RL THEORY VIRTUAL SEMINARS

Crush Optimism with Pessimism: Structured Bandits Beyond Asymptotic Optimality

Kwang-Sung Jun

join work with Chicheng Zhang

THE UNIVERSITY OF ARIZONA

Structured bandits

e.g., linear $\mathcal{A} = \{a^1, \dots, a^K \ \in \mathbb{R}^d\}$
$\mathcal{F} = \{a \mapsto \theta^\top a \colon \theta \in \mathbb{R}^d\}$

- **Input**: Arm set $\mathcal{A}$, hypothesis class $\mathcal{F} \subset (\mathcal{A} \to \mathbb{R})$

"the set of possible **configurations** of the **mean rewards**"

- **Initialize**: The environment chooses $f^* \in \mathcal{F}$ (unknown to the learner)

  For $t = 1, \dots, n$
  - Learner: chooses an arm $a_t \in \mathcal{A}$
  - Environment: generates the reward $r_t = f^*(a_t) +$ (zero-mean stochastic noise)
  - Learner: receives r_t

- **Goal**: Minimize the cumulative **regret**

$$\mathbb{E}\,\mathrm{Reg}_n = \mathbb{E}\left[n \cdot \left(\max_{a \in \mathcal{A}} f^*(a)\right) - \sum_{t=1}^{n} f^*(a_t)\right]$$

- Note: fixed arm set (=non-contextual), **realizability** $(f^* \in \mathcal{F})$

Structured bandits

- **Why relevant?**
  Techniques may transfer to RL (e.g., ergodic RL [Ok18])
- **Naive strategy: UCB**
  $\Longrightarrow$ $\frac{K}{\Delta}\log n$ regret bound (instance-dependent)
  - Scales with the number of arms K
  - Instead, the **complexity** of the hypothesis class $\mathcal{F}$ should appear.
- The **asymptotically optimal regret** is well-defined.
  - E.g., linear bandits : $c^* \cdot \log n$ for some well-defined $c^* \ll \frac{K}{\Delta}$.

The goal of this paper
Achieve the **asymptotic optimality** with improved **finite-time** regret for any $\mathcal{F}$.

(the worst-case regret is beyond the scope)

[Ok18] Ok et al., Exploration in Structured Reinforcement Learning, NeurIPS, 2018

Asymptotic optimality (instance-dependent)

- **Optimism** in the face of uncertainty (e.g., UCB, Thompson sampling)

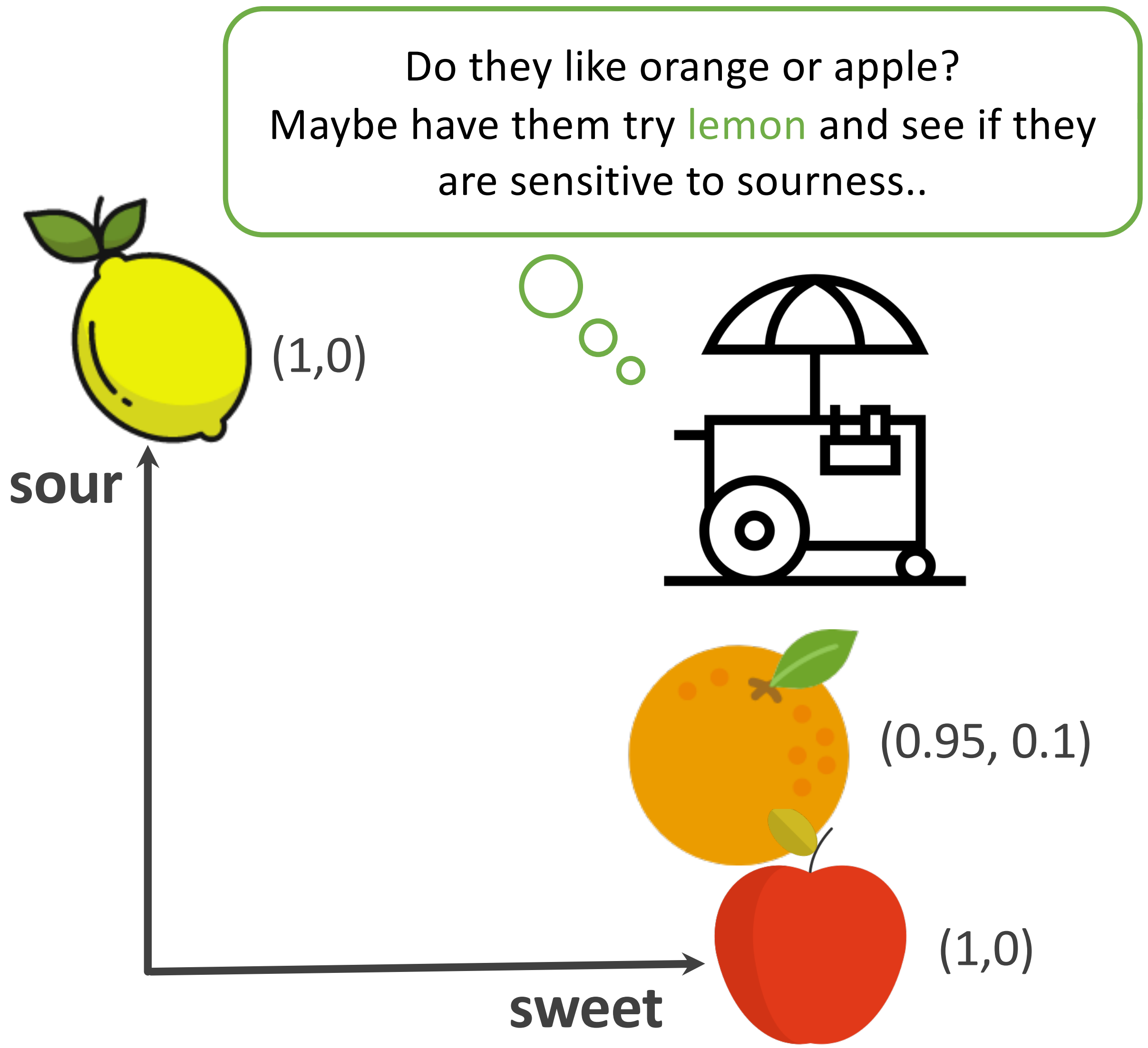
  $\Longrightarrow$ optimal asymptotic / worst-case regret in $\boldsymbol{K}$**-armed bandits**.
- Linear bandits: optimal worst-case rate = $d\sqrt{n}$
- Asymptotically optimal regret? $\Longrightarrow$ **No!**

The End of Optimism?
An Asymptotic Analysis of Finite-Armed Linear Bandits

Tor Lattimore
Indiana University, Bloomington

Csaba Szepesvári
University of Alberta, Edmonton

(AISTATS'17)



mean reward = 1*sweet + 0*sour

Asymptotic optimality: lower bound

- $\mathbb{E}\,\mathrm{Reg}_n \geq c(f^*) \cdot \log n$ (asymptotically)

$\Delta_a = \left(\max_{b \in \mathcal{A}} f^*(b)\right) - f^*(a)$

$$c(f^*) = \min_{\gamma_1, \dots, \gamma_K \geq 0} \sum_{a=1}^{K} \gamma_a \cdot \Delta_a$$

$$\text{s.t.} \qquad \gamma_{a^*(f)} = 0$$

$$\forall g \in \mathcal{C}(f^*), \qquad \sum_{a=1}^{K} \gamma_a \cdot \mathrm{KL}_\nu\big(f(a), g(a)\big) \geq 1$$

"competing" hypotheses

KL divergence with noise distribution ν

- $\gamma^* = (\gamma_1^*, \dots, \gamma_K^*) \geq 0$: the solution
- To be optimal, we must pull arm a like $\gamma_a^* \cdot \log n$ times.
- E.g., $\gamma_{\text{lemon}}^* = 8, \quad \gamma_{\text{orange}}^* = 0 \quad \Longrightarrow$ lemon is the **informative arm**!
- When $c(f^*) = 0$: **Bounded regret**! (except for pathological ones [Lattimore14])

Lattimore & Munos, Bounded regret for finite-armed structured bandits, 2014.

Existing asymptotically optimal algorithms

- Mostly uses forced exploration. [Lattimore+17,Combes+17,Hao+20]

  $\Longrightarrow$ ensures **every arm**'s pull count is an **unbounded** function of n such as $\frac{\log n}{1+\log\log n}$.

  $\Longrightarrow \mathbb{E}\,\mathrm{Reg}_n \lessapprox c(f^*) \cdot \log n + K \cdot \frac{\log n}{1+\log\log n}$

- Issues

  1. K appears in the regret* $\Longrightarrow$ what if K is exponentially large?
  2. **cannot** achieve **bounded** regret when $c(f^*) = 0$

- Parallel studies avoid forced exploration, but still depend on K. [Menard+20, Degenne+20]

*Dependence on K can be avoided in special cases (e.g., linear).

Contribution

Research Question

Assume $\mathcal{F}$ is finite. Can we design an algorithm that

- enjoys the **asymptotic optimality**
- adapts to **bounded regret** whenever possible
- does not necessarily depend on K?

Proposed algorithm:
CRush Optimism with Pessimism (CROP)

✓

✓

✓

- No forced exploration 😀
- The regret scales not with K but with $K_\psi \le K$ (defined in the paper).
- An interesting $\log\log n$ term in the regret*

* it's necessary (will be updated in arxiv)

Preliminaries

Assumptions

- $|\mathcal{F}| < \infty$
- The noise model

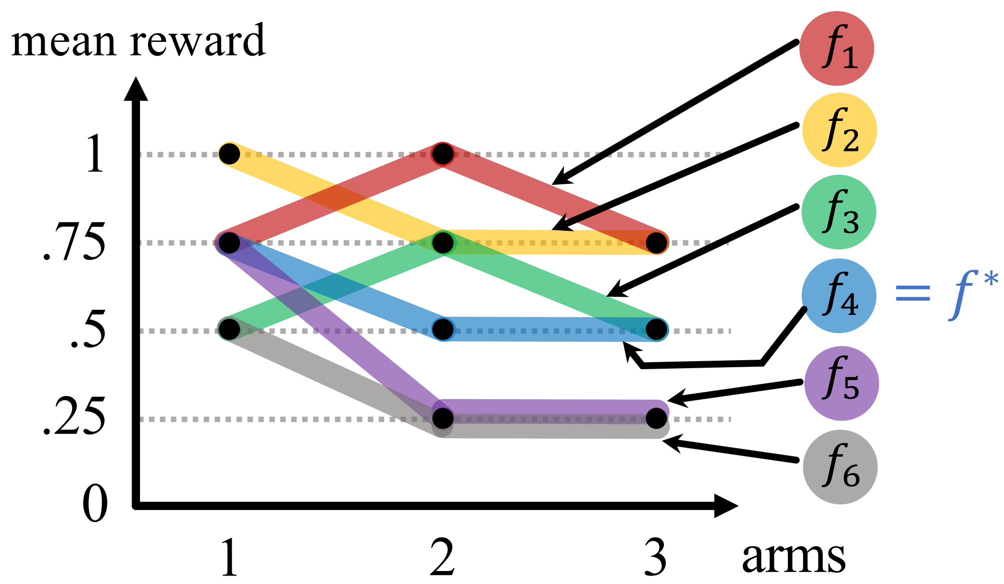
$$r_t = f^*(a_t) + \xi_t \qquad \text{where} \qquad \xi_t \text{ is 1-sub-Gaussian.}$$ (generalized to σ^2 in the paper)

- Notations: $a^*(f) \coloneqq \arg\max_{a \in \mathcal{A}} f(a)$, $\quad \mu^*(f) \coloneqq f\big(a^*(f)\big)$
- f **supports** arm a $\quad \Leftrightarrow \quad a^*(f) = a$
- f **supports** reward v $\quad \Leftrightarrow \quad \mu^*(f) = v$

- **[Assumption]** Every $f \in \mathcal{F}$ has a unique best arm (i.e., $|a^*(f)| = 1$)

Competing hypotheses

- $\mathcal{C}(f^*)$ consists of $f \in \mathcal{F}$ such that
  - (1) assigns the same reward to the best arm $a^*(f^*)$
  - (2) but supports a different arm $a^*(f) \neq a^*(f^*)$
- Importance: it's why we get $\log(n)$ regret!



Lower bound revisited

- Assume Gaussian rewards.
- $\mathbb{E}\,\mathrm{Reg}_n \geq c(f^*) \cdot \log n$, asymptotically.

$$\Delta_a = \left(\max_{b \in \mathcal{A}} f^*(b)\right) - f^*(a)$$

$$c(f^*) := \min_{\gamma_1,\ldots,\gamma_K \geq 0} \sum_{a=1}^{K} \gamma_a \cdot \Delta_a$$

$$\text{s.t.} \qquad \gamma_{a^*(f^*)} = 0$$

$$\forall g \in \mathcal{C}(f^*), \quad \boxed{\sum_{a=1}^{K} \gamma_a \cdot \frac{\left(f^*(a) - g(a)\right)^2}{2} \geq 1}$$

"competing" hypotheses

$\gamma_a \ln(n)$ samples for each $a \in \mathcal{A}$ can **distinguish** f^* from g confidently.

Finds arm pull allocations that (1) eliminate competing hypotheses and (2) 'reward'-efficient

Agrawal, Teneketzis, Anantharam. Asymptotically Efficient Adaptive Allocation Schemes for Controlled I.I.D. Processes: Finite Parameter Space, 1989.

Example: Cheating code

- $\epsilon > 0$: very small (like 0.0001)
- $\Lambda > 0$: not too small (like 0.5)

- The lower bound: $\Theta\left(\frac{\log_2 K}{\Lambda^2} \ln n\right)$
- UCB: $\Theta\left(\frac{K}{\epsilon} \ln n\right)$
- Exponential gap in K!

base arms (K_0) — rewards: $\{1-\epsilon, 1, 1+\epsilon\}$

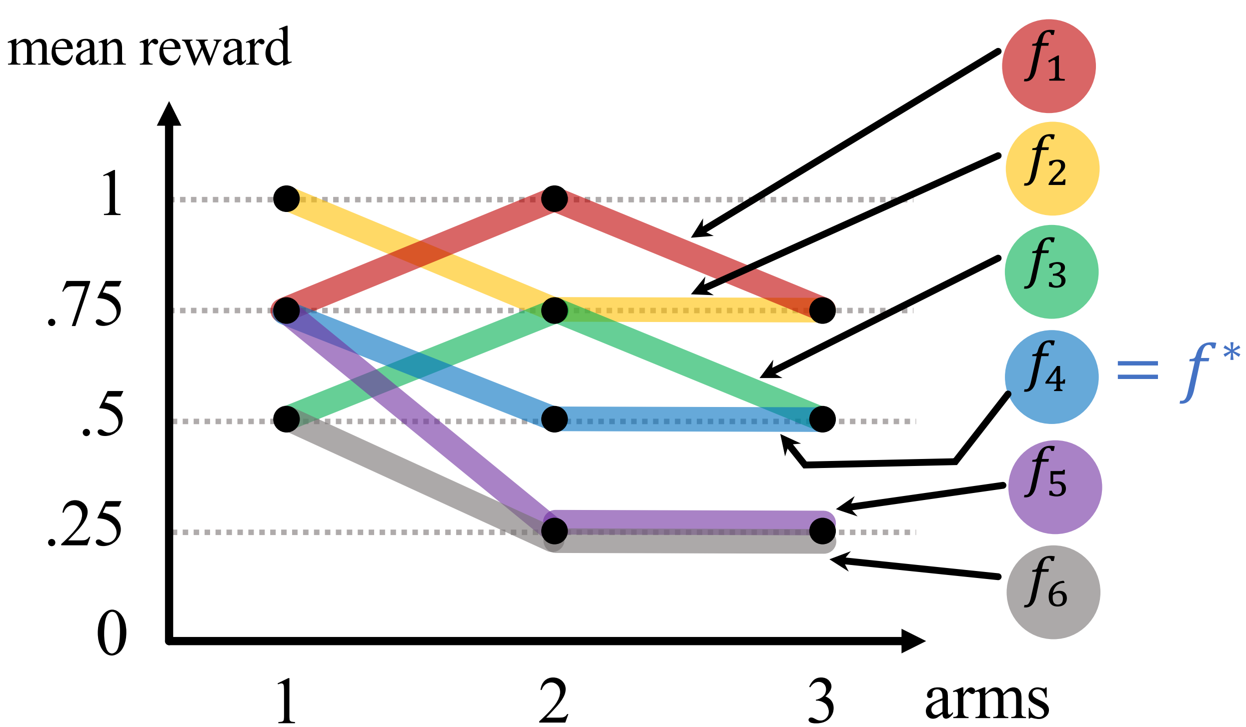
cheating arms ($\log_2 K_0$) — rewards: $\{0, \Lambda\}$

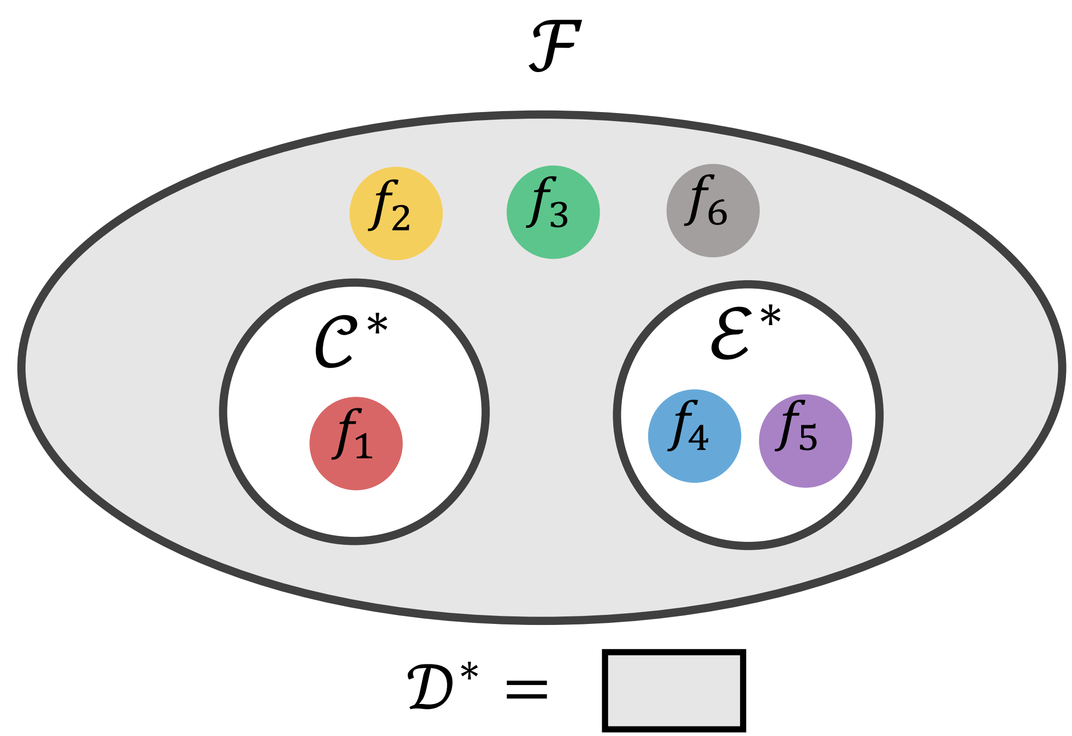
	A1	A2	A3	A4	A5	A6
f_1	$\mathbf{1}$	$1-\epsilon$	$1-\epsilon$	$1-\epsilon$	0	0
f_2	$1-\epsilon$	$\mathbf{1}$	$1-\epsilon$	$1-\epsilon$	0	Λ
f_3	$1-\epsilon$	$1-\epsilon$	$\mathbf{1}$	$1-\epsilon$	Λ	0
f_4	$1-\epsilon$	$1-\epsilon$	$1-\epsilon$	$\mathbf{1}$	Λ	Λ
f_5	$\mathbf{1+\epsilon}$	1	$1-\epsilon$	$1-\epsilon$	0	0
f_6	1	$\mathbf{1+\epsilon}$	$1-\epsilon$	$1-\epsilon$	0	Λ
f_7	$1-\epsilon$	$1-\epsilon$	$\mathbf{1+\epsilon}$	1	Λ	0
…	…	…	…	…	…	…

The function classes

regret contribution

- $\mathcal{C}(f^*)$: Competing $\Longrightarrow$ cannot distinguishable using $a^*(f^*)$, but supports a different arm — $\Theta(\log n)$
- $\mathcal{D}(f^*)$: Docile $\Longrightarrow$ distinguishable using $a^*(f^*)$ — $\Theta(1)$
- $\mathcal{E}(f^*)$: Equivalent $\Longrightarrow$ supports $a^*(f^*)$ and the reward $\mu^*(f^*)$ — can be $\Theta(\log\log n)$
- **[Proposition 2]** $\mathcal{F} = \mathcal{C}(f^*) \cup \mathcal{D}(f^*) \cup \mathcal{E}(f^*)$ (**disjoint** union)





CRush Optimism with Pessimism (CROP)

CROP: Overview

- The confidence set

confidence level: $1 - \text{poly}\left(\frac{1}{t}\right)$

$$L_t(f) := \sum_{s=1}^{t} \left(r_s - f(a_s)\right)^2$$

$$\mathcal{F}_t := \left\{ f \in \mathcal{F}\colon\ L_{t-1}(f) - \min_{g \in \mathcal{F}} L_{t-1}(g) \le \beta_t := \Theta(\,\ln(t|\mathcal{F}|)\,) \right\}$$

ERM

- Four important branches
  - `Exploit, Feasible, Fallback, Conflict`

- `Exploit`
  - Does every $f \in \mathcal{F}_t$ support the same best arm?
  - If yes, pull that arm.

CROP v1

At time t,

- Maintain a confidence set $\mathcal{F}_t \subseteq \mathcal{F}$
- If every $f \in \mathcal{F}_t$ agree on the best arm
  - `(Exploit)` pull that arm.
- Else: `(Feasible)`
  - Compute the pessimism: $\overline{f}_t = \arg\min_{f \in \mathcal{F}_t} \max_{a \in \mathcal{A}} f(a)$ (break ties by the cum. loss)
  - Compute $\gamma^* :=$ solution of the optimization problem $c(\overline{f}_t)$
  - (Tracking) Pull $a_t = \arg\min_{a \in \mathcal{A}} \frac{\text{pull\_count}(a)}{\gamma_a^*}$

Cf. optimism: $\tilde{f}_t = \arg\max_{f \in \mathcal{F}_t} \max_{a \in \mathcal{A}} f(a)$

Why pessimism?

Arms	A1	A2	A3	A4	A5
f_1	**1**	.99	.98	0	0
f_2	.98	**.99**	.98	<u>.25</u>	0
f_3	.97	.97	**.98**	.25	<u>.25</u>

- Suppose $\mathcal{F}_t = \{f_1, f_2, f_3\}$
- If I knew f^*, I could track $\gamma(f^*)$ (= the solution of $c(f^*)$)
- Which f should I track?

- **Pessimism**: either does the right thing, or eliminates itself.
- Other choices: may get stuck (so does ERM)

Key idea: the LB constraints prescribes how to distinguish f^* from those supporting **higher** rewards.

But we may still get stuck.

Arms	A1	A2	A3	A4	A5
f_1	**1**	.99	.98	0	0
f_2	.98	**.99**	.98	.25	0

$f^* =$ f_2

- Due to docile hypotheses.
- We must do something else.

$$\psi(f) := \arg\min_{\gamma\in[0,\infty)^K} \Delta_{\min}(f)\cdot\gamma_{a^*(f)} + \sum_{a\neq a^*(f)} \Delta_a(f)\cdot\gamma_a$$

$$\text{s.t.} \boxed{\forall\, g\in\mathcal{C}(f)\cup\mathcal{D}(f): \mu^*(g)\geq\mu^*(f),} \quad \sum_a \gamma_a \frac{\left(f(a)-g(a)\right)^2}{2}\geq 1$$

$$\gamma\geq\max\{\gamma(f),\phi(f)\}$$

- Includes **docile** hypotheses with best rewards higher $\mu^*(f)$

When to fallback to $\psi(f)$

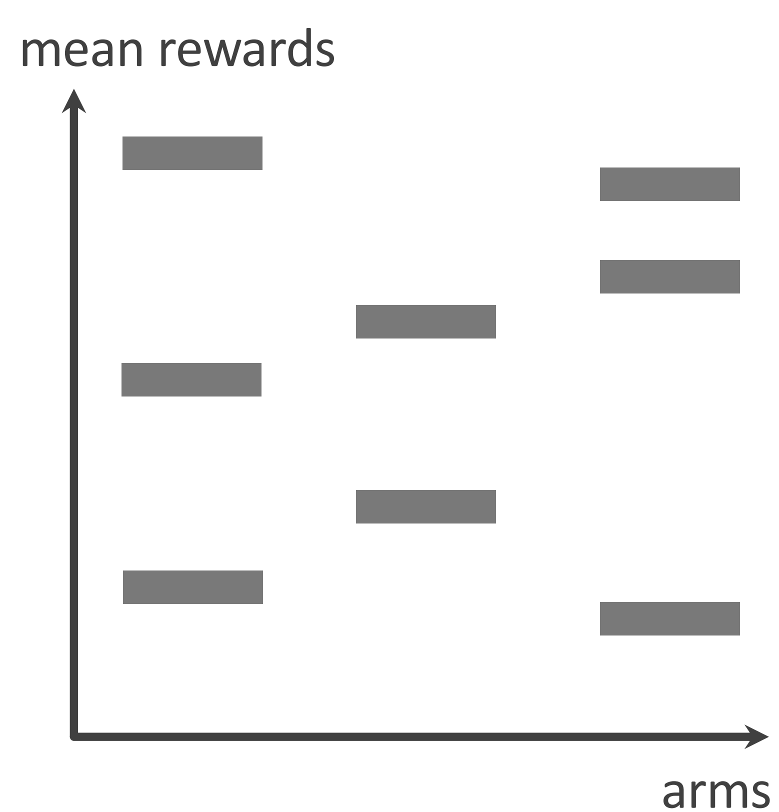
- $\mathcal{B}_t \coloneqq \{(a^*(f), \mu^*(f)) : f \in \mathcal{F}_t\} \implies$ induces a **partition** of $\mathcal{F}_t$
- **Optimistic set:** $\tilde{\mathcal{F}}_t$ = the partition containing the optimism
- **Pessimistic set**: $\bar{\mathcal{F}}_t$ = the partition containing the pessimism

- **Condition**: Use $\gamma(\bar{f}_t)$ if

$$\forall f \in \tilde{\mathcal{F}}_t, \qquad \sum_a \gamma_a(\bar{f}_t) \frac{\left(f(a) - \bar{f}_t(a)\right)^2}{2} \geq 1$$

  - otherwise, fallback to $\psi(\bar{f}_t)$.

- Then, we never get stuck
  - Crush optimism with pessimism (or end up crushing pessimism itself)



[a partition of $\mathcal{F}_t$]

optimistic set

pessimistic set

CROP v2

At time t,

- Maintain a confidence set $\mathcal{F}_t \subseteq \mathcal{F}$
- If every $f \in \mathcal{F}_t$ agree on the best arm
  - `(Exploit)` pull that arm.
- Else if $\gamma(\bar{f}_t)$ is sufficient to eliminate the optimistic set $\tilde{\mathcal{F}}_t$
  - `(Feasible)` $\pi_t = \gamma(\bar{f}_t)$
- Else
  - `(Fallback)` $\pi_t = \psi(\bar{f}_t)$
- (Tracking) Pull $a_t = \arg\min_{a \in \mathcal{A}} \frac{\text{pull\_count}(a)}{\pi_{t,a}}$

Still, we may not be asymptotical optimal

- Issue: Which informative arm to pull?

Arms	A1	A2	A3	A4	A5
f_1	**1**	.99	.98	0	0
f_2	.98	**.99**	.98	.25	0
f_3	.98	**.99**	.98	.25	.50

- If we follow f_2,
  - when $f^* = f_2$, it's fine.
  - when $f^* = f_3$, we have $(\text{suboptimal\_const}) \cdot \log(n)$ regret (and can be made arbitrarily suboptimal)

- **Intuition**: to guard against $\Theta(n)$ regret, we aim to be $\left(1 - \frac{1}{n}\right)$-confident.

  to guard against $\Theta(\log n)$ regret (w/ suboptimal const), we aim to be $\left(1 - \frac{1}{\square}\right)$-confident.

- **Solution**: construct a $\left(1 - \frac{1}{\log(n)}\right)$-confident set.

We build a refined confidence set

- $\dot{\mathcal{F}}_t = \{f \in \bar{\mathcal{F}}_t : L_{t-1}(f) - L_{t-1}(\bar{f}_t) \le \dot{\beta}_t = O(\log(|\mathcal{F}| \log t))\}$ confidence level: $1 - \mathrm{poly}\left(\frac{1}{\log(t)}\right)$
  - We have $\dot{\mathcal{F}}_t \subseteq \bar{\mathcal{F}}_t \subset \mathcal{F}_t$

- Ask: Compute $\gamma(f)$ for every $f \in \dot{\mathcal{F}}_t$. **Do they all agree, up to constant scaling?**
  - YES: CROP v2
  - NO: set $\pi_t = \phi(\bar{f}_t)$

$$\phi(f) = \arg\min_{\gamma_1,\ldots,\gamma_K \ge 0} \sum_{a=1}^{K} \gamma_a \cdot \Delta_a$$

$$\text{s.t.} \quad \gamma_{a^*(f)} = 0$$

$$\boxed{\forall g \in \mathcal{E}(f) : \gamma(g) \not\propto \gamma(f),} \quad \sum_{a=1}^{K} \gamma_a \cdot \frac{\left(f(a) - g(a)\right)^2}{2} \ge 1$$

distinguish those that give conflicting advice!

CROP v3 (final)

At time t,

- Maintain a confidence set $\mathcal{F}_t \subseteq \mathcal{F}$
- If every $f \in \mathcal{F}_t$ agree on the best arm
  - `(Exploit)` pull that arm.
- Else if $\exists f, g \in \dot{\mathcal{F}}_t$: $\gamma(f)$ and $\gamma(g)$ are not proportional to each other
  - `(Conflict)` $\pi_t = \phi(\bar{f}_t)$
- Else if $\gamma(\bar{f}_t)$ is sufficient to eliminate the optimistic set $\tilde{\mathcal{F}}_t$
  - `(Feasible)` $\pi_t = \gamma(\bar{f}_t)$
- Else
  - `(Fallback)` $\pi_t = \psi(\bar{f}_t)$
- (Tracking) Pull $a_t = \arg\min_{a \in \mathcal{A}} \frac{\text{pull\_count}(a)}{\pi_{t,a}}$

Main results

Main result

- Effective number of arms: $K_\psi =$ the number of arms with $\psi_a(f) \neq 0$ for some $f \in \mathcal{F}$
- **[Theorem 1] Anytime regret of CROP:**

$$\mathbb{E}\,\mathrm{Reg}_n = O\big(P_1 \ln n + P_2 \ln(\ln(n)) + P_3 \ln(|\mathcal{F}|) + K_\psi\big)$$

where

$$P_1 = \sum_a \Delta_a \cdot \gamma_a(f^*) \qquad \text{(from feasible)}$$

$$P_2 = \sum_a \Delta_a \cdot \max_{f \in \mathcal{E}(f^*)} \phi_a(f) \qquad \text{(from conflict)}$$

$$P_3 = \sum_a \Delta_a \cdot \max_{f \in \mathcal{F}} \psi_a(f) \qquad \text{(from fallback, mainly)}$$

- **[Corollary 1]** If $P_1 = 0$, then $P_2 = 0$. Thus, bounded regret.

Example: Cheating code

- $K_\psi \approx \log_2 K$
- CROP: $\frac{\ln(K)}{\Lambda^2} \ln n$
- Forced exploration: $\frac{\ln(K)}{\Lambda^2} \ln n + K$
- If $\Lambda = .5, K = 2^d$ and $n = K$
    - $\boldsymbol{d^2}$ vs $\mathbf{2^d}$
    - Exponential improvement!

$\overbrace{\text{A1, A2, A3, A3}}^{\text{base arms } (K_0)}$ $\overbrace{\text{A5, A6}}^{\text{cheating arms } (\log_2 K_0)}$

	A1	A2	A3	A3	A5	A6
f_1	1	$1-\epsilon$	$1-\epsilon$	$1-\epsilon$	0	0
f_2	$1-\epsilon$	1	$1-\epsilon$	$1-\epsilon$	0	Λ
f_3	$1-\epsilon$	$1-\epsilon$	1	$1-\epsilon$	Λ	0
f_4	$1-\epsilon$	$1-\epsilon$	$1-\epsilon$	1	Λ	Λ
f_5	1	$1+\epsilon$	$1-\epsilon$	$1-\epsilon$	0	0
f_6	1	$1-\epsilon$	$1+\epsilon$	$1-\epsilon$	0	0
f_7	1	$1-\epsilon$	$1-\epsilon$	$1+\epsilon$	0	0
…	…	…	…	…	…	…

Lower bound

- We pull some **uninformative** arm $\log(\log(n))$ times. Is it necessary?
- Existing lower bounds say: it can be anywhere between $\Theta(1)$ and $o(\log n)$.

- **Question**: Say an algorithm A is asymptotically optimal. Can it pull all **uninformative** arms $O(1)$ times?

- **[Theorem 2]**
  The answer is NO. There exists $\mathcal{F}'$ for which there exists an **uninformative** arm a with
  $$\mathbb{E}[\text{pull\_count}_n(a)] \geq c \cdot \ln\ln n$$

(conditions are more relaxed in the paper; will be updated in the arxiv in a few days)

The risk of naively mimicking the oracle

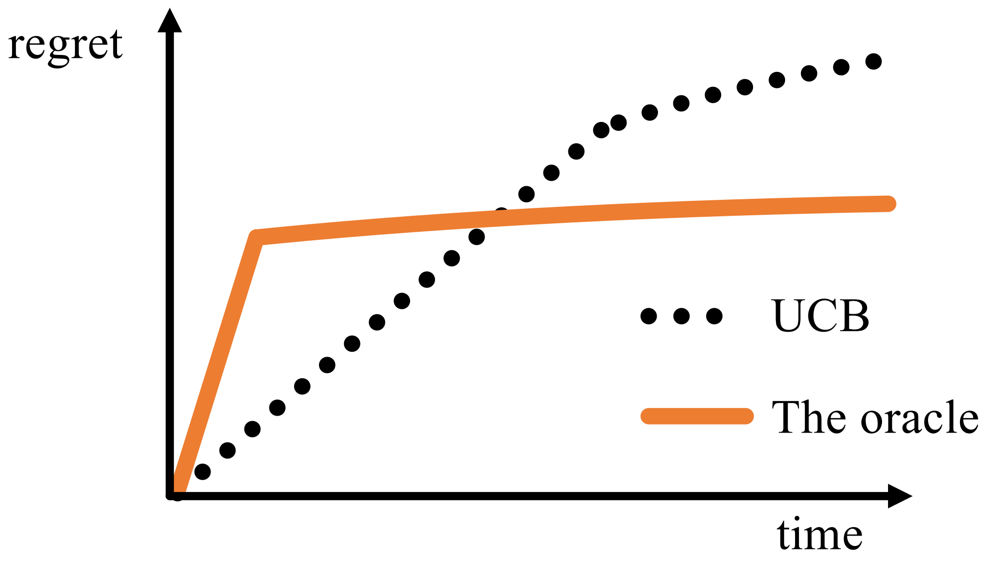
- **The oracle**: knows f^*
- At time t,
  - If $\forall a, \mathrm{pull\_count}_t(a) \geq \gamma_a(f^*) \cdot \ln t$
    - (Exploit) Pull $a^*(f^*)$
  - Else
    - (Explore) Track $\gamma(f^*)$

- Most existing algorithms try to mimic the oracle!
  - E.g., replace f^* with the ERM + forced exploration.

- CROP is not an exception

The risk of naively mimicking the oracle

- Regret of UCB: $O\left(\min\left\{\frac{K}{\epsilon}\ln(n), \epsilon n\right\}\right)$

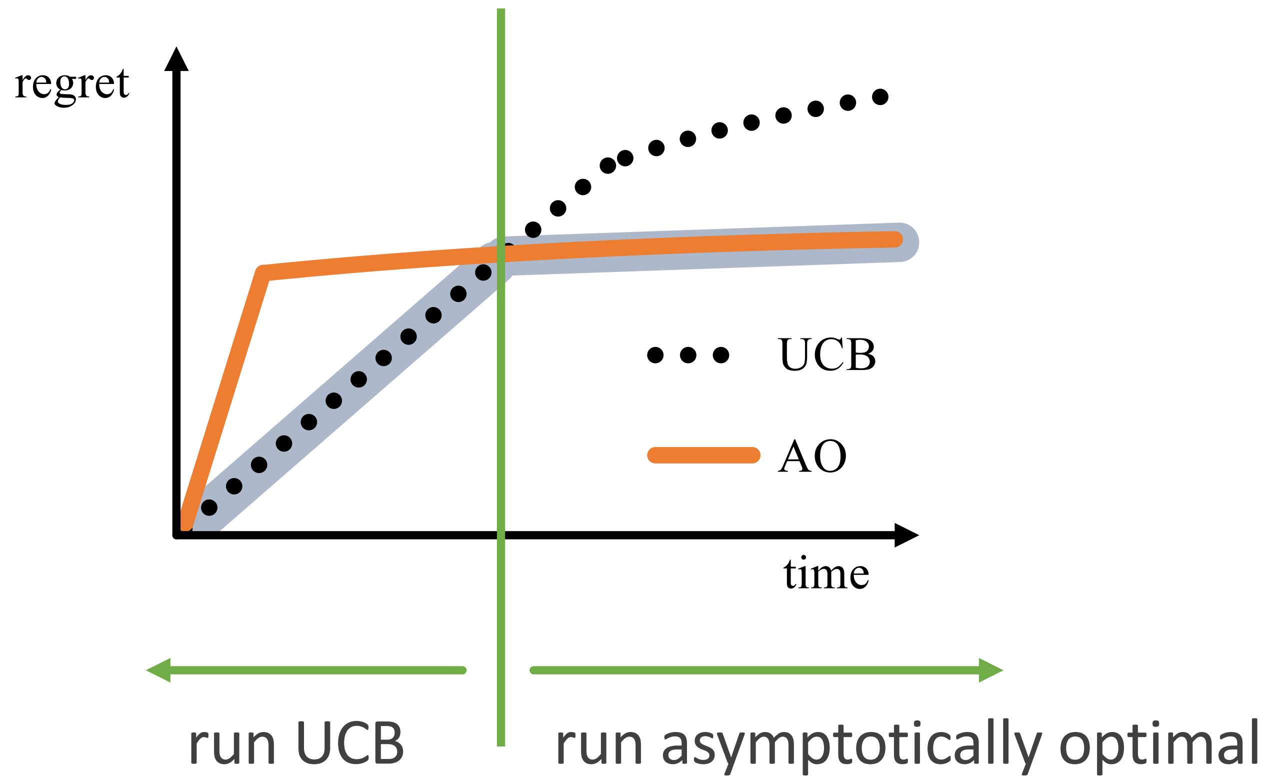


- The regret of the oracle: $O\left(\min\left\{\frac{\ln K}{\Lambda^2}\ln(n), n\right\}\right)$
  $\Longrightarrow$ Linear worst-case regret!
- Intuitively, if n is small, pulling ϵ-optimal arm is great!

	A1	A2	A3	A4	A5	A6
	base arms (K_0)				cheating arms ($\log_2 K_0$)	
f_1	$\mathbf{1}$	$1-\epsilon$	$1-\epsilon$	$1-\epsilon$	0	0
f_2	$1-\epsilon$	$\mathbf{1}$	$1-\epsilon$	$1-\epsilon$	0	Λ
f_3	$1-\epsilon$	$1-\epsilon$	$\mathbf{1}$	$1-\epsilon$	Λ	0
f_4	$1-\epsilon$	$1-\epsilon$	$1-\epsilon$	$\mathbf{1}$	Λ	Λ
f_5	$\mathbf{1}$	$1+\epsilon$	$1-\epsilon$	$1-\epsilon$	0	0
f_6	$\mathbf{1}$	$1-\epsilon$	$1+\epsilon$	$1-\epsilon$	0	0
f_7	$\mathbf{1}$	$1-\epsilon$	$1-\epsilon$	$1+\epsilon$	0	0
…	…	…	…	…	…	…

It may not be the end of optimism

- Can we achieve the best of both worlds? I.e., $O\left(\min\left\{\frac{\ln K}{\Lambda^2}\ln(n), \epsilon n\right\}\right)$
  - Yes, if we know ϵ



Summary

- CROP: Asymptotically optimal, adapt to bounded regret, with improved finite-time regret.
- Provides considerations when avoiding forced exploration.
- Reveals the danger of naively mimicking the oracle

- What next?
  - the worst-case regret simultaneously
  - can we use the pessimism for linear bandits?
  - can we even avoid solving the optimization problem?
  - Lower bounds for finite-time instance-dependent regret?
  - No explicit specification of confidence set construction/width?