_0/A times.

Compute the **empirical mean** $\hat{R}(a)$ for each arm a

In the remaining $T - T_0$ rounds, draw $\hat{a} = \mathrm{argmax}_a \hat{R}(a)$

$a^* = \mathrm{argmax}_a R(a)$ (true best arm)

After the exploration phase, we have $\boxed{|\hat{R}(a) - R(a)| \lesssim \sqrt{\frac{1}{N}} = \sqrt{\frac{A}{T_0}}}$

In the exploitation phase,

At any time $t \in$ exploitation phase, $R(a^*) - R(\hat{a})$

$$= \underbrace{\hat{R}(a^*) - \hat{R}(\hat{a})}_{\leq 0} + \underbrace{[R(a^*) - \hat{R}(a^*)]}_{\lesssim \sqrt{\frac{A}{T_0}}} + \underbrace{(\hat{R}(\hat{a}) - R(\hat{a}))}_{\sqrt{\frac{A}{T_0}}}$$

$$\text{Regret} \lesssim \text{cost of exploration} + \sum_{t \in \text{second phase}} (R(a^*) - R(\hat{a})) \lesssim T_0 + (T - T_0) \cdot 2\sqrt{\frac{A}{T_0}}$$

Regret Bound of Explore-then-Exploit and ϵ-Greedy

Theorem. Regret Bound of Explore-then-Exploit

Suppose that $R(a) \in [0,1]$ and w_t is 1-sub-Gaussian.

Then Explore-then-Exploit ensures with high probability,

$$\text{Regret} \lesssim T_0 + T\sqrt{\frac{A}{T_0}} \approx A^{1/3}T^{2/3} \text{ (choosing } T_0 = A^{1/3}T^{2/3}\text{)}$$

Theorem. Regret Bound of ϵ-Greedy (Your Exercise)

Suppose that $R(a) \in [0,1]$ and w_t is 1-sub-Gaussian.

Then ϵ-Greedy ensures with high probability,

$\varepsilon_t \approx \left(A/t\right)^{1/3}$

$$\text{Regret} \lesssim \epsilon T + \sqrt{\frac{AT}{\epsilon}} \approx A^{1/3}T^{2/3} \text{ (choosing } \epsilon = \left(\frac{A}{T}\right)^{1/3}\text{)}$$

Can We Do Better?

In explore-then-exploit and ϵ-greedy, the probability to choose arms do not depend on the estimated mean (except for the empirically best arm).

… Maybe, the probability of choosing arms can be adaptive to the estimated mean?

Solution: Refine the amount of exploration for each arm **based on the current mean estimation**.

(Has to do this carefully to avoid **under-exploration**)

Refined Exploration

$\boxed{1 \le \gamma_t' \le A}$

$\boxed{\gamma_t' = A \Rightarrow \pi_t(a) \le \frac{1}{A}}$

$\gamma_t' = 1 \Rightarrow \pi_t(b) = 1$

where $b = \arg\max_b \hat{R}_t(b)$

Boltzmann Exploration (Parameter: λ)

In each round, sample a_t according to

$$\begin{cases} \gamma_t' = A \Rightarrow \sum_a \pi_t(a) \le 1 \\ \gamma_t' = 1 \Rightarrow \sum_a \pi_t(a) \ge 1 \end{cases}$$

$$\pi_t(a) \propto \exp\left(\lambda \hat{R}_t(a)\right)$$

$$= e^{\lambda \hat{R}_t(a)} / \left(\sum_b e^{\lambda \hat{R}_t(b)}\right)$$

$$\pi_t(a) = \frac{1}{A + \lambda \mathrm{Gap}_t(a)} \le \frac{1}{A}$$

where $\hat{R}_t(a)$ is the empirical mean of arm a using samples up to time $t-1$.

Inverse Gap Weighting (Parameter: λ)

γ_t is a normalization factor that makes $\sum_a \pi_t(a) = 1$

$$\pi_t(a) = \frac{1}{\gamma_t - \lambda \hat{R}_t(a)} = \frac{1}{\gamma_t' + \lambda \mathrm{Gap}_t(a)}$$

where $\mathrm{Gap}_t(a) = \max_b \hat{R}_t(b) - \hat{R}_t(a) \ge 0$

$\boxed{\gamma_t = \gamma_t' + \max_b \hat{R}_t(b)}$

Refined Exploration

Variant of Inverse Gap Weighting Easier for Implementation (Parameter: λ)

$$\pi_t(a) = \begin{cases} \dfrac{1}{A + \lambda \mathrm{Gap}_t(a)} & \text{if } a \neq \operatorname{argmax} \hat{R}_t(a) \\ 1 - \displaystyle\sum_{a' \neq a} \pi_t(a') & \text{if } a = \operatorname{argmax} \hat{R}_t(a) \end{cases}$$

where $\mathrm{Gap}_t(a) = \max_b \hat{R}_t(b) - \hat{R}_t(a)$

Refined Exploration

- Boltzmann Exploration
  - A quite commonly used exploration strategy (like ϵ-greedy)
  - However, it's theoretically less desirable. For fixed parameter $\lambda \geq 2\log t$, there is always a problem instance making BE suffer $\Theta(T)$ regret
  - There is no known regret bound for it yet (?)

Cesa-Bianchi, Gentile, Lugosi, Neu. Boltzmann Exploration Done Right, 2017.
Bian and Jun. Maillard Sampling: Boltzmann Exploration Done Optimally. 2021.

$\sqrt{AT} \leq A^{1/3} T^{2/3}$

when $A \leq T$

- Inverse Gap Weighting
  - Less well-known
  - We can show a near-optimal regret bound $\sqrt{AT}$ for it, improving the $A^{1/3}T^{2/3}$ by ϵ-greedy

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

Guarantee of Inverse Gap Weighting

Inverse Gap Weighting ensures with high probability,

$$\mathrm{Regret} \lesssim \frac{A}{\lambda} + \lambda \log T \quad \approx \sqrt{AT \log T} \text{ (choosing } \lambda = \sqrt{\frac{T}{A \log T}}\text{)}$$

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

See supplementary materials for a formal proof.

Summary: MAB Based on Mean Estimation

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

Design a distribution $\pi_t(\cdot)$ based on the current mean estimation $\hat{R}_t(\cdot)$

EG $\quad \pi_t(a) = (1-\epsilon)\mathbb{I}\{a = \operatorname{argmax} \hat{R}_t(\cdot)\} + \frac{\epsilon}{A}$ $\quad A^{1/3}T^{2/3}$

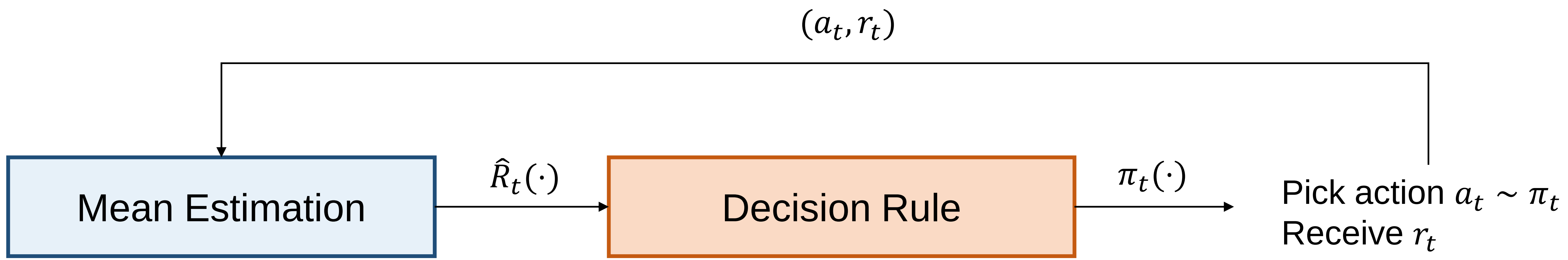
BE $\quad \pi_t(a) \propto \exp\left(\lambda \hat{R}_t(a)\right)$ $\quad \lambda_t$ increasing over t XXX

IGW $\quad \pi_t(a) = \frac{1}{\gamma_t - \lambda \hat{R}_t(a)}$ $\quad \sqrt{AT \log T}$

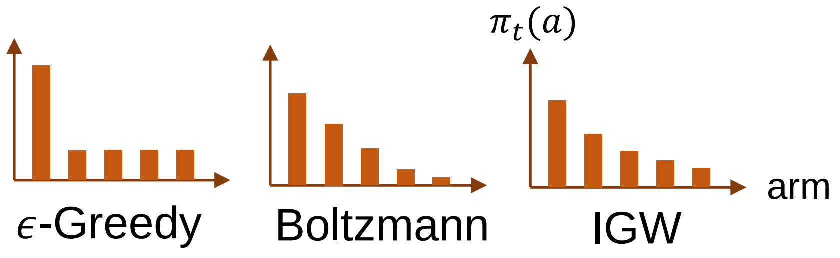
Sample an arm $a_t \sim \pi_t$ and receive the corresponding reward r_t.

Refine the mean estimation $\hat{R}_{t+1}(\cdot)$ with the new sample (a_t, r_t).

Summary: MAB Based on Mean Estimation



$$\hat{R}_t(a) = \frac{\sum_{s=1}^{t-1} \mathbb{I}\{a_s = a\}\, r_s}{\sum_{s=1}^{t-1} \mathbb{I}\{a_s = a\}}$$



$$\pi_t(a) = (1-\epsilon)\mathbb{I}\{a = \operatorname{argmax} \hat{R}_t(\cdot)\} + \frac{\epsilon}{A}$$

$$\pi_t(a) \propto \exp\big(\lambda \hat{R}_t(a)\big)$$

$$\pi_t(a) = \frac{1}{\gamma_t - \lambda \hat{R}_t(a)}$$

Summary: MAB Based on Mean Estimation

- All 3 methods are based on the same **mean estimation**
  - ϵ-Greedy, Boltzmann exploration, Inverse gap weighting

- The key difference is in the **decision rule**, i.e., the mapping from estimated means $\hat{R}_t$ to a distribution π_t.
  - The shape of the mapping makes differences

- There is a scalar hyperparameter that allows for a tradeoff between exploration and exploitation (ϵ in EG, λ in BE or IGW)

Some Experiments

$T = 10000$ rounds

$A = 2$ arms

Reward mean $R = [0.5, 0.5 - \Delta]$

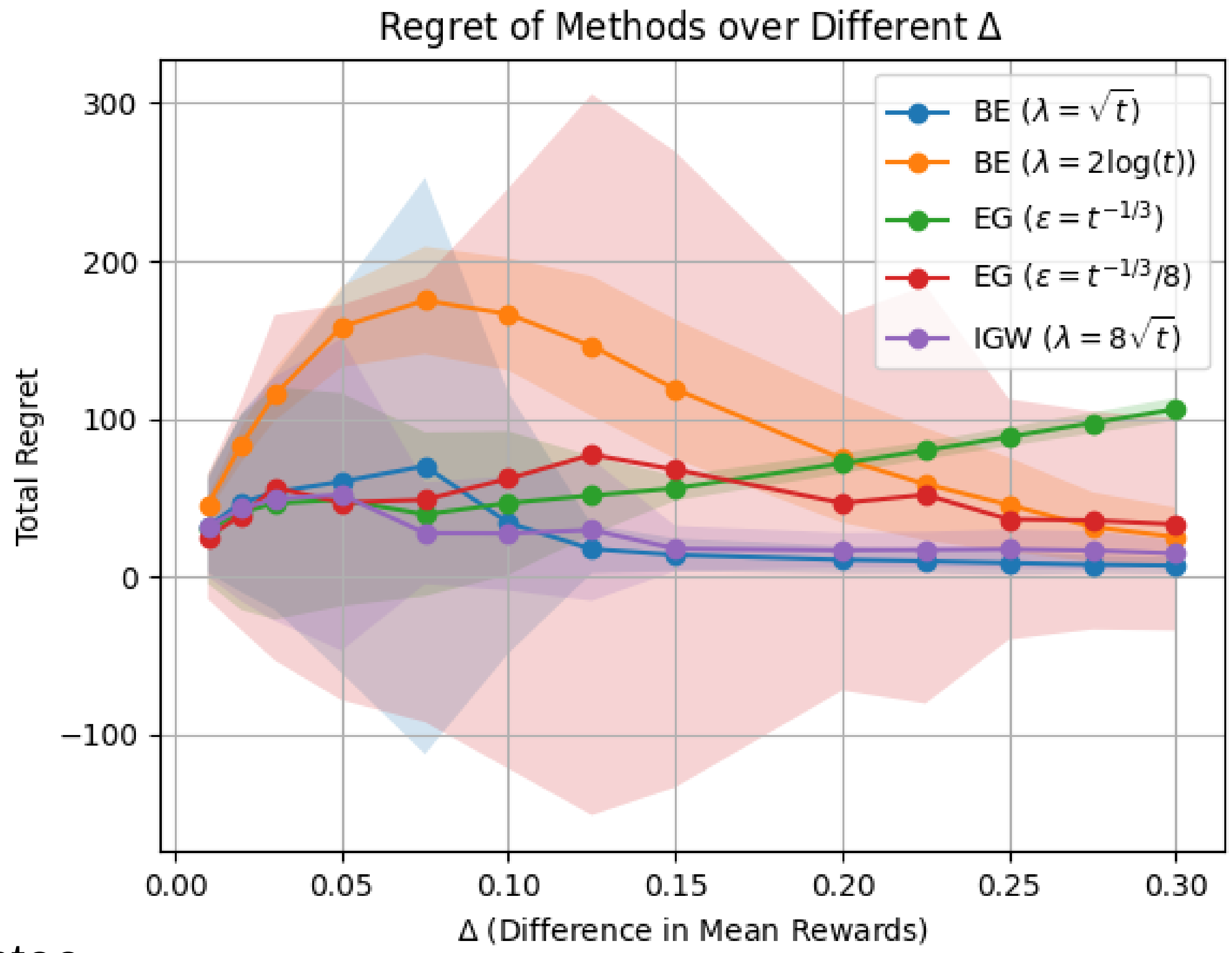
Bernoulli distribution

Time-dependent parameters

30 random seeds

Observations:

- Bound from theory could be loose
  -- theory captures **worst-case** guarantee
- Most algorithms seem to have its worst regret at some intermediate Δ value
  -- will be studied in Homework 1



Contextual Bandits

Based on reward function estimation

Multi-Armed Bandits vs. Contextual Bandits

Multi-Armed Bandit

4.2

?

?

all-user recommendation

Context → **Contextual Bandit**

4.2

?

?

personalized recommendation

e.g. the user’s history,
location, social network
activity, …

Contextual Bandits Generalizes MAB and SL

Multi-Armed Bandit → 4.2, ?, ?

No input, bandit feedback

- Generalization
- **Exploration**
- Credit assignment

Input → **Supervised Learning** → 4.2, 5.0, 3.1

Takes input, full-information feedback

- **Generalization**
- Exploration
- Credit assignment

Input → **Contextual Bandit** → 4.2, ?, ?

Takes input, bandit feedback

- **Generalization**
- **Exploration**
- Credit assignment

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

noise

$|R(x,a)| \leq 1$

$|w_t| \leq 1$

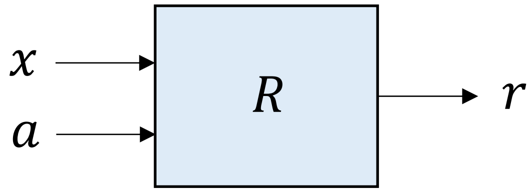
Discussion

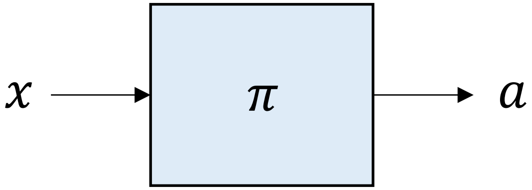
- Contextual bandits is a minimal simultaneous generalization of supervised learning (SL) and multi-armed bandits (MAB)
- We learned a lot about SL in machine learning courses
- We just learned some simple MAB algorithms
  - 3 strategies based on mean estimation

- **Question:** Can you design a contextual bandits algorithm based on the techniques you know for SL and MAB?

Two ways to leverage SL techniques in CB

x: context, a: action, r: reward





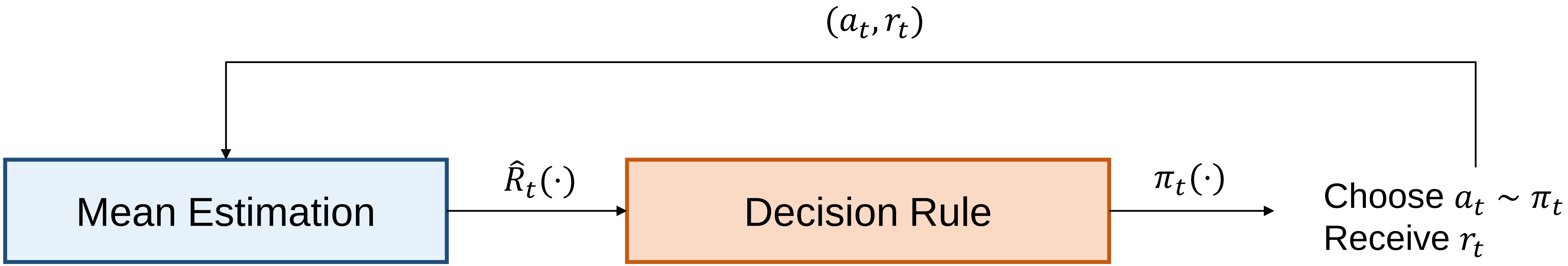
Learn a mapping from
(context, action) to reward

Learn a mapping from
context to action (or action distribution)

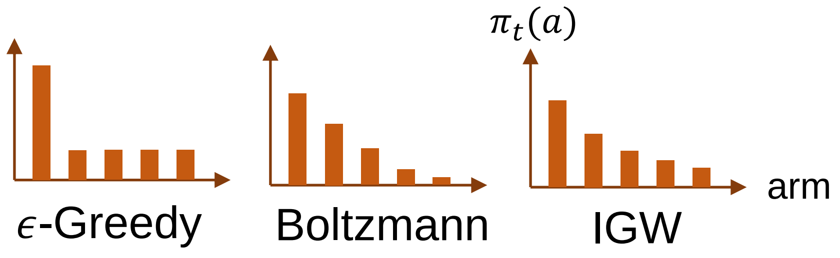
CB with **regression oracle**
Value-based approach
(discussed next)

CB with **classification oracle**
Policy-based approach
(slightly later in the course)

Recall: MAB Based on Mean Estimation



$$\hat{R}_t(a) = \frac{\sum_{s=1}^{t-1} \mathbb{I}\{a_s = a\} r_s}{\sum_{s=1}^{t-1} \mathbb{I}\{a_s = a\}}$$

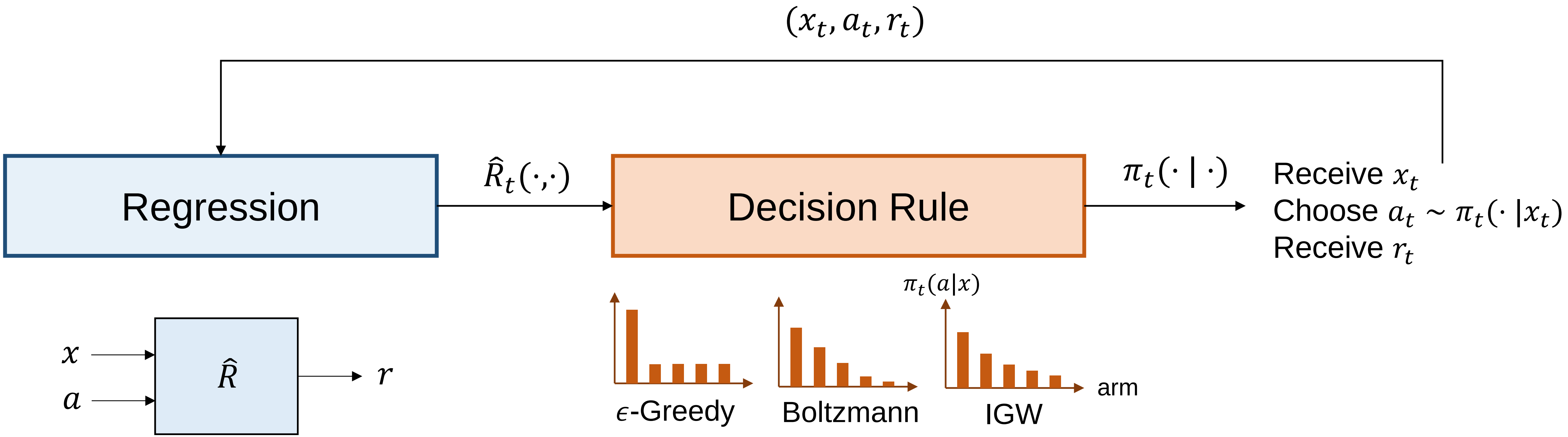


$$\pi_t(a) = (1-\epsilon)\mathbb{I}\{a = \operatorname{argmax} \hat{R}_t(\cdot)\} + \frac{\epsilon}{A}$$

$$\pi_t(a) \propto \exp\big(\lambda \hat{R}_t(a)\big)$$

$$\pi_t(a) = \frac{1}{\gamma_t - \lambda \hat{R}_t(a)}$$

CB Based on Reward Function Estimation (Regression)



Train a $\hat{R}$ such that $r_i \approx \hat{R}(x_i, a_i)$

$$\pi_t(a|x) = (1-\epsilon)\mathbb{I}\{a = \operatorname{argmax} \hat{R}_t(x,\cdot)\} + \frac{\epsilon}{A}$$

$$\pi_t(a|x) \propto \exp\left(\lambda \hat{R}_t(x,a)\right)$$

$$\pi_t(a|x) = \frac{1}{\gamma_t - \lambda \hat{R}_t(x,a)}$$

CB Based on Reward Function Estimation

Instantiate a regression procedure $\hat{R}_1$

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

Receive context x_t

Design a distribution $\pi_t(\cdot|x_t)$ based on the estimated reward $\hat{R}_t(x_t,\cdot)$

EG $\quad \pi_t(a|x_t) = (1-\epsilon)\mathbb{I}\{a = \operatorname{argmax} \hat{R}_t(x_t,\cdot)\} + \frac{\epsilon}{A}$

BE $\quad \pi_t(a|x_t) \propto \exp\left(\lambda \hat{R}_t(x_t, a)\right)$

IGW $\quad \pi_t(a|x_t) = \frac{1}{\gamma_t - \lambda \hat{R}_t(x_t, a)}$

Sample an action $a_t \sim \pi_t(\cdot\,|x_t)$ and receive the corresponding reward r_t.

Refine the reward estimator $\hat{R}_{t+1}(\cdot,\cdot)$ with the new sample (x_t, a_t, r_t).

Regret in Contextual Bandits

(space of context)

deterministic policy: $\mathcal{X} \longrightarrow A$ $\quad \pi(x) \in A$

randomized policy: $\mathcal{X} \longrightarrow \Delta_A$ $\quad \pi(a|x) \in [0,1]$

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

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

Benchmark policy: $\pi^\star(x) = \underset{a \in \mathcal{A}}{\operatorname{argmax}}\ R(x, a)$

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

Regret in Contextual Bandits

$\hat{R}_t$ sees $(x_1, a_1, r_1) \cdots (x_{t-1}, a_{t-1}, r_{t-1})$

Regret Bound of ϵ-Greedy

ϵ-Greedy ensures

$$\text{Regret} \lesssim \epsilon T + \sqrt{\frac{AT \cdot \text{Err}}{\epsilon}}$$

Regression error

$$\text{Err} = \sum_{t=1}^{T} \left(\hat{R}_t(x_t, a_t) - R(x_t, a_t) \right)^2$$

Regret Bound of Inverse Gap Weighting

IGW ensures

$$\text{Regret} \lesssim \frac{AT}{\lambda} + \lambda \cdot \text{Err}$$

Will be proven in HW1

Summary

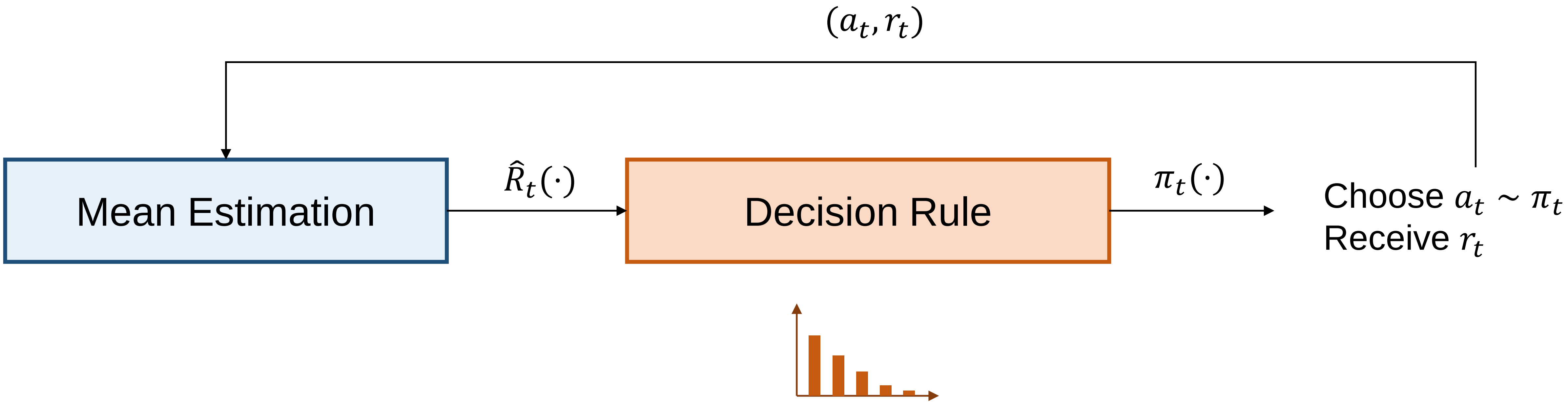
- Contextual bandits (CB) simultaneously generalizes supervised learning (SL) and multi-armed bandits (MAB). It captures the challenges of **generalization** and **exploration** in online RL.

- Any MAB algorithm based on "**mean estimation**" can be lifted as a CB algorithm with "**reward function estimation**" by leveraging a regression oracle.
  - This gives a general framework for value-based CB

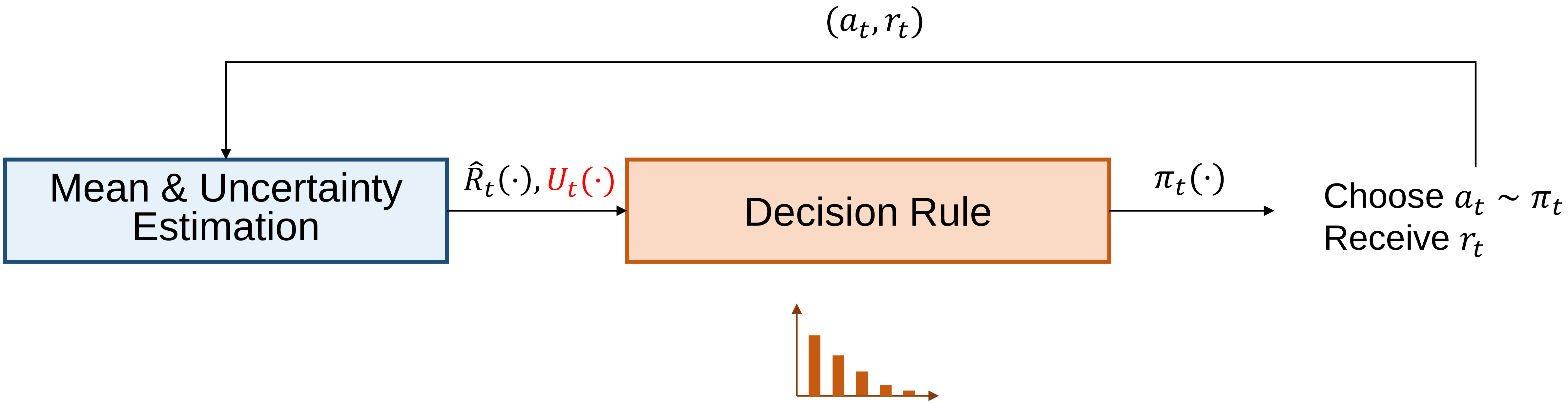
Multi-Armed Bandits

Based on mean and uncertainty estimation

Recall: MAB Based on Mean Estimation



MAB Based on Mean and Uncertainty Estimation



$U_t(a)$: measures the uncertainty of $\hat{R}_t(a)$

$$\left|\hat{R}_t(a) - R(a)\right| \leq \sqrt{\frac{2\log(2/\delta)}{N_t(a)}} \triangleq U_t(a)$$

times taking a before time t

This inequality is used in the **math analysis** of ϵ-Greedy and IGW, but not in their **algorithm**.

Useful Idea: “Optimism in the Face of Uncertainty”

In words:

Act according to the **best plausible world**.

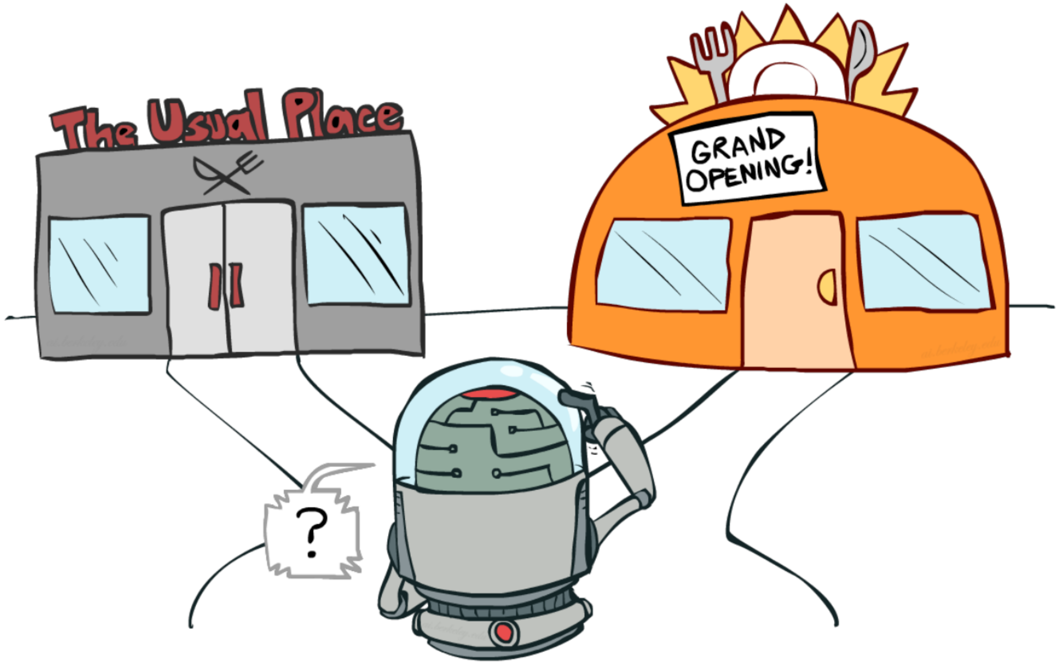


Image source: UC Berkeley CS188

Another Idea: "Optimism in the Face of Uncertainty"

In words:

Act according to the **best plausible world**.

At time t, suppose that arm a has been drawn for $N_t(a)$ times, with empirical mean $\hat{R}_t(a)$.

What can we say about the true mean $R(a)$?

$$\left| R(a) - \hat{R}_t(a) \right| \leq \sqrt{\frac{2 \log(2/\delta)}{N_t(a)}} \quad \text{w.p.} \geq 1 - \delta$$

What's the most optimistic mean estimation for arm a?

$$\hat{R}_t(a) + \sqrt{\frac{2 \log(2/\delta)}{N_t(a)}}$$

Upper Confidence Bound (UCB)

UCB (Parameter: δ) Usually decreases over time as $\delta_t \sim 1/t$ (drives continual exploration)

$1/t^2$

In the first A rounds, draw each arm once.

For the remaining rounds: in round t, draw

Exploration Bonus

$$a_t = \operatorname{argmax}_a \;\; \hat{R}_t(a) + \sqrt{\frac{2\log(2/\delta)}{N_t(a)}}$$

where $\hat{R}_t(a)$ is the empirical mean of arm a using samples up to time $t-1$.

$N_t(a)$ is the number of samples of arm a up to time $t-1$.

P Auer, N Cesa-Bianchi, P Fischer. **Finite-time analysis of the multiarmed bandit problem**, 2002.

Regret Bound of UCB

Theorem. Regret Bound of UCB

UCB ensures with high probability,

$$\text{Regret} \lesssim \sqrt{AT} \ .$$

Visualizing UCB

True mean: [0.2, 0.4, 0.6, 0.7] animation

UCB Regret Bound Analysis

UCB Regret Bound Analysis

Summary: Algorithms We Learned So Far

	Regret Bound	**Approach**
Explore-then-Exploit ϵ-Greedy Boltzmann Exploration Inverse Gap Weighting	$A^{1/3}\,T^{2/3}$ $A^{1/3}\,T^{2/3}$ X $\sqrt{AT}$	Mean estimation + decision rule
Upper Confidence Bound Thompson Sampling Arm Elimination	$\sqrt{AT}$	Mean and uncertain estimation + decision rule

Thompson Sampling – A Bayesian Approach for MAB

Assumptions:

- At the beginning, the environment draws a parameter $\theta^\star$ from some prior distribution $\theta^\star \sim P_{\text{prior}}$
- In every round, the reward vector $\boldsymbol{r_t} = (r_t(1), \dots, r_t(A))$ is generated from $\boldsymbol{r_t} \sim P_{\theta^\star}$

E.g., Gaussian Case

- At the beginning, $\theta^\star(a) \sim \mathcal{N}(0, 1)$ for all $a \in \{1, \dots, A\}$.
- In every round, the reward of arm a is generated by $r_t(a) \sim \mathcal{N}(\theta^\star(a), 1)$.

For the learner, P_{prior} is known; $\theta^\star$ is unknown; P_θ is known for any θ.

Thompson Sampling

William Thompson. **On the likelihood that one unknown probability exceeds another in view of the evidence of two samples**, 1933.

In words:

Randomly pick an arm according to the probability you **believe** it is the optimal arm.

At time t, after seeing $\mathcal{H}_t = (a_1, r_1(a_1), a_2, r_2(a_2), \ldots, a_{t-1}, r_{t-1}(a_{t-1}))$, the learner has a **posterior distribution** for $\theta^\star$:

$$P(\theta^\star = \theta | \mathcal{H}_t) = \frac{P(\mathcal{H}_t, \theta^\star = \theta)}{P(\mathcal{H}_t)} = \frac{P_\theta(\mathcal{H}_t) P_{\text{prior}}(\theta)}{P(\mathcal{H}_t)} \propto P_\theta(\mathcal{H}_t) P_{\text{prior}}(\theta)$$

In math:

Sample a_t according to $\pi_t(a) = \int_\theta P(\theta | \mathcal{H}_t)\, \mathbb{I}\{a^\star(\theta) = a\} = \mathbb{E}_{\theta \sim P(\cdot | \mathcal{H}_t)}[\mathbb{I}\{a^\star(\theta) = a\}]$

Implementation: Sample $\theta_t \sim P(\cdot \mid \mathcal{H}_t)$, and choose $a_t = a^\star(\theta_t)$.

Also called "**Posterior Sampling**"

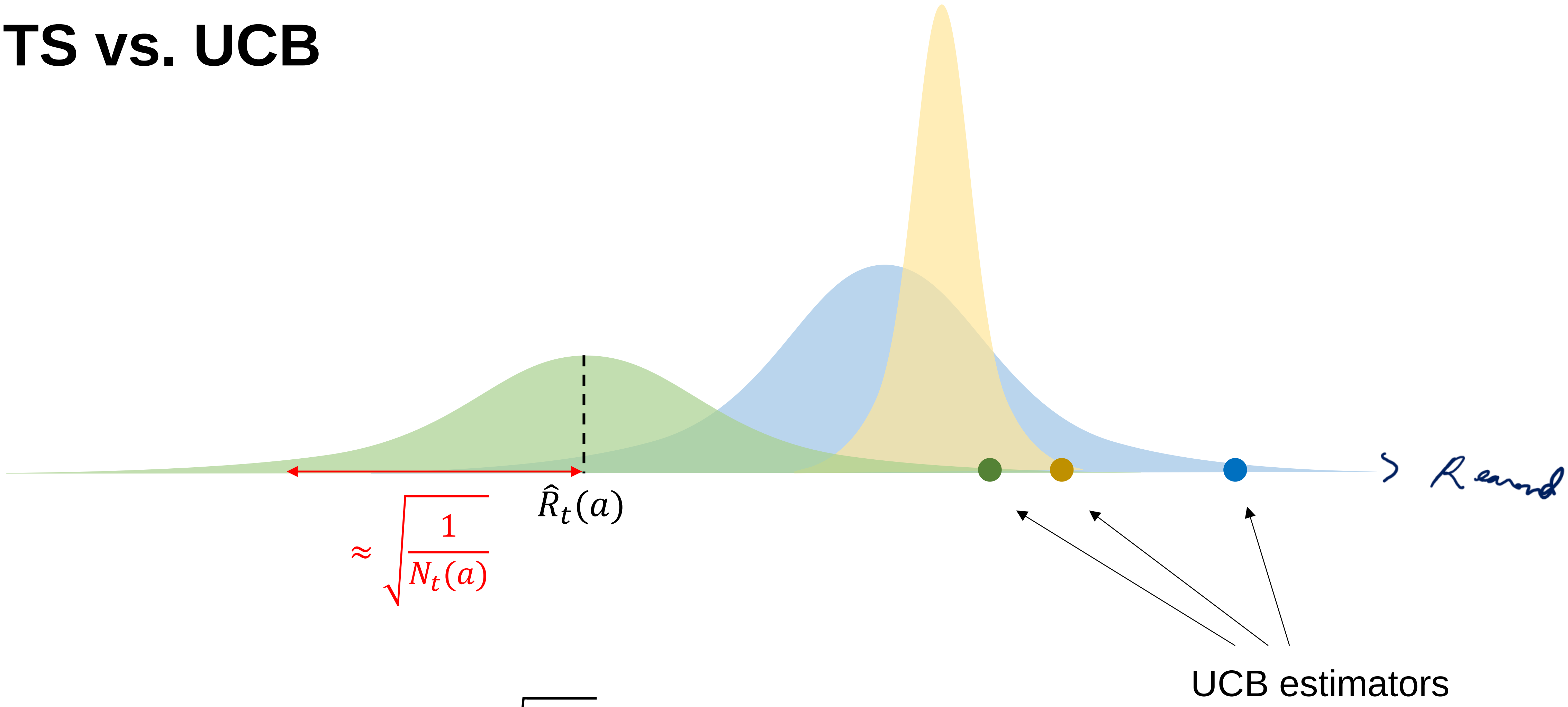
Gaussian Thompson Sampling

Gaussian prior $\theta^\star(a) \sim \mathcal{N}(0, 1)$ + **Gaussian reward** $r_t(a) \sim \mathcal{N}(\theta^\star(a), 1)$:

$$P(\theta^\star(a) = \theta(a) \mid \mathcal{H}_t) = \mathcal{N}\left(\hat{R}_t(a), \frac{1}{N_t(a)+1}\right) \text{ where } \hat{R}_t(a) = \frac{\sum_{s=1}^{t-1} \mathbb{I}\{a_t = a\} r_t(a)}{N_t(a)+1}$$

Empirical mean assuming 1 fake sample with reward 0

TS vs. UCB



UCB: $a_t \approx \text{argmax}_a \; \hat{R}_t(a) + c\sqrt{\frac{1}{N_t(a)}}$

Gaussian TS: $a_t \approx \text{argmax}_a \; \hat{R}_t(a) + c\sqrt{\frac{1}{N_t(a)}} n_t(a)$ with $n_t(a) \sim \mathcal{N}(0,1)$

More on Thompson Sampling

For **Bernoulli** reward, we assume the **Beta** prior:
https://gdmarmerola.github.io//ts-for-bernoulli-bandit/

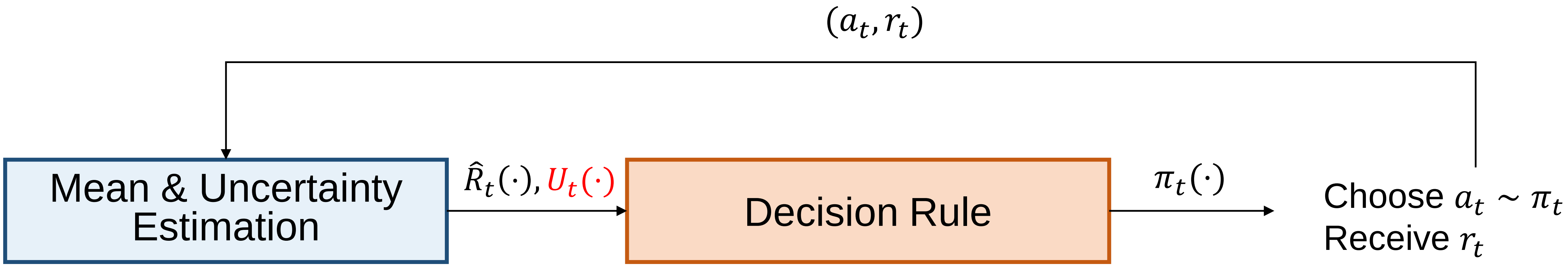
Thompson sampling is empirically strong

Chapelle and Li. An Empirical Evaluation of Thompson Sampling. 2011.

Yang. A Study on Multi-Arm Bandit Problem with UCB and Thompson Sampling Algorithm. 2024.

Wang and Chen. Thompson Sampling for Combinatorial Semi-Bandits. 2018.

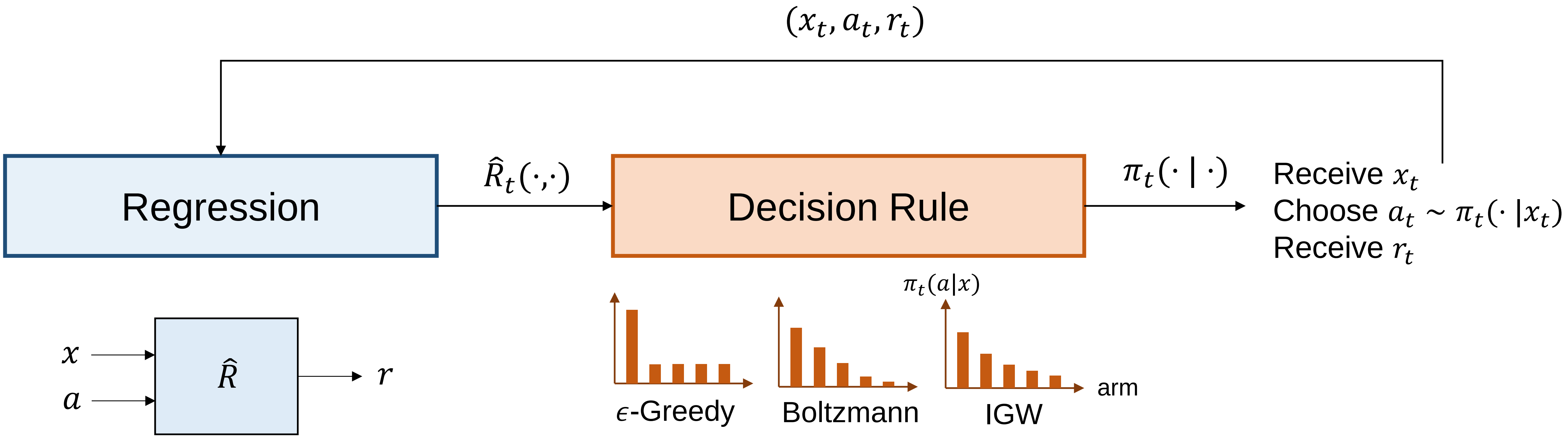
Extension to Contextual Bandits?



- Linear UCB, Linear TS   (using linear regression)
- In UCB or TS or Linear UCB or Linear TS, we really did not perform "**uncertainty estimation**" – the uncertainty measure is directly derived from Hoeffding's bound or prior knowledge about reward distribution.
- When general function approximation is used, it's no longer easy to "derive" uncertainty measure, so it really needs to be "estimated".
- Let's talk about this more in MDP.

Summary

The Most Important Slide for Value-Based Bandits



Train a $\hat{R}$ such that $r_i \approx \hat{R}(x_i, a_i)$

Exploration by Modifying the Reward

- Add exploration bonus (UCB) or perturbation (TS) that scales with the degree of uncertainty.