Elements of Interactive Decision Making

Chen-Yu Wei (University of Virginia)

2026 IEEE East Asian School of Information Theory

Outline

Part I. Reinforcement Learning and the Exploration-Exploitation tradeoff

- Reinforcement learning vs. supervised learning
- Multi-armed bandits
- Linear bandits

Part II. Information-theoretic algorithm design for the Exploration-Exploitation tradeoff

- Decision-making with structured observation (DMSO)
- Bayesian setting
- Frequentist setting

Part III. Beyond model estimation

- Value estimation / policy estimation
- Open problems

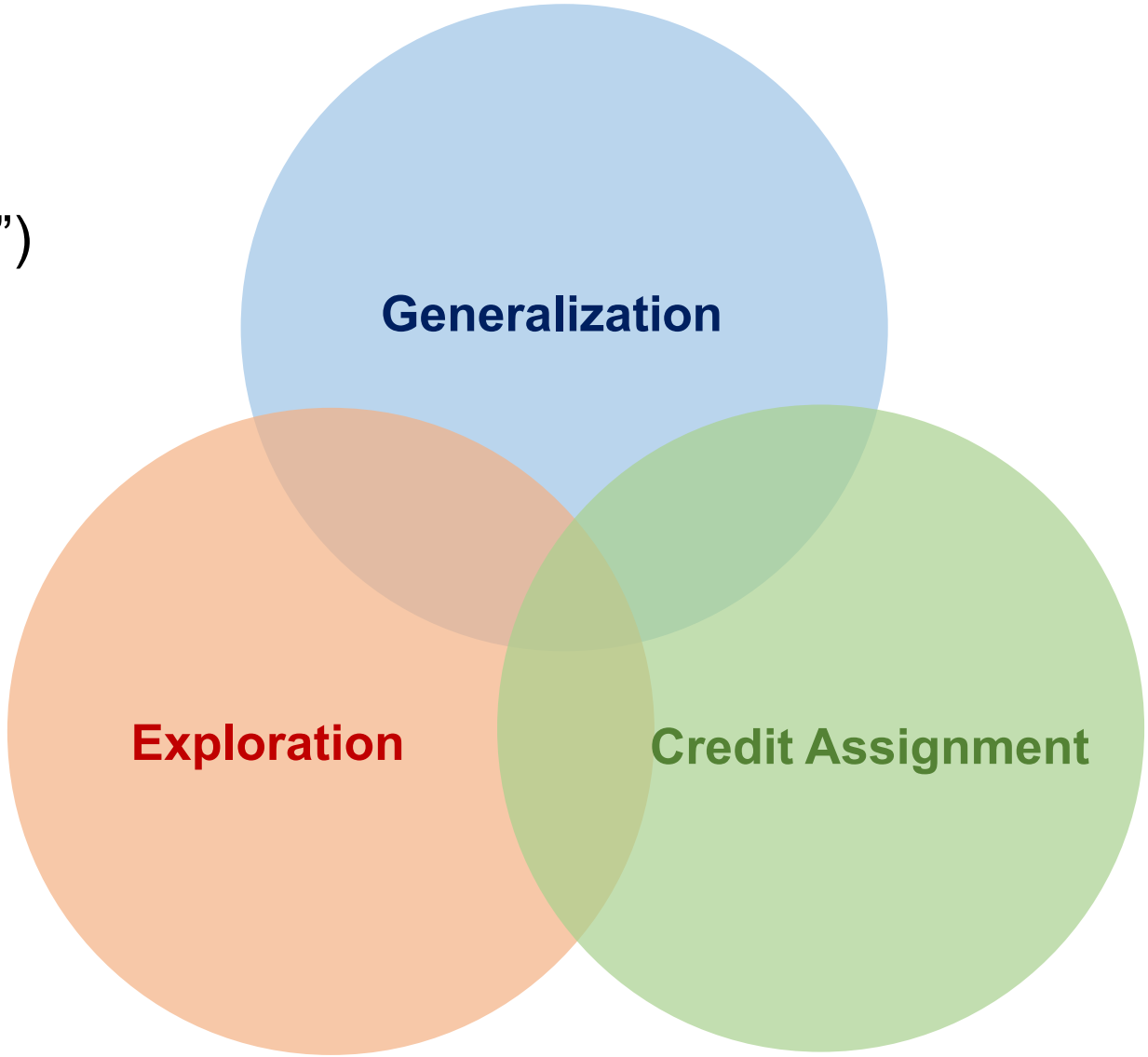
Goal

- Get an idea of the **challenges of exploration** in bandits/RL
- Get an idea of how information theory can be used in bandits/RL algorithm design and analysis
- Get some theoretical open problems to work on

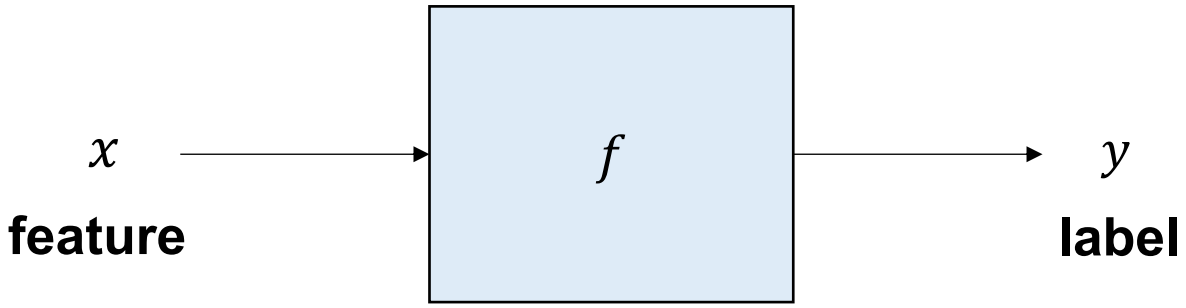
Part I. RL and the Exploration-Exploitation Tradeoff

Core Challenges of RL

Today, we focus on “exploration”
(or the “exploration-exploitation tradeoff”)



Supervised Learning

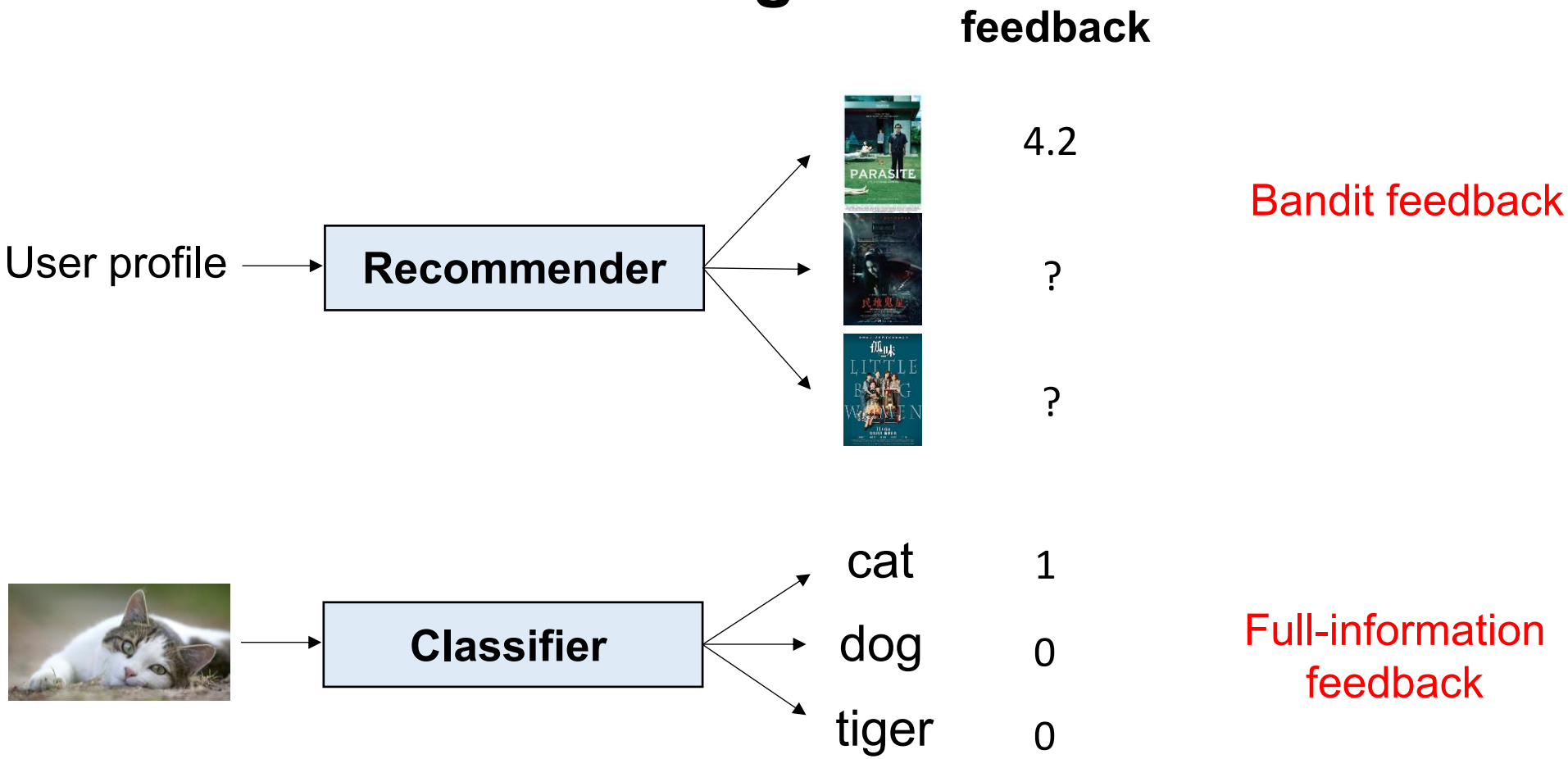


f () = Cat

f (temperature, humidity,...) = 1000mm precipitation

Given a lot of (x, y) pairs, find an f that such that $f(x) \approx y$

Reinforcement Learning

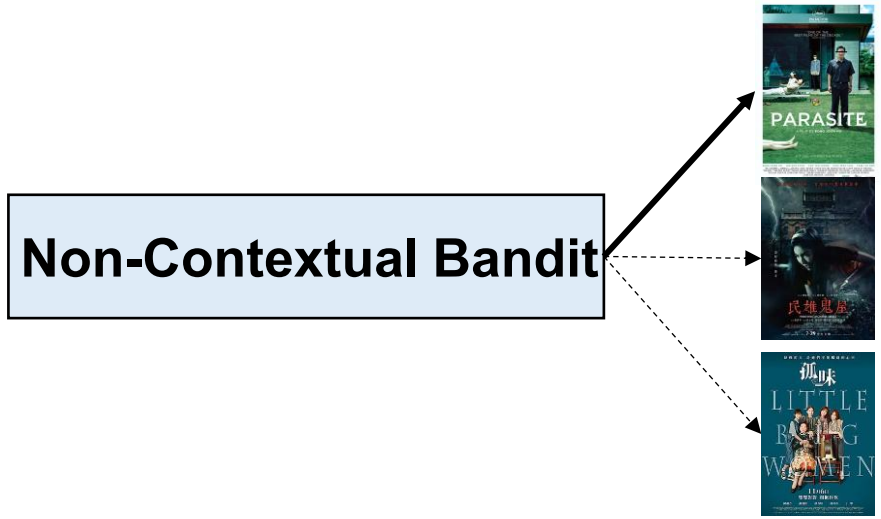


RL usually deals with bandit feedback

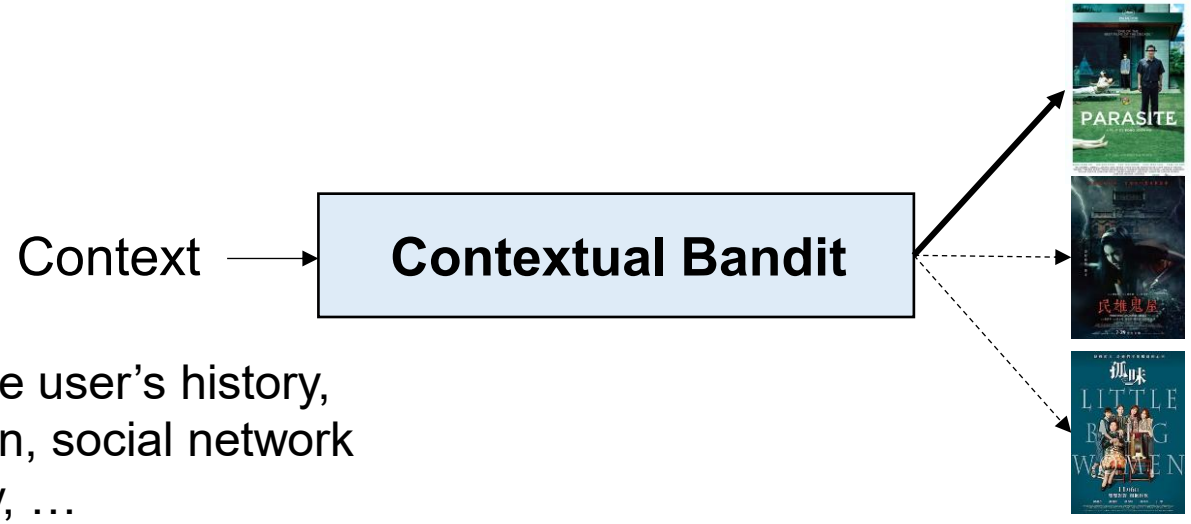
Bandit Feedback

- Needs **exploration**

Non-Contextual Bandits vs. Contextual Bandits



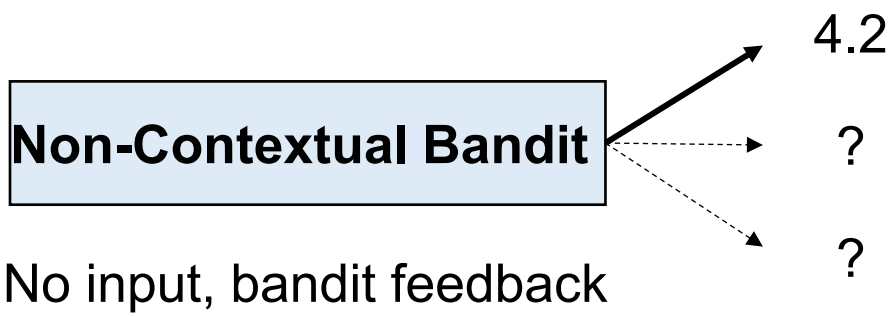
all-user recommendation



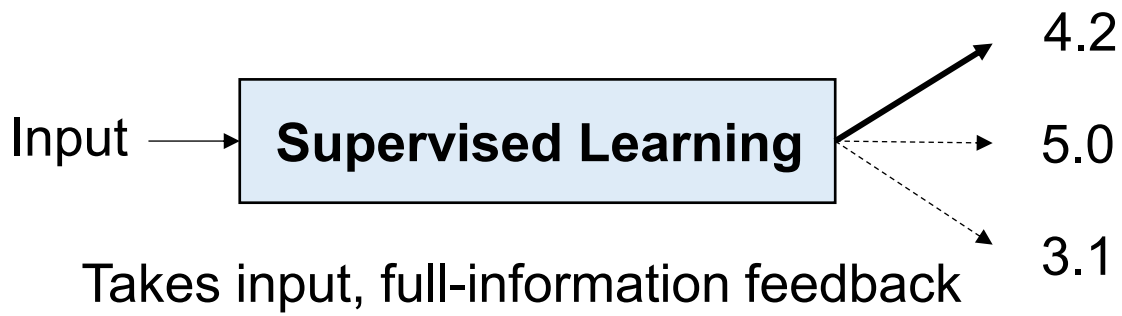
personalized recommendation

E.g. the user's history, location, social network activity, …

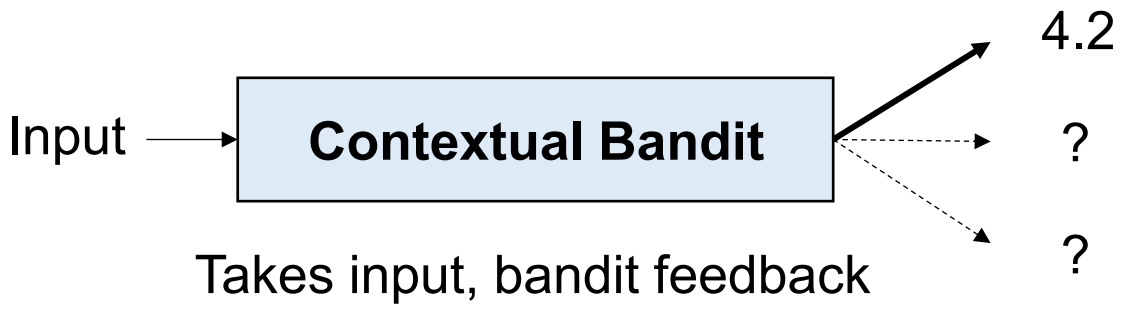
A Comparison



No input, bandit feedback



Takes input, full-information feedback



Takes input, bandit feedback

Multi-Armed Bandits

A non-contextual bandit problem

Multi-Armed Bandits (MAB)

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

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

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

Learner observes $r_t = R(a_t) + \text{zero mean noise}$

Arm = Action

$R(a)$ is the ground-truth expected reward function

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

Hope: achieve Regret $= o(T)$

The Exploration and Exploitation Trade-off in MAB

- To perform as well as the best arm, the learner has to pull it 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**

Multi-Armed Bandits Algorithms

Based on mean estimation

A Simple Strategy: ϵ-Greedy

ϵ-Greedy (Parameter: ϵ)

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 up to time $t-1$.

The parameter ϵ balances exploitation and exploration

Regret Bound Achieved by ϵ-Greedy

Theorem. Regret Bound of ϵ-Greedy

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

Then ϵ-Greedy ensures with high probability,

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

Minimax regret bound: $\sqrt{AT}$

Can We Do Better?

In 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?

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

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

Exploration adaptive to Empirical Mean

Boltzmann Exploration

In each round, sample a_t according to

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

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

λ controls the degree of exploration (often adaptive over time)

However, theory has shown that Boltzmann exploration cannot attain $\sqrt{AT}$ regret, unless with several modifications.

Nicolò Cesa-Bianchi, Claudio Gentile, Gábor Lugosi, Gergely Neu. Boltzmann exploration done right. 2017.

Summary: MAB Based on Mean Estimation

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

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

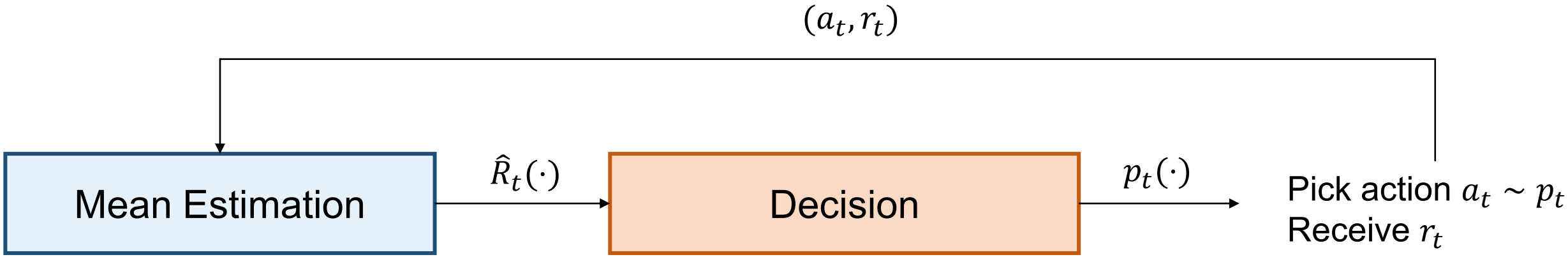
EG $p_t(a) = (1 - \epsilon_t)\mathbb{I}\left\{a = \underset{a'}{\operatorname{argmax}}\, \hat{R}_t(a')\right\} + \frac{\epsilon_t}{A}$

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

Sample an arm $a_t \sim p_t$ 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



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

EG: $\pi_t(a) = (1-\epsilon)\mathbb{I}\left\{a = \underset{a'}{\text{argmax}}\, \hat{R}_t(a')\right\} + \frac{\epsilon}{A}$

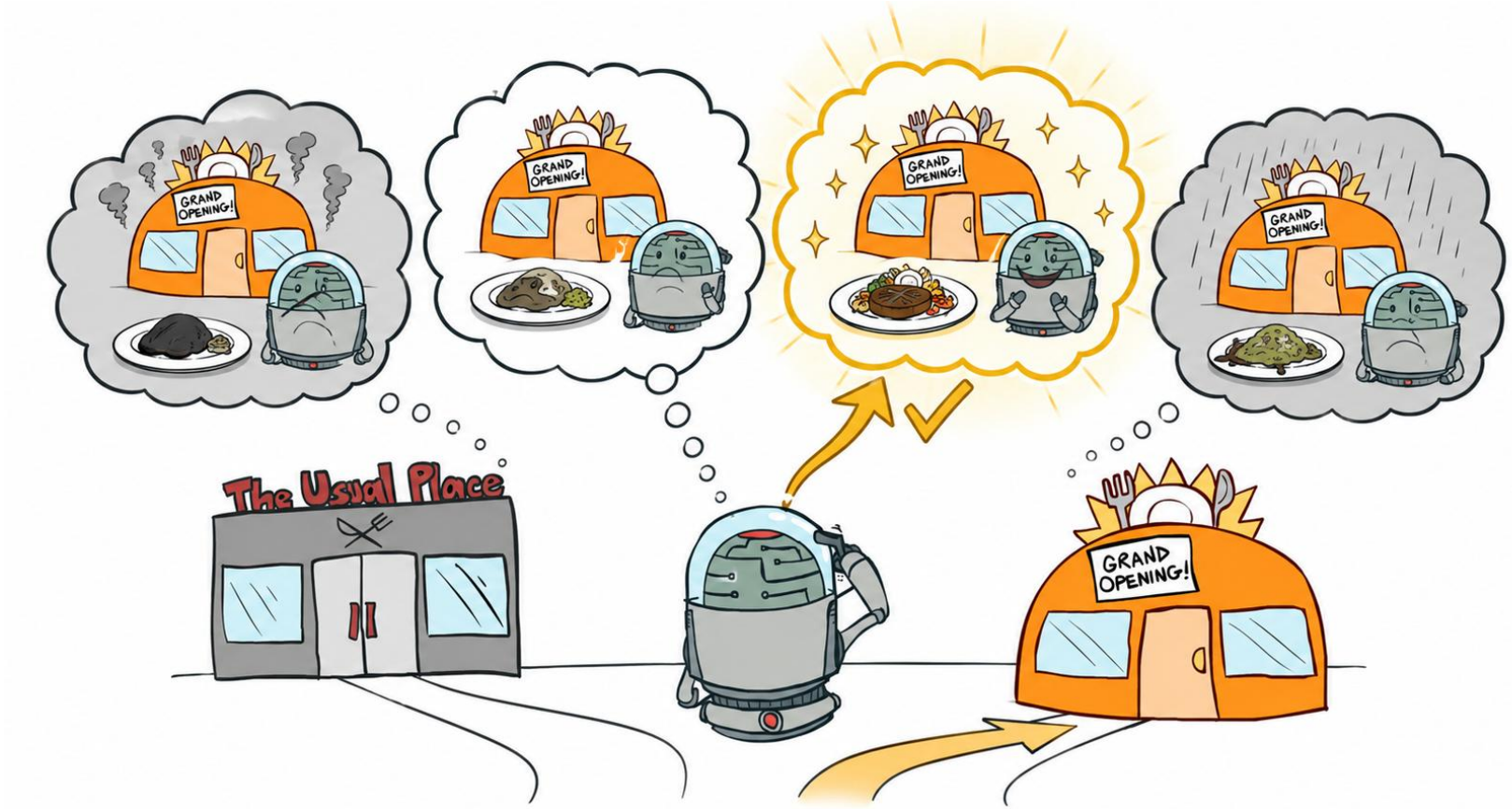
BE: $\pi_t(a) \propto \exp\left(\lambda \hat{R}_t(a)\right)$

Multi-Armed Bandits Algorithms

Based on mean estimation and uncertainty quantification

Useful Idea: “Optimism in the Face of Uncertainty”

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



Another Idea: "Optimism in the Face of Uncertainty"

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

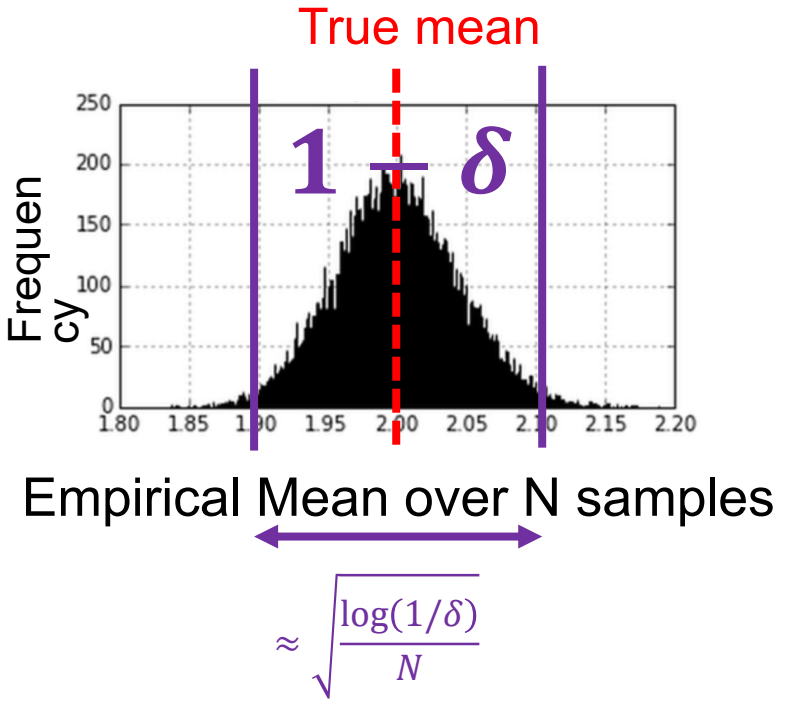
At time t, suppose that arm a has been drawn for $N_t(a)$ times, with empirical mean $\hat{R}_t(a)$.

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

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

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

$$\hat{R}_t(a) + c\sqrt{\frac{\log(1/\delta)}{N_t(a)}}$$



Upper Confidence Bound (UCB)

UCB (Parameter: c, δ)

In round t, draw

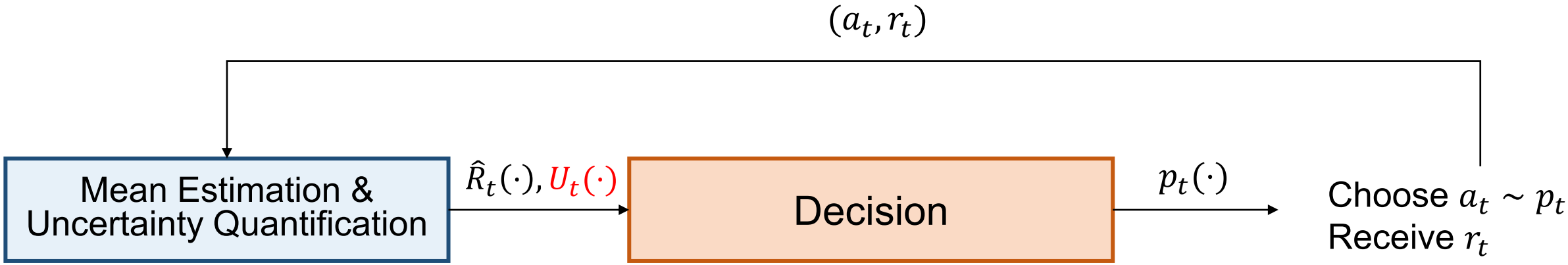
Exploration Bonus
= Amount of Uncertainty

$$a_t = \text{argmax}_a \quad \hat{R}_t(a) + c\sqrt{\frac{\log(1/\delta)}{N_t(a)}}$$

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 $t-1$.

MAB Based on Mean Estimation and Uncertainty Quantification



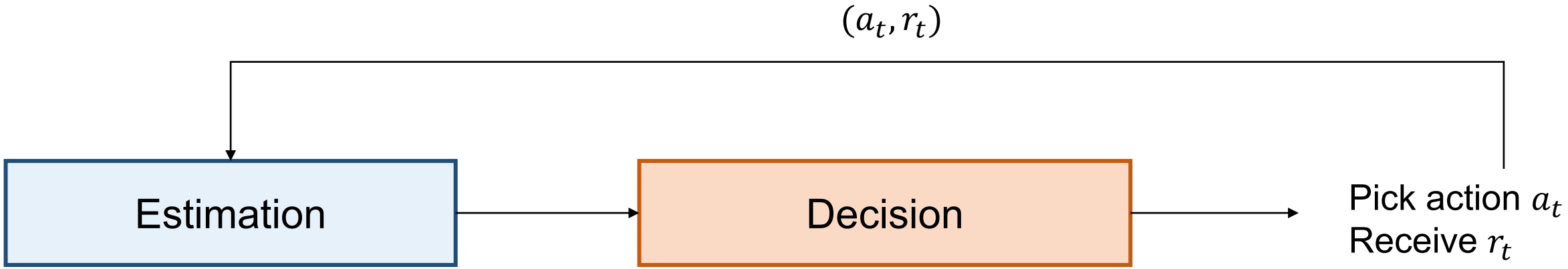
$U_t(a)$: quantifies our uncertainty of $\hat{R}_t(a)$

Our choice of $U_t(a)$: $\left|\hat{R}_t(a) - R(a)\right| \leq c\sqrt{\frac{\log(1/\delta)}{N_t(a)}} \triangleq U_t(a)$

Visualizing UCB

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

MAB Algorithms We Discussed So Far



	Estimation	Decision	Regret Bound
ϵ-Greedy	Mean $\hat{R}_t$	$\text{argmax}\ \hat{R}_t$ + uniform	$A^{1/3}T^{2/3}$
Boltzmann	Mean $\hat{R}_t$	$\text{softmax}\ \hat{R}_t$	--
UCB	Mean $\hat{R}_t$ + Uncertainty U_t	$\text{argmax}\left(\hat{R}_t + U_t\right)$	$\sqrt{AT}$

Linear Bandits

Linear Bandits

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

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

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

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

Same protocol as MAB, but with additional assumptions:

- Each arm a is associated with a known **feature vector** $\phi(a) \in \mathbb{R}^d$.
- There exists a **weight vector** $\theta^\star \in \mathbb{R}^d$ such that $R(a) = \phi(a)^\top \theta^\star$ for all arm a.

Linear Bandits

- Underlying assumption: actions sharing features will have similar reward.
- The learner only needs to **learn the weights** for each dimension of the feature, but not the reward of each individual arm.

	Multi-armed Bandits	**Linear Bandits**
Estimation	Empirical mean $\hat{R}_t(a)$	Linear regression $\hat{R}_t(a) = \phi(a)^\top \hat{\theta}_t$
Decision (ϵ-Greedy)	Uniform over arms	Optimal design over features
Decision (UCB)	$U_t(a) \approx 1/\sqrt{N_t(a)}$	$U_t(a) \approx \|\phi(a)\|_{\Lambda_t^{-1}}$ where $\Lambda_t = \sum_{s<t} \phi(a_s)\phi(a_s)^\top$

Linear ϵ-Greedy: $\mathrm{Reg}_T = \tilde{O}(d^{2/3}T^{2/3})$

Linear UCB: $\mathrm{Reg}_T = \tilde{O}(d\sqrt{T})$

Part II. Information-Theoretic Lens of Decision-Making Problems

Outline

Part I. Reinforcement Learning and the Exploration-Exploitation tradeoff

- Reinforcement learning vs. supervised learning
- Multi-armed bandits
- Linear bandits

Part II. Information-theoretic algorithm design for the Exploration-Exploitation tradeoff

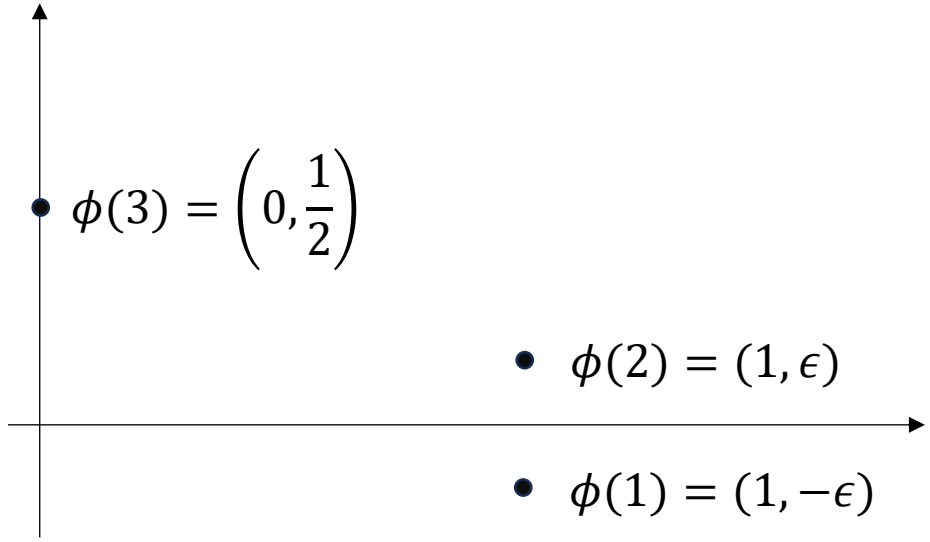
- Decision-making with structured observation (DMSO)
- Bayesian setting
- Frequentist setting

Part III. Beyond model estimation

- Value estimation / policy estimation
- Open problems

Is Optimism All We Need (1/3)?

Consider the following (somewhat contrived) **3-arm linear bandit** example:



Features of the 3 arms $(0 < \epsilon < 0.02)$

Assume the observed reward is $\phi(a)^\top \theta^\star + \mathcal{N}(0,1)$

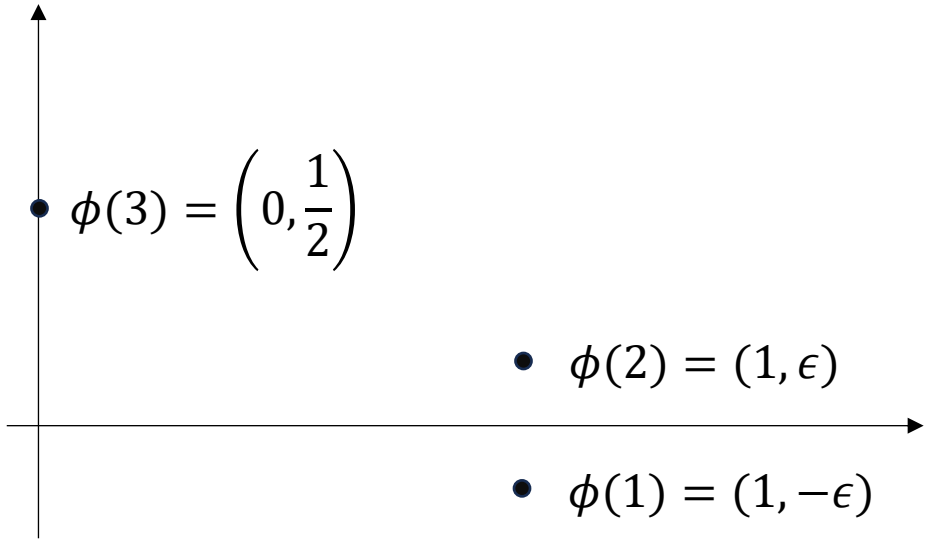
we have the prior knowledge that the true weight vector is either $\theta^\star = (1, -1)$ or $\theta^\star = (1,1)$

$\theta^\star$	Optimal arm
$(1, -1)$	1
$(1, 1)$	2

Lattimore and Szepesvari. The End of Optimism? An Asymptotic Analysis of Finite-Armed Linear Bandits. 2016.

Is Optimism All We Need (1/3)?

$\theta^\star = (1, \theta_2^\star)$ where $\theta_2^\star = -1$ or 1



Taking arm 1 observes $\mathcal{N}(1 - \epsilon\theta_2^\star,\ 1)$

Taking arm 2 observes $\mathcal{N}(1 + \epsilon\theta_2^\star,\ 1)$

Taking arm 3 observes $\mathcal{N}\left(\frac{1}{2}\theta_2^\star,\ 1\right)$

Optimistic algorithm would only choose arm 1 or 2

$\Rightarrow$ Need to tell apart $\mathcal{N}(1-\epsilon, 1)$ and $\mathcal{N}(1+\epsilon, 1)$

$\Rightarrow$ Need $\gtrsim \frac{1}{\epsilon^2}$ samples

$\Rightarrow$ Regret $\gtrsim \frac{1}{\epsilon^2} \times \boldsymbol{\epsilon} \geq \frac{1}{\epsilon}$

one-step regret before telling apart

If we choose arm 3

$\Rightarrow$ Only need to tell apart $\mathcal{N}\left(-\frac{1}{2}, 1\right)$ and $\mathcal{N}\left(\frac{1}{2}, 1\right)$

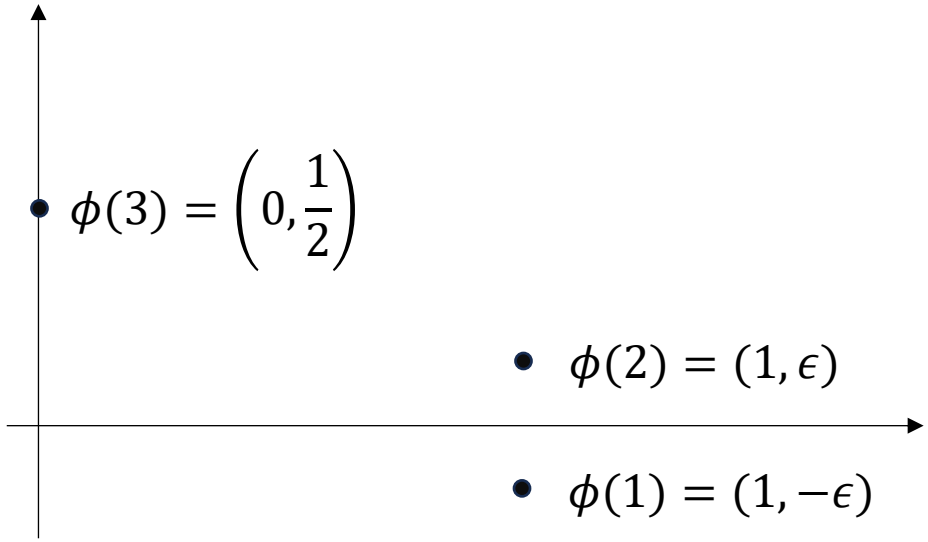
$\Rightarrow$ Only need $\tilde{O}(1)$ samples

$\Rightarrow$ Regret $\lesssim \tilde{O}(1)$

Is Optimism All We Need (1/3)?

$\theta^\star = (1, \theta_2^\star)$ where $\theta_2^\star = -1$ or 1

Main message: Some clearly sub-optimal action could help identify the optimal action quicker.



Lattimore and Szepesvari. The End of Optimism? An Asymptotic Analysis of Finite-Armed Linear Bandits. 2016.

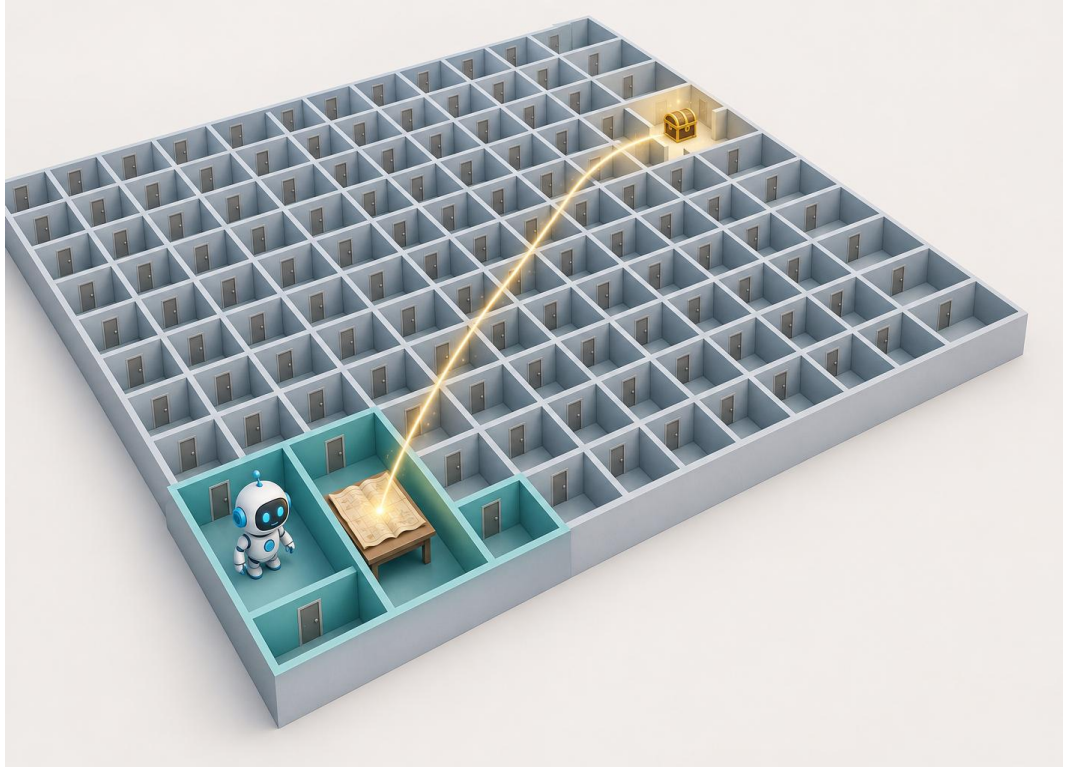
Is Optimism All We Need (2/3)?

Robot navigation

One of the 97 rooms has treasure, but the robot doesn't know where it is.

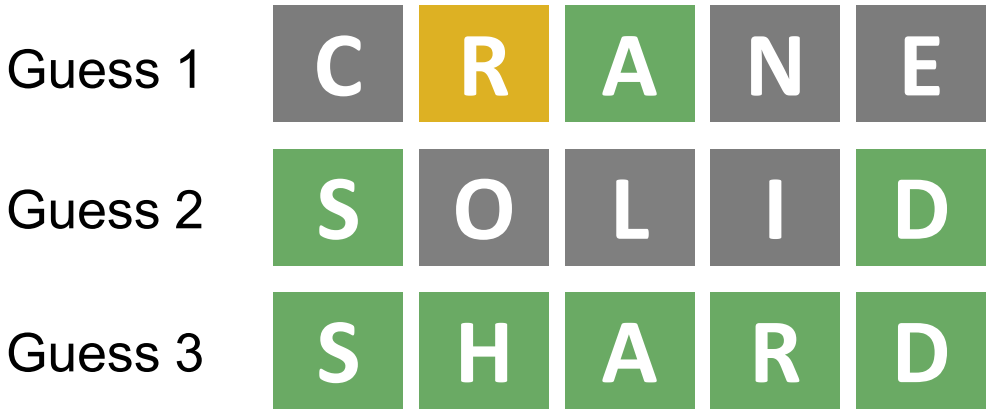
On the other hand, it is known that one of Room 98, 99, 100 contains a map, which contains the information of the treasure location.

Notice: this problem is more general than MAB/LB, because the robot could receive **observations other than rewards** (i.e., the content of the map)



Is Optimism All We Need (3/3)?

Wordle



Green indicates a correct letter in the correct spot.

Yellow indicates that the letter is in the word but in the wrong spot.

Gray means the word doesn't contain that letter in any spot.

Is Optimism All We Need (3/3)?

Wordle

Candidates: **LIGHT**, **MIGHT**, **NIGHT**, **RIGHT**, **SIGHT**

Optimism-based decision guess some plausible answer, like **LIGHT**

But guess like **SNARL** could more quickly identify which letter is in the answer

Decision Making with Structured Observation (DMSO)

Decision Making with Structured Observation (DMSO)

Given to learner: decision set Π,
observation set $\mathcal{O}$,
model set $\mathcal{M}$,
payoff function f: $\mathcal{M} \times \Pi \to [0,1]$

Every model $M \in \mathcal{M}$ is a mapping $\Pi \to \Delta(\mathcal{O})$.
$M(\pi) \in \Delta(\mathcal{O})$ is the distribution over observations if we play decision π in model M.

Environment chooses $M^\star \in \mathcal{M}$ (NOT revealed to Learner)

For $t = 1, 2, \ldots, T$:
Learner takes a decision $\pi_t \in \Pi$
Learner observes an observation $o_t \sim M^\star(\pi_t)$

$$\pi_M := \operatorname*{argmax}_{\pi \in \Pi} f_M(\pi)$$

$$\text{Regret} = \mathbb{E}\left[\sum_{t=1}^{T} \left(f_{M^\star}(\pi_{M^\star}) - f_{M^\star}(\pi_t)\right)\right]$$

For simplicity, we assume $|\mathcal{M}|, |\Pi| < \infty$ throughout the presentation

Foster, Kakade, Qian, Rakhlin. The Statistical Complexity of Interactive Decision Making. 2021.

A model M **describes a world**, specifying the observation distribution under any decision π.

E.g., A 3-arm bandit may look like the following in the DMSO framework:

$\Pi = \{\text{arm } 1, \text{arm } 2, \text{arm } 3\}$ $\quad\quad$ $\mathcal{O} = \mathbb{R}$

$$\mathcal{M} = \left\{ \begin{array}{l} M_1(\text{arm } 1) = \text{Bernoulli}(0.5) \\ M_1(\text{arm } 2) = \text{Normal}(0.4, 1) \\ M_1(\text{arm } 3) = \text{Deterministic}(0.3) \end{array} \right. \quad \left. \begin{array}{l} M_2(\text{arm } 1) = \text{Normal}(0.2,1) \\ M_2(\text{arm } 2) = \text{Bernoulli}(0.6) \\ M_2(\text{arm } 3) = \text{Unif}([0.3, 0.8]) \end{array} \right\} \ \ldots$$

$$f_{M_1}(\text{arm } 1) = 0.5$$
$$f_{M_1}(\text{arm } 2) = 0.4$$
$$f_{M_1}(\text{arm } 3) = 0.3$$

$$f_{M_2}(\text{arm } 1) = 0.2$$
$$f_{M_2}(\text{arm } 2) = 0.6$$
$$f_{M_2}(\text{arm } 3) = 0.55$$

DMSO Examples

Problem	Decision π	Observation o	Payoff function $f_M(\pi)$
Multi-armed bandit	Arm a	Noisy reward r	$R_M(a)$
Linear bandit	Arm a	Noisy reward r	$\phi(a)^\top \theta_M$
Wordle	A word	Colors of each position	$\mathbf{1}_M$[guess is correct]
Robot Navigation	Room to enter	Things in the room, map info (if there is map)	100 × $\mathbf{1}_M$[∃treasure in the room]
Contextual bandit	Policy $x \mapsto a$	Noisy reward r	$\mathbb{E}_x[R_M(x, \pi(x))]$
Episodic MDP	Policy $s \mapsto a$	Trajectory$(s_1, a_1, r_1, s_2, a_2, \dots)$	$\mathbb{E}_{s_1}[V_M^\pi(s_1)]$

The Bayesian Setting

Decision Making with Structured Observation (DMSO)

Given to learner: decision set Π,
observation set $\mathcal{O}$,
model set $\mathcal{M}$,
payoff function $f\colon \mathcal{M} \times \Pi \to \mathbb{R}$
model prior $\mu_1 \in \Delta(\mathcal{M})$

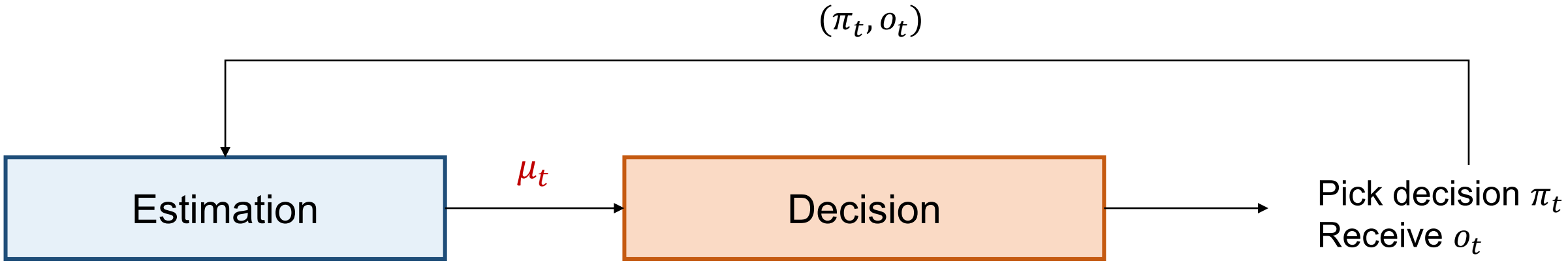
Environment samples $M^\star \sim \mu_1$

For $t = 1, 2, \ldots, T$:
Learner takes a decision $\pi_t \in \Pi$
Learner observes an observation $o_t \sim M^\star(\pi_t)$

$$\text{BayesRegret} = \mathbb{E}\left[\sum_{t=1}^{T} \left(f_{M^\star}(\pi_{M^\star}) - f_{M^\star}(\pi_t)\right)\right]$$

Including the expectation over $M^\star \sim \mu_1$

Algorithm Design: Estimation Part



Model posterior update:

$\mu_{t+1}(M) \propto \mu_t(M) \cdot M(o_t|\pi_t)$

Balancing exploitation and exploration

How?

Notation:

$M(\pi) \in \Delta(\mathcal{O})$

$M(o|\pi) \in [0,1]$ is an entry of $M(\pi)$

Goal of Exploration: Reduce Uncertainty

Model uncertainty makes decision-making difficult.

M_1	M_2
Arm 1 = Bernoulli (0.5)	Arm 1 = Bernoulli (0.5)
Arm 2 = Bernoulli (0.4)	Arm 2 = Bernoulli (0.6)

If μ_t = Uniform $\{M_1, M_2\}$

- The learner will face a difficult situation: Having ½ probability to make a *wrong* decision and suffer 0.1 regret

If μ_t is biased towards either model, the learner will have a clearer picture which arm is better.

Goal of Exploration: Reduce Uncertainty

Model uncertainty can make decision-making difficult:

M_1	M_2
Arm 1 = Bernoulli (0.5)	Arm 1 = Bernoulli (0.5)
Arm 2 = ~~Bernoulli (0.4)~~ Bernoulli (0.3)	Arm 2 = Bernoulli (0.6)

-1000

If μ_t = Uniform $\{M_1, M_2\}$, which arm would you pick?

What if μ_t = Uniform $\{M_1, M_2\}$ but Bernoulli (0.4) is changed to Bernoulli (0.3)?

Arm 2 helps us distinguish models, but gives less reward in expectation…

What if it is further changed to a deterministic -1000?

Quantifying Information Gain

$h_{t-1} \coloneqq (\pi_1, o_1, \pi_2, \ldots, \pi_{t-1}, o_{t-1})$

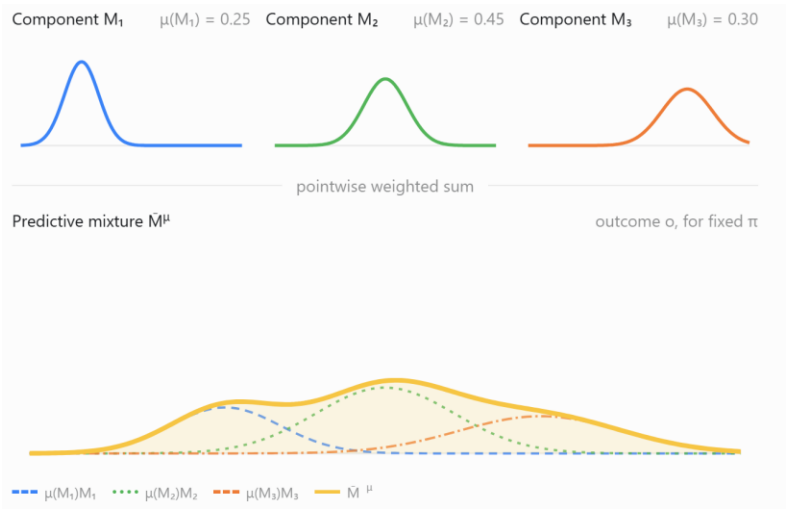
For any decision π, define the **information gain**

$$\begin{aligned}\mathrm{IG}_t(\pi) &= H(M^\star \mid h_{t-1}) - \mathbb{E}_{M\sim\mu_t}\,\mathbb{E}_{o\sim M(\pi)}\left[H(M^\star \mid h_{t-1}, \pi, o)\right] \\ &= I(M^\star; O_t \mid h_{t-1}, \pi) \\ &= \mathbb{E}_{M\sim\mu_t}\left[\mathrm{KL}\left(M(\pi) \;\|\; \overline{M}^{\mu_t}(\pi)\right)\right]\end{aligned}$$

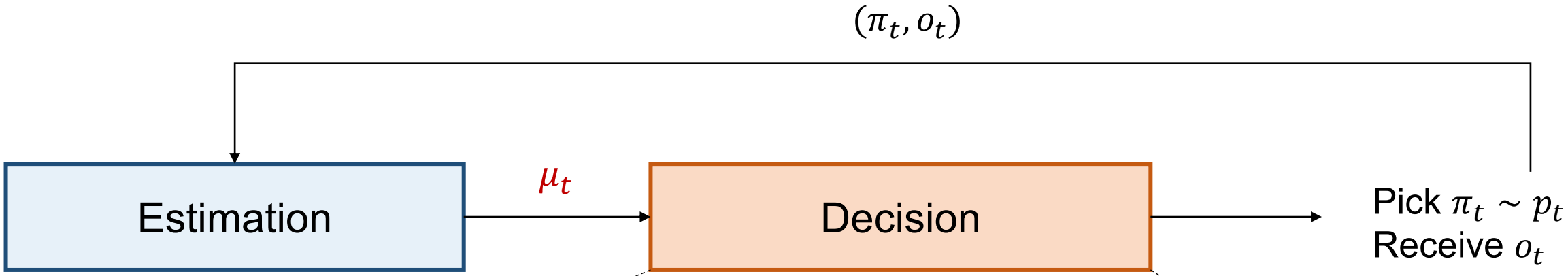
$\overline{M}^{\mu}(\pi)$ is the mixture model with weights μ:

$$\overline{M}^{\mu}(o|\pi) = \sum_{M\in\mathcal{M}} \mu(M)\, M(o|\pi)$$

Remember this notation --- we will use it a lot later.



Algorithm Design: Decision Part



Model posterior update:

$$\mu_{t+1}(M) \propto \mu_t(M) \cdot M(o_t|\pi_t)$$

$$p_t = \operatorname*{argmin}_{p \in \Delta(\Pi)} \mathbb{E}_{\pi \sim p} \mathbb{E}_{M \sim \mu_t} \left[\underbrace{f_M(\pi_M) - f_M(\pi)}_{\text{one-step regret}} - \lambda \cdot \underbrace{\mathrm{KL}\left(M(\pi) \;\|\; \overline{M}^{\mu_t}(\pi) \right)}_{\text{Information gain}} \right]$$

E2D.Bayes

Bayesian Regret Analysis

$$\mathrm{AIR}_\lambda(\mu_t) \coloneqq \min_{p\in\Delta(\Pi)} \mathbb{E}_{\pi\sim p}\mathbb{E}_{M\sim\mu_t}\left[f_M(\pi_M) - f_M(\pi) - \lambda\cdot \mathrm{KL}\left(M(\pi) \;\|\; \overline{M}^{\mu_t}(\pi)\right)\right]$$

$$\mathrm{DEC}_\lambda \coloneqq \max_{\mu\in\Delta(\mathcal{M})} \mathrm{AIR}_\lambda(\mu)$$

$$= \max_{\mu\in\Delta(\mathcal{M})}\min_{p\in\Delta(\Pi)} \mathbb{E}_{\pi\sim p}\mathbb{E}_{M\sim\mu}\left[f_M(\pi_M) - f_M(\pi) - \lambda\cdot \mathrm{KL}\left(M(\pi) \;\|\; \overline{M}^{\mu}(\pi)\right)\right]$$

Theorem

$$\mathrm{BayesRegret} \;\le\; \mathbb{E}\left[\sum_{t=1}^{T}\mathrm{AIR}_\lambda(\mu_t)\right] + \lambda\, H(\mu_1) \;\le\; T\cdot\mathrm{DEC}_\lambda + \lambda\log|\mathcal{M}|$$

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(f_{M^\star}(\pi_{M^\star}) - f_{M^\star}(\pi_t)\right)\right]$$

$$= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_{M\sim\mu_t}[f_M(\pi_M) - f_M(\pi_t)]\right]$$

$$= \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}_{M\sim\mu_t}[f_M(\pi_M) - f_M(\pi_t) - \lambda\cdot \mathrm{IG}_t(\pi_t)]\right] + \left[\sum_{t=1}^{T}\lambda\cdot \mathrm{IG}_t(\pi_t)\right]$$

$$= \mathbb{E}\left[\sum_{t=1}^{T}\min_{p}\mathbb{E}_{\pi\sim p}\mathbb{E}_{M\sim\mu_t}[f_M(\pi_M) - f_M(\pi) - \lambda\cdot \mathrm{IG}_t(\pi)]\right] + \left[\sum_{t=1}^{T}\lambda\left(H(\mu_t) - H(\mu_{t+1})\right)\right]$$

$$\leq \mathbb{E}\left[\sum_{t=1}^{T}\mathrm{AIR}_\lambda(\mu_t)\right] + \lambda\, H(\mu_1)$$

Bayesian Regret Analysis

If $(\mathcal{M}, \Pi, \mathcal{O}, f)$ is a MAB problem: $\mathrm{DEC}_\lambda(\mathcal{M}, \Pi, \mathcal{O}, f) \leq \frac{A}{\lambda}$

$$\mathrm{BayesRegret} \leq \frac{AT}{\lambda} + \lambda \log |\mathcal{M}| \approx \sqrt{AT \log|\mathcal{M}|}$$

If $(\mathcal{M}, \Pi, \mathcal{O}, f)$ is a LB problem: $\mathrm{DEC}_\lambda(\mathcal{M}, \Pi, \mathcal{O}, f) \leq \frac{d}{\lambda}$

$$\mathrm{BayesRegret} \leq \frac{dT}{\lambda} + \lambda \log |\mathcal{M}| \approx \sqrt{dT \log|\mathcal{M}|}$$

Russo and Van Roy. An Information-Theoretic Analysis of Thompson Sampling. 2014.

Russo and Van Roy. Learning to Optimize via Information-Directed Sampling. 2014.

Foster, Kakade, Qian, Rakhlin. The Statistical Complexity of Interactive Decision Making. 2021.

Bayesian Regret Analysis

$$\text{BayesRegret} \le T \cdot \text{DEC}_\lambda + \lambda\, H(\mu_1) \overset{\text{optimal } \lambda}{\approx} \sqrt{T\ \text{PriceOfInfo} \cdot \text{TotalInfo}}$$

$\frac{1}{4\lambda}$ (Price of information)

Total amount of information

$$\text{DEC}_\lambda = \text{OneStepRegret} - \lambda \cdot \text{IG} \overset{\text{AM-GM}}{\le} \frac{1}{4\lambda} \frac{(\text{OneStepRegret})^2}{\text{IG}}$$

Russo and Van Roy. An Information-Theoretic Analysis of Thompson Sampling. 2014.

Russo and Van Roy. Learning to Optimize via Information-Directed Sampling. 2014.

Foster, Kakade, Qian, Rakhlin. The Statistical Complexity of Interactive Decision Making. 2021.

The Frequentist Setting

Decision Making with Structured Observation (DMSO)

Given to learner: decision set Π,
observation set $\mathcal{O}$,
model set $\mathcal{M}$,
payoff function $f\colon \mathcal{M} \times \Pi \to \mathbb{R}$

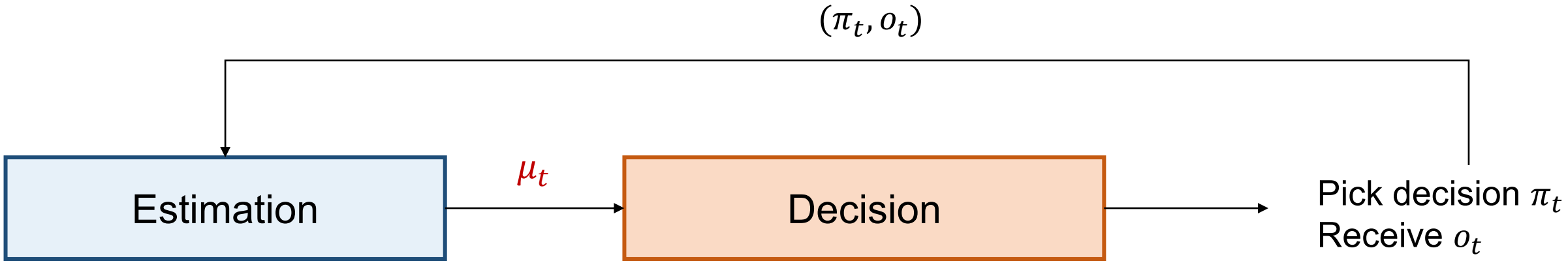
Environment chooses $M^\star \in \mathcal{M}$ (NOT revealed to Learner)

For $t = 1, 2, \ldots, T$:
Learner takes a decision $\pi_t \in \Pi$
Learner observes an observation $o_t \sim M^\star(\pi_t)$

$$\text{Regret} = \mathbb{E}\left[\sum_{t=1}^{T} \left(f_{M^\star}(\pi_{M^\star}) - f_{M^\star}(\pi_t)\right)\right]$$

For fixed $M^\star$

Algorithm Design: Estimation Part



Model posterior update:

$\mu_{t+1}(M) \propto \mu_t(M) \cdot M(o_t|\pi_t)$

with arbitrary $\mu_1(M)$ (e.g. uniform)

same as in the Bayesian setting!

Guarantee for the Estimation Procedure

Theorem

For a fixed $M^\star$ and the update rule $\mu_{t+1}(M) \propto \mu_t(M) \cdot M(o_t|\pi_t)$ with $o_t \sim M^\star(\pi_t)$,

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathrm{KL}\left(M^\star(\pi_t) \,\|\, \overline{M}^{\mu_t}(\pi_t)\right)\right] \leq \log\frac{1}{\mu_1(M^\star)}$$

This recovers the guarantee in the Bayesian setting: Further taking $\mathbb{E}_{M^\star\sim\mu_1}$, the right-hand side becomes

$$\mathbb{E}_{M^\star\sim\mu_1}\left[\log\frac{1}{\mu_1(M^\star)}\right] = H(\mu_1)$$

Guarantee for the Estimation Procedure

$\mu_{t+1}(M) \propto \mu_t(M) \cdot M(o_t|\pi_t)$ $\Leftrightarrow$ $\mu_{t+1}(M) \propto \mu_t(M) \exp(-\ell_t(M))$ with $\ell_t(M) = \log \frac{1}{M(o_t|\pi_t)}$

In fact, the guarantee can be even more general.

For **arbitrary** choice of $M^\star$ and **arbitrary** $o_1, o_2, \ldots, o_T$:

$$\sum_{t=1}^{T} \log \frac{1}{\overline{M}^{\mu_t}(o_t|\pi_t)} - \sum_{t=1}^{T} \log \frac{1}{M^\star(o_t|\pi_t)} \leq \log \frac{1}{\mu_1(M^\star)}$$

$\mathbb{E}_{o_t \sim M^\star(\pi_t)}$

For **arbitrary** choice of $M^\star$ and $o_t \sim M^\star(\pi_t)$:

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathrm{KL}\left(M^\star(\pi_t) \,\|\, \overline{M}^{\mu_t}(\pi_t)\right)\right] \leq \log \frac{1}{\mu_1(M^\star)}$$

For $M^\star \sim \mu_1$ and $o_t \sim M^\star(\pi_t)$:

$\mathbb{E}_{M^\star \sim \mu_1}$

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathrm{KL}\left(M^\star(\pi_t) \,\|\, \overline{M}^{\mu_t}(\pi_t)\right)\right] \leq H(\mu_1)$$

Guarantee for the Estimation Procedure

$$\mu_{t+1}(M) \propto \mu_t(M) \cdot M(o_t|\pi_t) \quad \Leftrightarrow \quad \mu_{t+1}(M) \propto \mu_t(M) \exp(-\ell_t(M)) \text{ with } \ell_t(M) = \log \frac{1}{M(o_t|\pi_t)}$$

Implication:

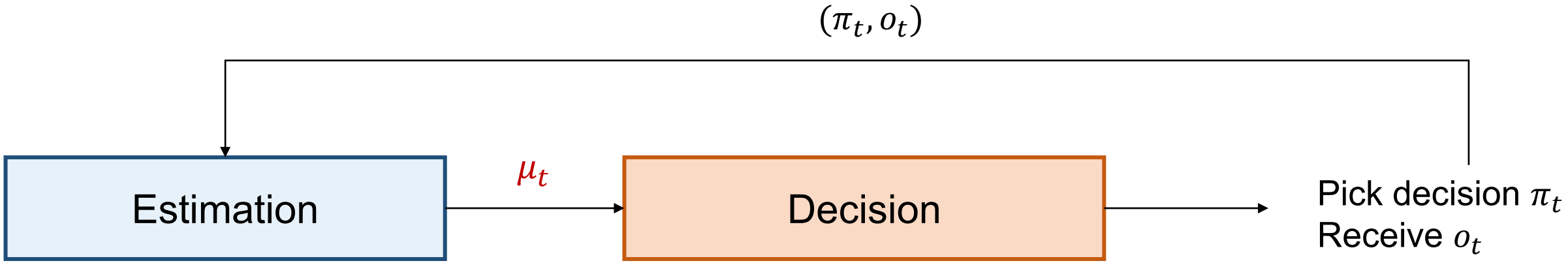
The estimation procedure (i.e., posterior update) need not rely on distributional assumptions / have distributional interpretations.

We can more generally view it as

A loss-minimization procedure with a particular loss function

(loss function = reflect how well a model explains the observation)

Algorithm Design: Estimation Part



Model posterior update:

$$\mu_{t+1}(M) \propto \mu_t(M) \cdot M(o_t|\pi_t)$$

Guarantee for μ_t:

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathrm{KL}\left(M^\star(\pi_t) \,\|\, \overline{M}^{\mu_t}(\pi_t)\right)\right] \leq \log|\mathcal{M}|$$

Algorithm Design: Decision Part

Just do the same as in the Bayesian setting?

$$p_t = \operatorname*{argmin}_{p \in \Delta(\Pi)} \mathbb{E}_{M \sim \mu_t} \mathbb{E}_{\pi \sim p} \left[\underbrace{f_M(\pi_M) - f_M(\pi)}_{\text{one-step regret}} - \lambda \cdot \underbrace{\mathrm{KL}\left(M(\pi) \;\|\; \overline{M}^{\mu_t}(\pi) \right)}_{\text{Information gain}} \right]$$

Unfortunately, this doesn't work well. (See Zhang (2021) for an example)

We will modify the decision rule as

$$p_t = \operatorname*{argmin}_{p \in \Delta(\Pi)} \max_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p} \left[f_M(\pi_M) - f_M(\pi) - \lambda \cdot \mathrm{KL}\left(M(\pi) \;\|\; \overline{M}^{\mu_t}(\pi) \right) \right]$$

Zhang. Feel-Good Thompson Sampling for Contextual Bandits and Reinforcement Learning. 2021.

Interpretation of the Decision Rule

$$p_t = \operatorname*{argmin}_{p \in \Delta(\Pi)} \max_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p} \left[f_M(\pi_M) - f_M(\pi) - \lambda \cdot \underbrace{\mathrm{KL}\left(M(\pi) \;\|\; \overline{M}^{\mu_t}(\pi) \right)}_{\text{Discrepancy between } M \text{ and } \overline{M}^{\mu_t} \text{ under decision } \pi} \right]$$

$$\min_{p \in \Delta(\Pi)} \max_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p} \left[f_M(\pi_M) - f_M(\pi) \right]$$

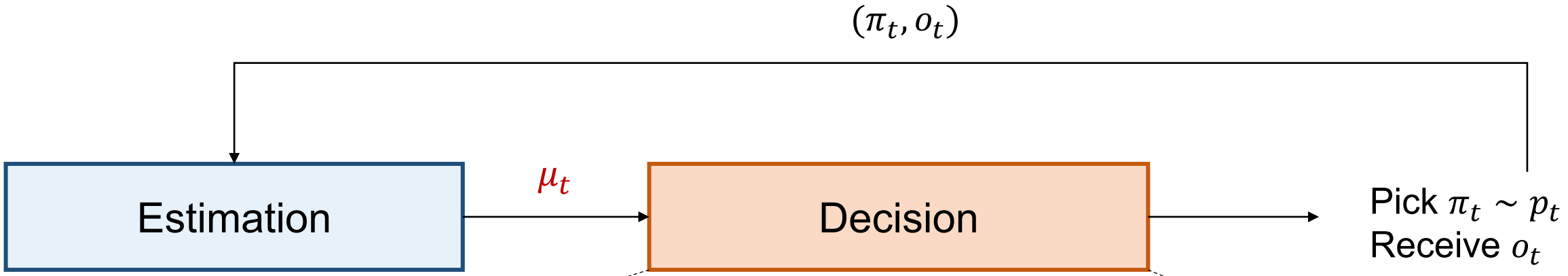
(Impossible to make this small)

M_1	M_2
Arm 1 = Ber(0.5)	Arm 1 = Ber(0.5)
Arm 2 = Ber(0.4)	Arm 2 = Ber(0.6)

$$\min_{p \in \Delta(\Pi)} \max_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p} \left[f_M(\pi_M) - f_M(\pi) - \lambda \cdot \mathrm{Discrepancy}_\pi(M, \mathrm{EstimationFromThePast}) \right]$$

(Regularized game, anchored at the learner's model estimation from past data)

Frequentist Decision Rule



Model posterior update:

$$\mu_{t+1}(M) \propto \mu_t(M) \cdot M(o_t|\pi_t)$$

$$p_t = \operatorname*{argmin}_{p \in \Delta(\Pi)} \max_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p} \left[f_M(\pi_M) - f_M(\pi) - \lambda \cdot \mathrm{KL}\left(M(\pi) \;\|\; \overline{M}^{\mu_t}(\pi) \right) \right]$$

one-step regret

Information gain
(discrepancy from the estimated model)

E2D

Foster, Kakade, Qian, Rakhlin. The Statistical Complexity of Interactive Decision Making. 2021.

Frequentist Regret Analysis

$$\overline{\mathrm{AIR}}_\lambda(\mu_t) \coloneqq \min_{p\in\Delta(\Pi)} \max_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\left[f_M(\pi_M) - f_M(\pi) - \lambda\cdot \mathrm{KL}\left(M(\pi) \;\|\; \overline{M}^{\mu_t}(\pi)\right)\right]$$

Theorem

$$\text{Regret} \;\le\; \mathbb{E}\left[\sum_{t=1}^{T} \overline{\mathrm{AIR}}_\lambda(\mu_t)\right] + \lambda\log|\mathcal{M}| \le \;\; T\cdot \mathrm{DEC}_\lambda + \lambda\log|\mathcal{M}|$$

The same upper bound we established for the Bayesian setting

Foster, Kakade, Qian, Rakhlin. The Statistical Complexity of Interactive Decision Making. 2021.

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(f_{M^\star}(\pi_{M^\star}) - f_{M^\star}(\pi_t)\right)\right]$$

$$= \mathbb{E}\left[\sum_{t=1}^{T} f_{M^\star}(\pi_{M^\star}) - f_{M^\star}(\pi_t) - \lambda \cdot \mathrm{KL}\left(M^\star(\pi_t), \overline{M}^{\mu_t}(\pi_t)\right)\right] + \left[\sum_{t=1}^{T} \lambda \cdot \mathrm{KL}\left(M^\star(\pi_t), \overline{M}^{\mu_t}(\pi_t)\right)\right]$$

$$\leq \mathbb{E}\left[\sum_{t=1}^{T} \max_{M\in\mathcal{M}} f_M(\pi_M) - f_M(\pi_t) - \lambda \cdot \mathrm{KL}\left(M(\pi_t), \overline{M}^{\mu_t}(\pi_t)\right)\right] + \lambda \log|\mathcal{M}|$$

$$= \mathbb{E}\left[\sum_{t=1}^{T} \min_{p\in\Delta(\Pi)} \max_{M\in\mathcal{M}} \mathbb{E}_{\pi\sim p}\left[f_M(\pi_M) - f_M(\pi) - \lambda \cdot \mathrm{KL}\left(M(\pi), \overline{M}^{\mu_t}(\pi)\right)\right]\right] + \lambda \log|\mathcal{M}|$$

$$= \mathbb{E}\left[\sum_{t=1}^{T} \overline{\mathrm{AIR}}_\lambda(\mu_t)\right] + \lambda \log|\mathcal{M}|$$

$$\overline{\mathrm{AIR}}_\lambda(\mu_t) = \min_{p\in\Delta(\Pi)} \max_{\mu\in\Delta(\mathcal{M})} \mathbb{E}_{\pi\sim p}\, \mathbb{E}_{M\sim\mu}\left[f_M(\pi_M) - f_M(\pi) - \lambda\cdot \mathrm{KL}\left(M(\pi), \overline{M}^{\mu_t}(\pi)\right)\right]$$

$$= \max_{\mu\in\Delta(\mathcal{M})} \min_{p\in\Delta(\Pi)} \mathbb{E}_{\pi\sim p}\, \mathbb{E}_{M\sim\mu}\left[f_M(\pi_M) - f_M(\pi) - \lambda\cdot \mathrm{KL}\left(M(\pi), \overline{M}^{\mu_t}(\pi)\right)\right]$$

$$\leq \max_{\mu\in\Delta(\mathcal{M})} \min_{p\in\Delta(\Pi)} \mathbb{E}_{\pi\sim p}\, \mathbb{E}_{M\sim\mu}\left[f_M(\pi_M) - f_M(\pi) - \lambda\cdot \mathrm{KL}\left(M(\pi), \overline{M}^{\mu}(\pi)\right)\right]$$

$$= \mathrm{DEC}_\lambda$$

How Fundamental Is This Regret Bound?

We have designed algorithms achieving a regret upper bound of

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(f_{M^\star}(\pi_{M^\star}) - f_{M^\star}(\pi_t)\right)\right] \leq \quad T \cdot \mathrm{DEC}_\lambda + \lambda \cdot \log|\mathcal{M}|$$

Is there a matching / nearly-matching lower bound?

How Fundamental Is This Regret Bound?

Qualitatively, DEC_λ is a necessary price to pay.

$$\mathrm{DEC}_\lambda = \max_{\mu \in \Delta(\mathcal{M})} \min_{p \in \Delta(\Pi)} \mathbb{E}_{\pi \sim p} \mathbb{E}_{M \sim \mu} \Big[\underbrace{f_M(\pi_M) - f_M(\pi)}_{\text{one-step regret}} - \lambda \cdot \underbrace{\mathrm{KL}\Big(M(\pi), \overline{M}^{\mu}(\pi)\Big)}_{\text{Information gain}} \Big]$$

Large DEC_λ means there exists a model distribution μ such that **for any decision π:**

- The expected one-step regret is large
  ⇒ the models under μ disagree a lot on the optimal decisions
- The information gain is small
  ⇒ it's hard for the learner to reduce the entropy of their posterior

} Not precise

However, $\log |\mathcal{M}|$ may not always be necessary.

Arm 1 = Ber(0.5)	Arm 1 = Ber(0.5)
Arm 2 = Ber(0.4)	Arm 2 = Ber(0.6)

Summary

Decision Making with Structured Observations (DMSO)

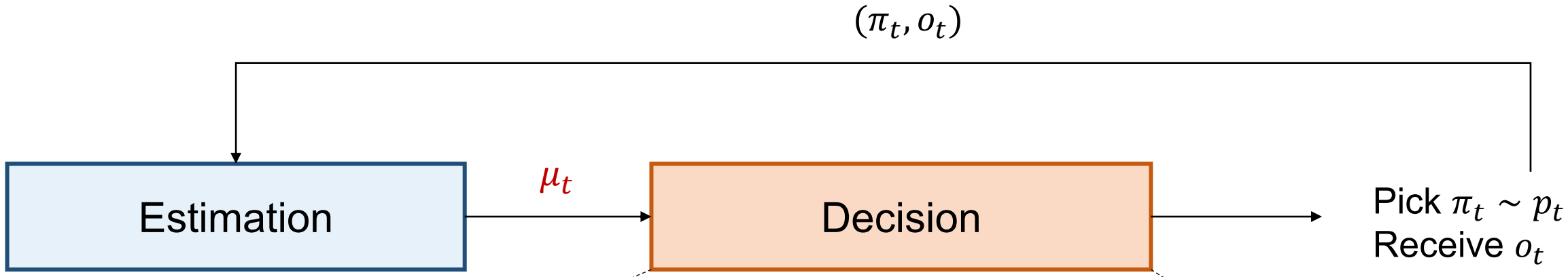
Given to learner: decision set Π,
observation set $\mathcal{O}$,
model set $\mathcal{M}$,
payoff function f: $\mathcal{M} \times \Pi \rightarrow \mathbb{R}$

Environment chooses $M^\star \in \mathcal{M}$ (NOT revealed to Learner)

For $t = 1, 2, \ldots, T$:
Learner takes a decision $\pi_t \in \Pi$
Learner observes an observation $o_t \sim M^\star(\pi_t)$

$$\text{Regret} = \mathbb{E}\left[\sum_{t=1}^{T} \left(f_{M^\star}(\pi_{M^\star}) - f_{M^\star}(\pi_t)\right)\right]$$

Summary: Algorithm in the Bayesian Setting



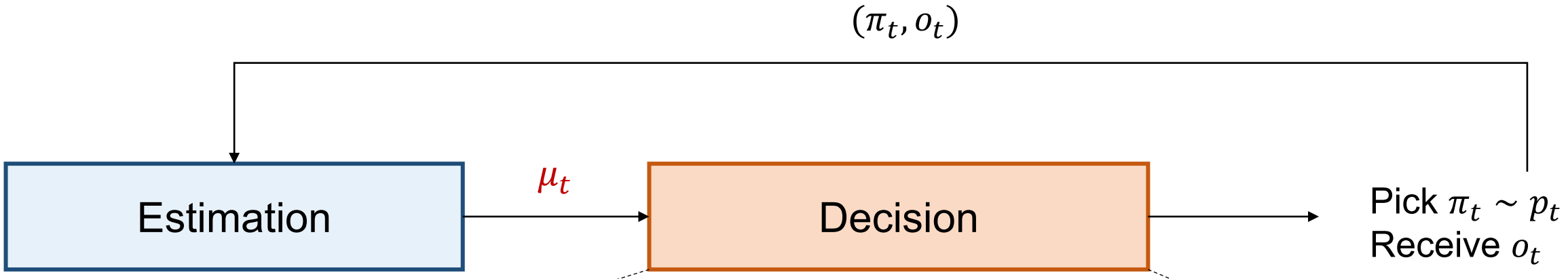
Model posterior update:

$\mu_{t+1}(M) \propto \mu_t(M) \cdot M(o_t|\pi_t)$

$$p_t = \operatorname*{argmin}_{p \in \Delta(\Pi)} \mathbb{E}_{\pi \sim p} \mathbb{E}_{M \sim \mu_t} \left[\underbrace{f_M(\pi_M) - f_M(\pi)}_{\text{one-step regret}} - \lambda \cdot \underbrace{\mathrm{KL}\left(M(\pi) \;\|\; \overline{M}^{\mu_t}(\pi) \right)}_{\text{Information gain}} \right]$$

E2D.Bayes

Summary: Algorithm in the Frequentist Setting



Model posterior update:

$\mu_{t+1}(M) \propto \mu_t(M) \cdot M(o_t|\pi_t)$

$$p_t = \operatorname*{argmin}_{p \in \Delta(\Pi)} \max_{M \in \mathcal{M}} \mathbb{E}_{\pi \sim p} \left[\underbrace{f_M(\pi_M) - f_M(\pi)}_{\text{one-step regret}} - \lambda \cdot \underbrace{\mathrm{KL}\left(M(\pi) \;\|\; \overline{M}^{\mu_t}(\pi) \right)}_{\text{Information gain}} \right]$$

(discrepancy from the estimated model)

E2D

Summary: Regret Upper Bound

For both Bayesian and Frequentist settings, we show

$$\text{Regret} = \mathbb{E}\left[\sum_{t=1}^{T}\left(f_{M^\star}(\pi_{M^\star}) - f_{M^\star}(\pi_t)\right)\right] \le T \cdot \text{DEC}_\lambda + \lambda \log |\mathcal{M}|$$

$\frac{1}{4\lambda}$ (Price of information)

Total amount of information

$$\text{optimal } \lambda \approx \sqrt{T \cdot \text{PriceOfInfo} \cdot \text{TotalInfo}}$$

Remark: A DEC that better characterizes the bound

$$\mathrm{CDEC}_\epsilon = \max_{\mu\in\Delta(\mathcal{M})} \min_{p\in\Delta(\Pi)} \max_{M} \left\{ \mathbb{E}_{\pi\sim p}[f_M(\pi_M) - f_M(\pi)] \;\middle|\; D^2_{\mathrm{H}}\left(M(\pi), \overline{M}^{\mu}(\pi)\right) \le \epsilon^2 \right\}$$

(replacing the "regularization" in DEC_λ by "constraint")

It has been shown that the minimax regret is

$$\Omega\left(T \cdot \mathrm{CDEC}_{1/\sqrt{T}}\right) \qquad \text{and} \qquad \mathrm{O}\left(T \cdot \mathrm{CDEC}_{1/\sqrt{T}} \log|\mathcal{M}|\right)$$

Foster, Golowich, Han. Tight Guarantees for Interactive Decision Making with the Decision-Estimation Coefficient. 2023.

Part III. Beyond Model Estimation

Outline

Part I. Reinforcement Learning and the Exploration-Exploitation tradeoff

- Reinforcement learning vs. supervised learning
- Multi-armed bandits
- Linear bandits

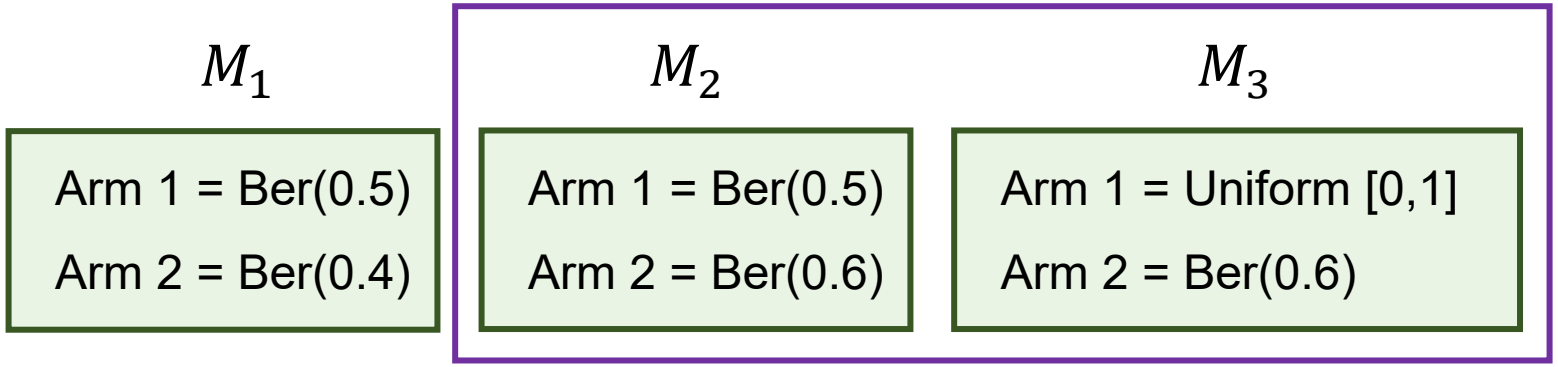
Part II. Information-theoretic algorithm design for the Exploration-Exploitation tradeoff

- Decision-making with structured observation (DMSO)
- Bayesian setting
- Frequentist setting

Part III. Beyond model estimation

- Value estimation / policy estimation
- Open problems

Not All Information Is Useful

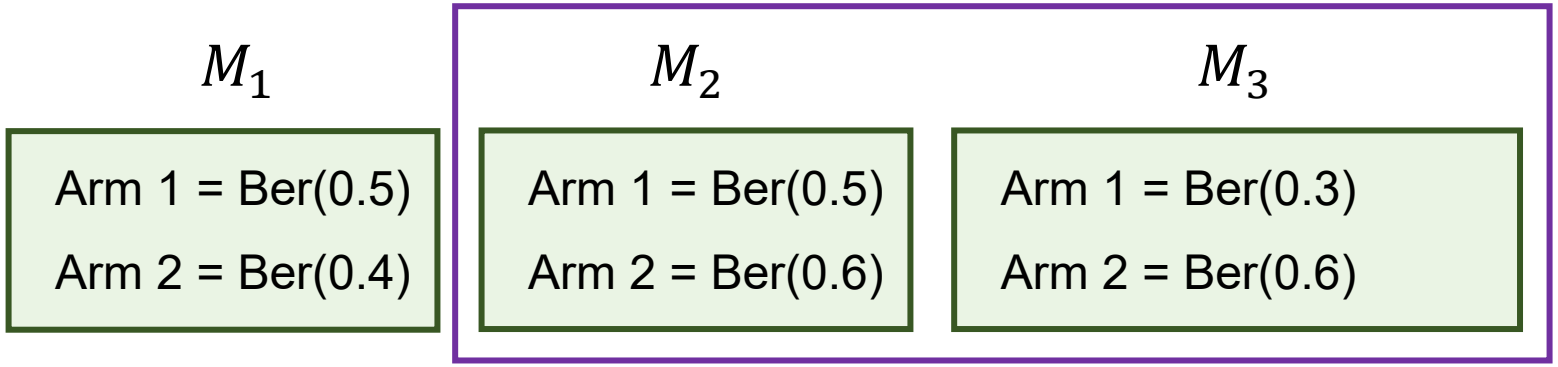


These models always induce the same **expected value**

Distinguishing between M_2 and M_3 does reduce the entropy of posterior, but **may not be useful** for regret minimization…

(This is a source of gap between existing upper/lower regret bounds for DMSO)

Not All Information Is Useful

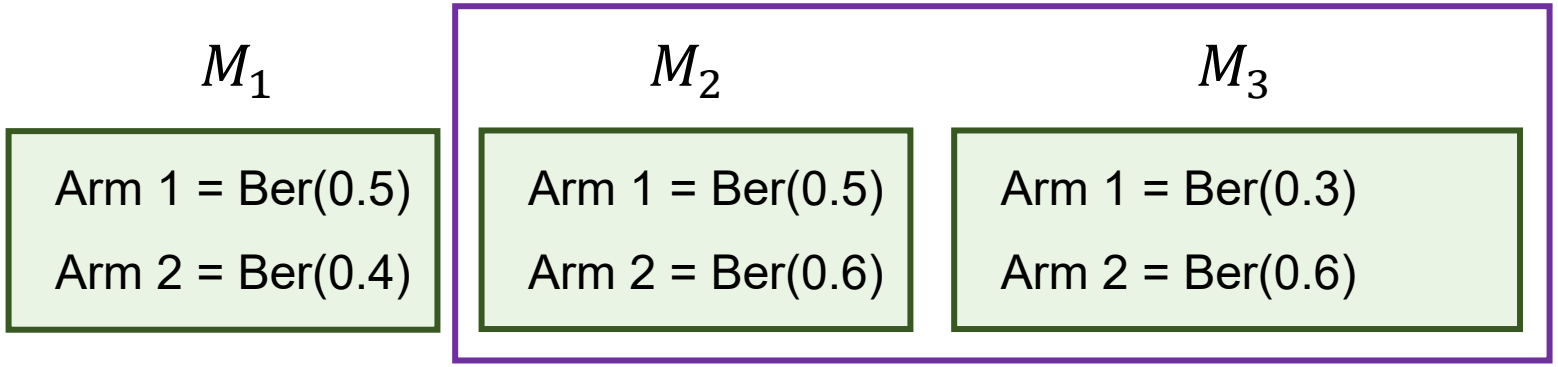


These models induce the same **optimal decision**

Distinguishing between M_2 and M_3 does reduce the entropy of posterior,
but **may not be useful** for regret minimization…

(This is a source of gap between existing upper/lower regret bounds for DMSO)

Not All Information Is Useful



These two models induce the same **optimal decision**

Question: can we define some "decision-relevant" information gain?

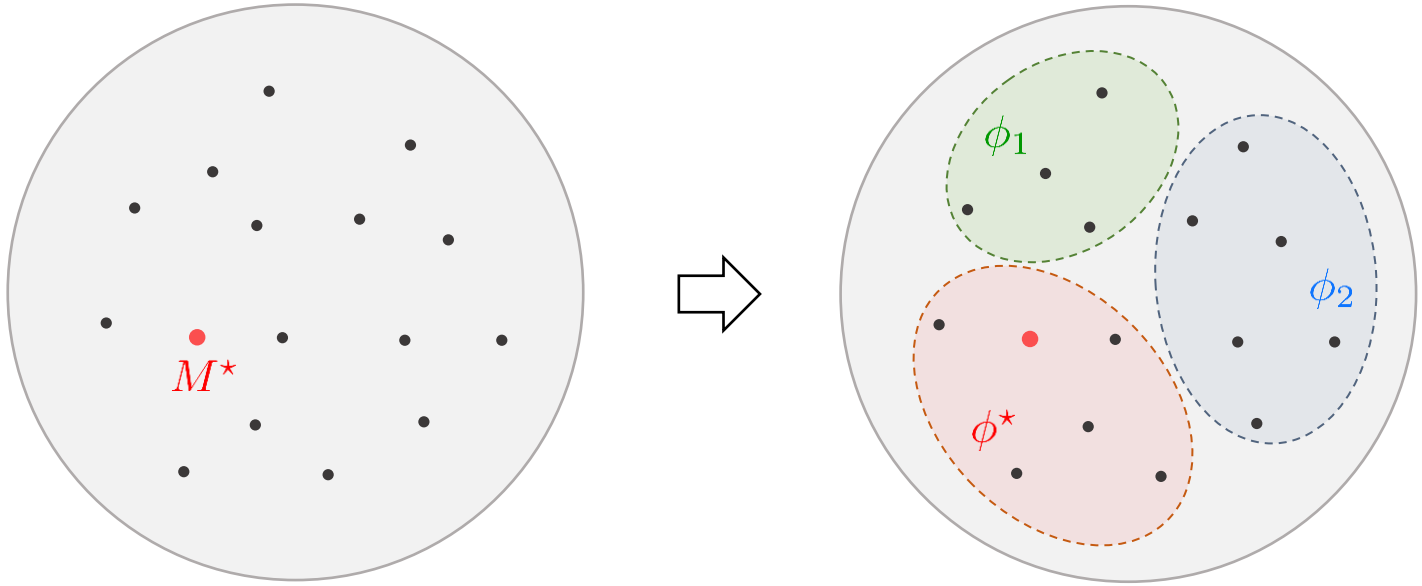
(Still relatively open, I believe)

Coarser Estimation

Comparison

	Estimation	Decision
ϵ-Greedy	Mean $\hat{R}_t$	$\operatorname{argmax} \hat{R}_t$ + uniform
Boltzmann	Mean $\hat{R}_t$	$\operatorname{softmax} \hat{R}_t$
UCB	Mean $\hat{R}_t$ + Uncertainty U_t	$\operatorname{argmax} \left(\hat{R}_t + U_t\right)$
E2D	Mixture Model $\overline{M}^{\mu_t}$	min max {……}

Coarser Estimation



For example,

$\phi_f = \{M\colon\ f_M = f\}$ (grouping models with the same payoff function)

$\phi_\pi = \{M\colon\ \pi_M = \pi\}$ (grouping models with the same optimal policy)

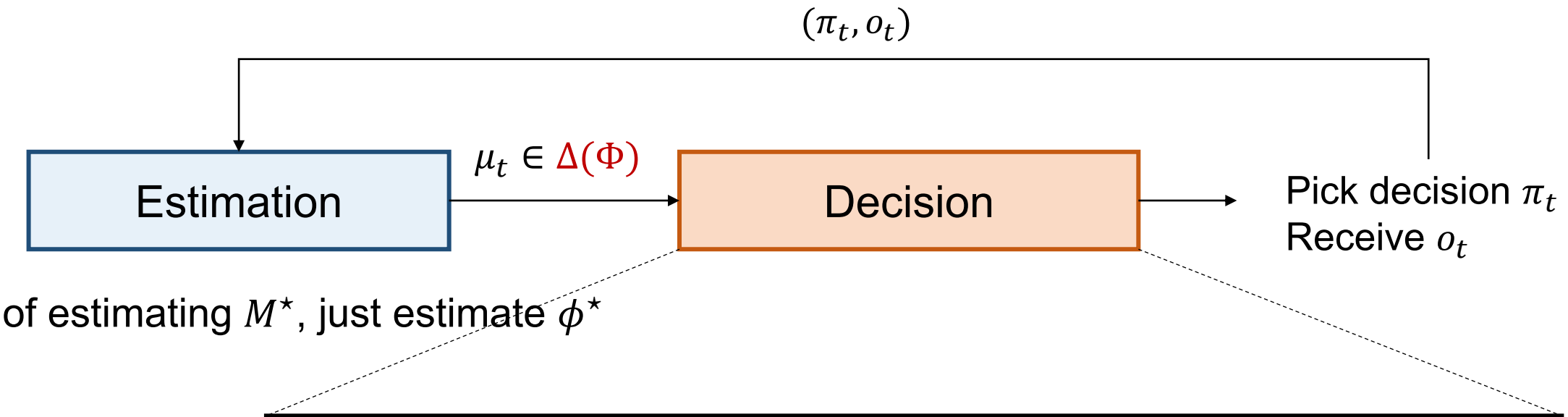
(there are other natural grouping ways in problems with more structure, e.g., MDPs)

Then define $\Phi = \{\phi_f\}$ or $\Phi = \{\phi_\pi\}$

Coarser Estimation

Unclear how to do Bayesian update over $\mu_t \in \Delta(\Phi)$

Sol. Hallucinate some model distribution ν whose marginal distribution is close to μ_t



Instead of estimating $M^\star$, just estimate $\phi^\star$

$$p_t = \operatorname*{argmin}_{p \in \Delta(\Pi)} \max_{\nu \in \Delta(\mathcal{M})} \mathbb{E}_{\pi \sim p} \mathbb{E}_{M \sim \nu} \left[\dots\dots - \lambda \left(I_\nu(\phi^\star; o_t \mid \pi) + \mathrm{KL}(\nu_\Phi, \mu_t) \right) \right]$$

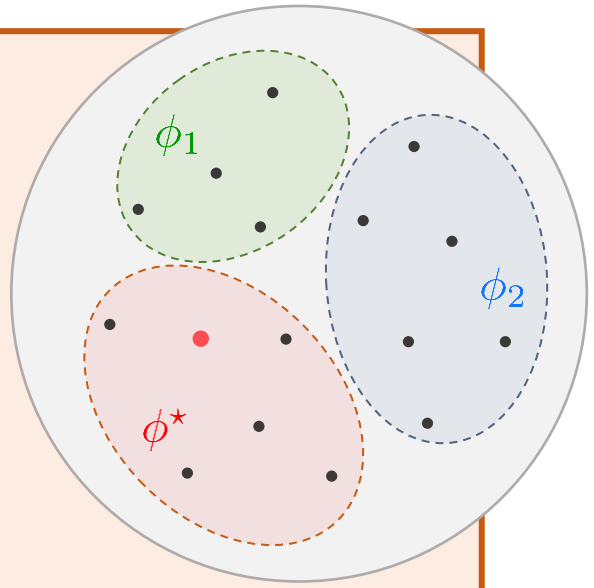
$$\| \quad \mathbb{E}_{\phi \sim \nu_\Phi} \left[\mathrm{KL} \left(\overline{M}^{\nu(\cdot \mid \phi)}(\pi) \;\middle\|\; \overline{M}^{\nu}(\pi) \right) \right]$$

Xu and Zeevi. Bayesian Design Principles for Frequentist Sequential Learning. 2023.

Liu, Wei, Zimmert. Decision Making in Hybrid Environments: A Model Aggregation Approach. 2025.

Coarser Estimation Allows for Time-Varying Environments

Given to learner: decision set Π,
observation set $\mathcal{O}$,
model set $\mathcal{M}$,
payoff function $f\colon \mathcal{M} \times \Pi \to \mathbb{R}$



Environment chooses $\phi^\star \in \Phi$ (NOT revealed to Learner)

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

Environment chooses $M_t \in \phi^\star$ arbitrarily
Learner takes a decision $\pi_t \in \Pi$
Learner observes an observation $o_t \sim M_t(\pi_t)$

Of course, this is not for free:

$$\min \max \left[\mathrm{reg}_t - \lambda \cdot \mathrm{IG}_t^{\mathcal{M}} \right] \leq \min \max \left[\mathrm{reg}_t - \lambda \cdot \mathrm{IG}_t^{\Phi} \right] \quad \text{(data-processing ineq.)}$$

(DEC_λ) $\qquad$ $(\Phi - \mathrm{DEC}_\lambda)$

Coarser Estimation Allows for Time-Varying Environments

Regret Upper Bound for the original DMSO:

$$T \cdot \mathrm{DEC}_\lambda + \lambda \log |\mathcal{M}|$$

Estimating $M^\star \in \mathcal{M}$

$$\mathrm{DEC}_\lambda \leq \Phi - \mathrm{DEC}_\lambda$$
$$\log |\mathcal{M}| \geq \log |\Phi|$$

$$T \cdot \Phi - \mathrm{DEC}_\lambda + \lambda \log |\Phi|$$

Estimating $\phi^\star \in \Phi$

Coarser Estimation Allows for Time-Varying Environments

If the grouping is based on decisions (e.g., arms), the framework handles **completely non-stationary** reward.

Adversarial Multi-Armed Bandits

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

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

Environment chooses $(R_t(1), R_t(2), \dots, R_t(A))$ arbitrarily.

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

Learner observes $r_t = R_t(a_t) + \text{noise}$

$$\textbf{Regret} := \max_a \sum_{t=1}^{T} R_t(a) - \sum_{t=1}^{T} R_t(a_t) \leq O\left(\sqrt{AT \log A}\right)$$

Xu and Zeevi. Bayesian Design Principles for Frequentist Sequential Learning. 2023.

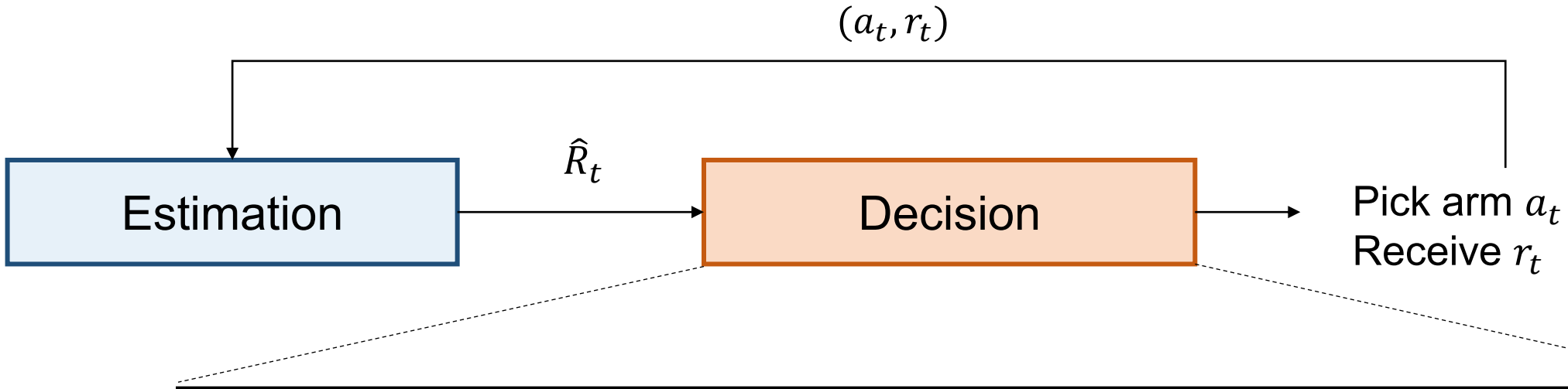
Coarser Estimation Allows for Time-Varying Environments

Adversarial MAB is a classic setting, and the most classic algorithm is **EXP3**.

- Algorithms for this setting does not estimate reward distribution or reward function.
- Instead, they maintain a distributions over decisions (arms).
- Solving the min-max program we just discussed under e.g. Gaussian reward assumptions reproduces a similar update rule as EXP3.

E2D with Mean Estimation

The Estimation Module Is Flexible



$$p_t = \underset{p \in \Delta(\mathcal{A})}{\operatorname{argmin}} \max_{f \in \mathbb{R}^A} \mathbb{E}_{a \sim p} \left[f(a_f) - f(a) - \lambda \left(f(a) - \hat{R}_t(a) \right)^2 \right]$$

$$\Rightarrow \quad p_t(a) = \frac{1}{\gamma - \lambda \hat{R}_t(a)}$$ where γ is a normalization factor that makes $\sum_a p_t(a) = 1$

(Inverse-Gap Weighting)

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

A New (?) MAB Algorithm

	Estimation	Decision	Regret Bound
ϵ-Greedy	Mean $\hat{R}_t$	$\text{argmax}\ \hat{R}_t(a)$ + uniform	$A^{1/3}T^{2/3}$
Boltzmann	Mean $\hat{R}_t$	$\propto \exp\left(\hat{R}_t(a)\right)$	--
IGW	Mean $\hat{R}_t$	$\frac{1}{\gamma - \hat{R}_t(a)}$	$\sqrt{AT}$
UCB	Mean $\hat{R}_t$ + Uncertainty U_t	$\text{argmax}\left(\hat{R}_t(a) + U_t(a)\right)$	$\sqrt{AT}$

The power of IGW actually lies in contextual bandits problems.

Open Problems

1. Better Characterization of the Minimax Regret in DMSO

Given to learner: decision set Π,
observation set $\mathcal{O}$,
model set $\mathcal{M}$,
payoff function f: $\mathcal{M} \times \Pi \rightarrow \mathbb{R}$

Environment chooses $M^\star \in \mathcal{M}$ (NOT revealed to Learner)

For $t = 1, 2, \ldots, T$:
Learner takes a decision $\pi_t \in \Pi$
Learner observes an observation $o_t \sim M^\star(\pi_t)$

$$\text{Minimiax regret} = \min_{\text{algorithm}} \max_{M^\star} \mathbb{E}\left[\sum_{t=1}^{T} \left(f_{M^\star}(\pi_{M^\star}) - f_{M^\star}(\pi_t)\right)\right]$$

1. Better Characterization of the Minimax Regret in DMSO

Lower Bound

$$T \cdot \mathrm{CDEC}_{T^{-1/2}}$$

Upper Bounds

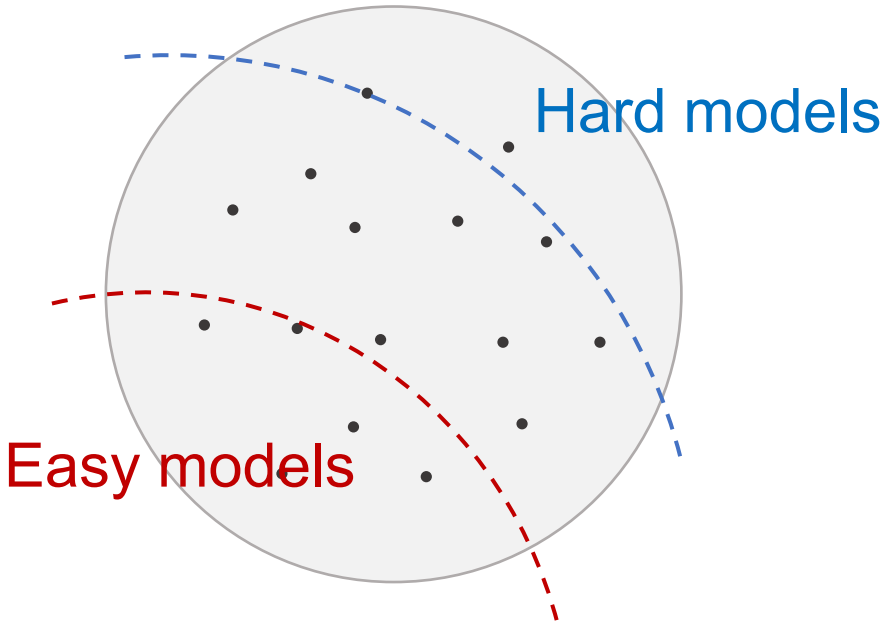
$$T \cdot \mathrm{CDEC}_{T^{-1/2}} \ \log|\mathcal{M}|$$

$$T \cdot \mathrm{DEC}_{\lambda} + \lambda \log|\mathcal{M}|$$

$$T \cdot \Phi - \mathrm{DEC}_{\lambda} + \lambda \log|\Phi|$$

CDEC: constrained DEC (replacing the "regularization" in DEC_{λ} by "constraint")

2. Beyond Minimax Guarantee



Better regret bound if $\mathcal{M}^\star$ is an easy model?