Crush Optimism with Pessimism: Structured Bandits Beyond Asymptotic Optimality

Kwang-Sung Jun (presenter) and 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 $r_t = f^*(a_t) + \text{(zero-mean stochastic noise)}$
  - 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: stochastic bandits with **realizability**

Structured bandits

- **Why relevant?**
  Techniques developed here may extend to RL (e.g., ergodic RL [Ok+18])

- **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 $\mathcal{F}$.

Asymptotic optimality

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

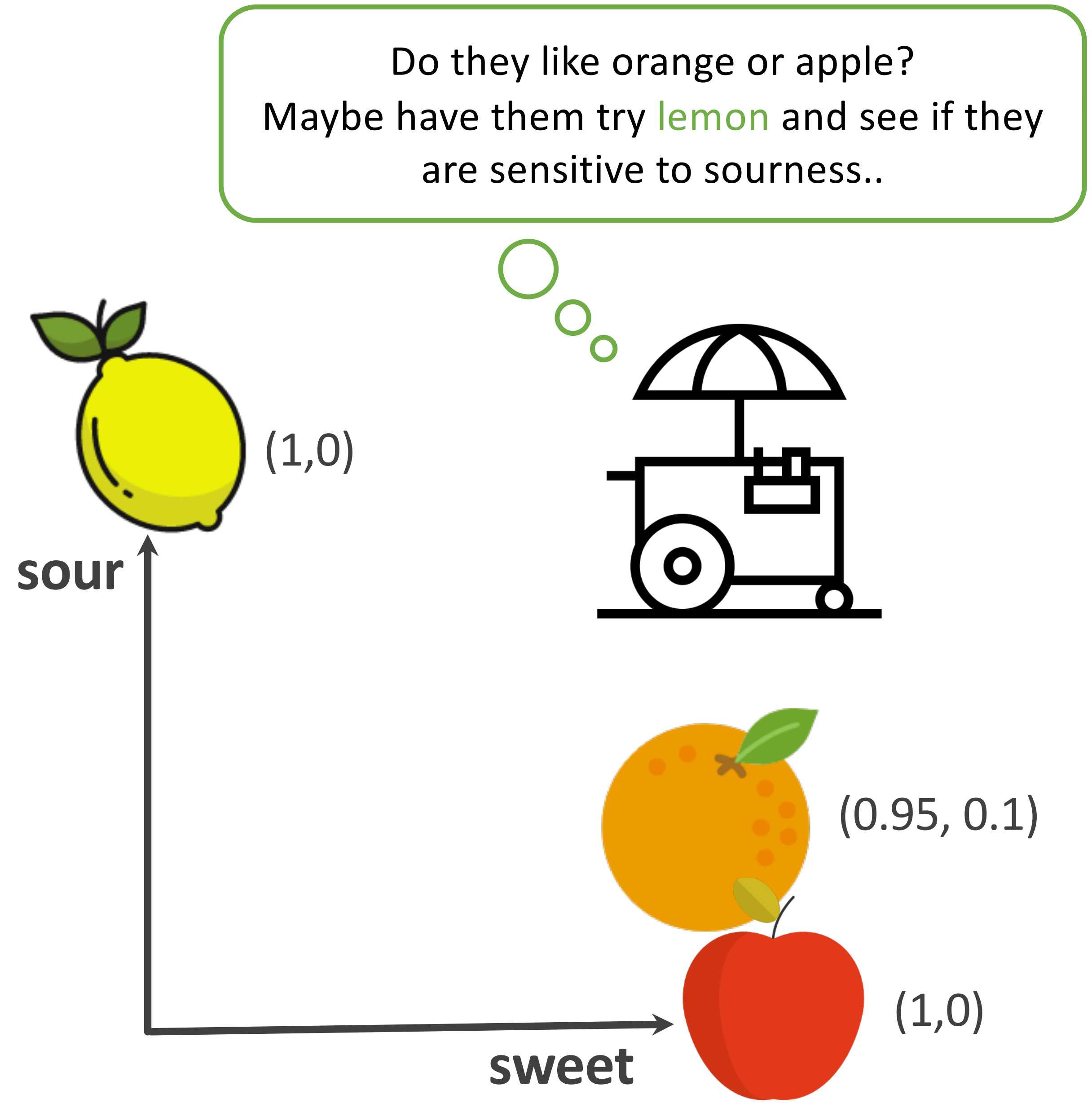
  $\Longrightarrow$ optimal asymptotic / worst-case regret in $\boldsymbol{K}$**-armed bandits**.
- Linear bandits: optimal worst-case rate = $d\sqrt{n}$
- Asymptotically optimal regret? $\Longrightarrow$ **No!**

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)



Asymptotic optimality: lower bound

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

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

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

$$\text{s.t.}\ \ \forall g \in \mathcal{C}(f), \quad \sum_{a=1}^{K} \gamma_a \cdot \mathrm{KL}_\nu\big(f(a), g(a)\big) \geq 1$$

”competing” hypotheses

KL divergence with noise distribution ν

- $\gamma^* = (\gamma_1^*, \ldots, \gamma_K^*) \geq 0$: the solution
- Suggests that we must pull arm a like $\gamma_a^* \cdot \log n$ times.
- What if $c(f^*) = 0$? **Bounded regret**! (except for pathological ones)

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 $K_\psi \leq K$ (defined in the paper).
- An interesting $\log \log n$ term in the regret*

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

CROP, just the core part.

At time t,

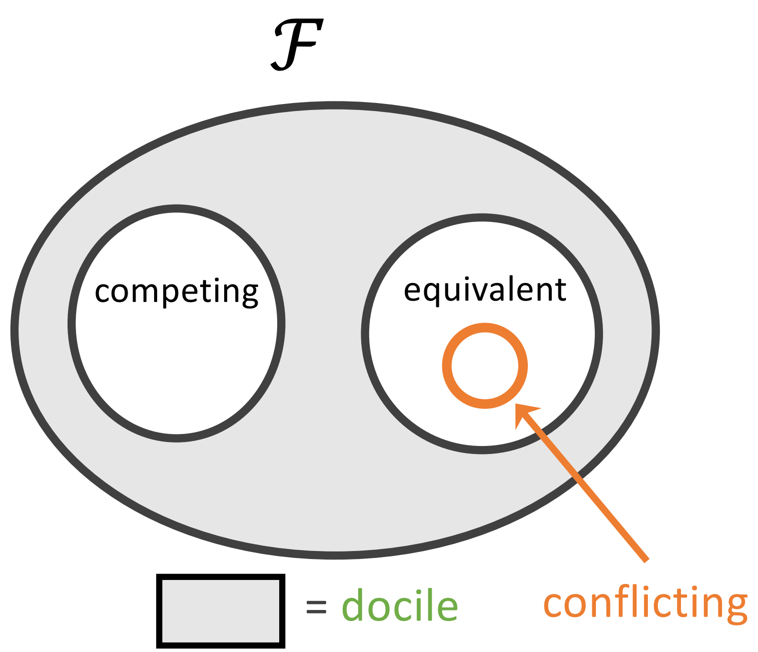
- Maintain a confidence set $\mathcal{F}_t \subseteq \mathcal{F}$
- Do all $f \in \mathcal{F}_t$ agree on the best arm?
  - YES: pull that arm.
  - NO:
    - Compute the pessimism: $\overline{f}_t = \arg\min_{f \in \mathcal{F}_t} \max_{a \in \mathcal{A}} f(a)$
    - Compute γ^* := solution of the optimization problem $c(\overline{f}_t)$
    - (Tracking) Pull $a_t = \arg\min_{a \in \mathcal{A}} \frac{\text{pull\_count}(a)}{\gamma_a^*}$

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

Most existing approaches:
Replace $\overline{f}_t$ with the empirical risk minimizer.
$\Longrightarrow$ requires forced sampling!

Preview of the paper, no details.

- Full version of CROP deals with
  - "docile" hypotheses. $\Longrightarrow$ bounded terms in the regret
  - "conflicting" hypotheses $\Longrightarrow$ $\log\log n$ terms



- The **risk** of naively pursuing the asymptotic optimality
  - The oracle who plays according to $\gamma^*(f^*)$ may suffer a **linear regret** in the worst-case sense (finite time).

- **It may not be the end of optimism**: we achieve **both** the worst-case and the asymptotic optimality by leveraging **optimism** (for a special $\mathcal{F}$ only).

Come to our poster after this session, find me in rocket chat, or email me/Chicheng for questions/discussions!