Optimal Multi-Objective Best Arm Identification with Fixed Confidence

Zhirui Chen, P. N. Karthik, Yeow Meng Chee, and **Vincent Y. F. Tan**

National University of Singapore, IIT Hyderabad

Information Theory and Application (ITA) Workshop 2025

Feb, 2025

Background: Multi-Objective Optimization

- Consider a $K = 3$ arm bandit problem.

Background: Multi-Objective Optimization

- Consider a $K = 3$ arm bandit problem.
- There are $M = 2$ users.

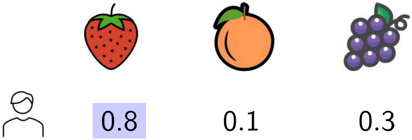
Background: Multi-Objective Optimization

- Consider a $K = 3$ arm bandit problem.
- There are $M = 2$ users.
- Each user has their own preference.

Background: Multi-Objective Optimization

- Consider a $K = 3$ arm bandit problem.
- There are $M = 2$ users.
- Each user has their own preference.

$$M = 2, \; K = 3$$



Background: Multi-Objective Optimization

- Consider a $K = 3$ arm bandit problem.
- There are $M = 2$ users.
- Each user has their own preference.

$M = 2,\ K = 3$

			
	0.8	0.1	0.3
	0.1	0.2	0.9

Background: Multi-Objective Optimization

- Consider a $K = 3$ arm bandit problem.
- There are $M = 2$ users.
- Each user has their own preference.

$M = 2, \ K = 3$

			
	$\boxed{0.8}$	0.1	0.3
	0.1	0.2	$\boxed{0.9}$

$\boxed{i_1^* = 1}, \ \boxed{i_2^* = 3}$

Background: Multi-Objective Optimization

- Consider a $K = 3$ arm bandit problem.
- There are $M = 2$ users.
- Each user has their own preference.

$M = 2,\ K = 3$

User 1	0.8	0.1	0.3
User 2	0.1	0.2	0.9

$i_1^* = 1$, $i_2^* = 3$

- Aim to find $i_1^*, \ldots, i_M^* \in [K]$ via bandit feedback.

Problem Statement

- Arm set: $[K] = \{1, \ldots, K\}$;

Problem Statement

- Arm set: $[K] = \{1, \ldots, K\}$;
- Objective set: $[M] = \{1, \ldots, M\}$;

Problem Statement

- Arm set: $[K] = \{1, \ldots, K\}$;
- Objective set: $[M] = \{1, \ldots, M\}$;
- Confidence level: $\delta \in (0, 1)$;

Problem Statement

- Arm set: $[K] = \{1, \ldots, K\}$;
- Objective set: $[M] = \{1, \ldots, M\}$;
- Confidence level: $\delta \in (0, 1)$;
- Mean reward of arm $i \in [K]$ under objective $m \in [M]$: $\mu_{i,m} \in \mathbb{R}$;

Problem Statement

- Arm set: $[K] = \{1, \ldots, K\}$;
- Objective set: $[M] = \{1, \ldots, M\}$;
- Confidence level: $\delta \in (0, 1)$;
- Mean reward of arm $i \in [K]$ under objective $m \in [M]$: $\mu_{i,m} \in \mathbb{R}$;

$M = 2, \; K = 3$

			
	0.8	0.1	0.3
	0.1	0.2	0.9

$\mu_{1,1} = 0.8$, $\quad \mu_{2,1} = 0.1$, $\quad \mu_{3,1} = 0.3$, $\quad i_1^* = 1$

$\mu_{2,1} = 0.1$, $\quad \mu_{2,2} = 0.2$, $\quad \mu_{3,2} = 0.9$, $\quad i_2^* = 3$

Problem Statement

- $I^* = (i_1^*, \cdots, i_M^*) \in [K]^M$ is the vector of best arms, where

$$i_m^* = \arg\max_{i \in [K]} \mu_{i,m}.$$

Problem Statement

- $I^* = (i_1^*, \cdots, i_M^*) \in [K]^M$ is the vector of best arms, where

$$i_m^* = \arg\max_{i \in [K]} \mu_{i,m}.$$

- For $t \in \mathbb{N}$, agent pulls arm $A_t \in [K]$ and obtains M rewards

$$X_{A_t,m}(t) \sim \mathcal{N}(\mu_{A_t,m}, 1) \qquad \forall\, m \in [M].$$

Problem Statement

- $I^* = (i_1^*, \cdots, i_M^*) \in [K]^M$ is the vector of best arms, where

$$i_m^* = \arg\max_{i \in [K]} \mu_{i,m}.$$

- For $t \in \mathbb{N}$, agent pulls arm $A_t \in [K]$ and obtains M rewards

$$X_{A_t,m}(t) \sim \mathcal{N}(\mu_{A_t,m}, 1) \qquad \forall\, m \in [M].$$

- Based on the history of arm pulls and rewards up to time t, agent can decide whether to stop at the time step t.

Problem Statement

- $I^* = (i_1^*, \cdots, i_M^*) \in [K]^M$ is the vector of best arms, where

$$i_m^* = \arg\max_{i \in [K]} \mu_{i,m}.$$

- For $t \in \mathbb{N}$, agent pulls arm $A_t \in [K]$ and obtains M rewards

$$X_{A_t,m}(t) \sim \mathcal{N}(\mu_{A_t,m}, 1) \qquad \forall\, m \in [M].$$

- Based on the history of arm pulls and rewards up to time t, agent can decide whether to stop at the time step t.
- Once the agent stops, it recommends the empirically best arm $\hat{i}_m$ for each objective $m \in [M]$.

Problem Statement

- $I^* = (i_1^*, \cdots, i_M^*) \in [K]^M$ is the vector of best arms, where

$$i_m^* = \arg\max_{i \in [K]} \mu_{i,m}.$$

- For $t \in \mathbb{N}$, agent pulls arm $A_t \in [K]$ and obtains M rewards

$$X_{A_t,m}(t) \sim \mathcal{N}(\mu_{A_t,m}, 1) \qquad \forall\, m \in [M].$$

- Based on the history of arm pulls and rewards up to time t, agent can decide whether to stop at the time step t.
- Once the agent stops, it recommends the empirically best arm $\hat{i}_m$ for each objective $m \in [M]$.
- Objective:

$$\min_{\pi} \ \mathbb{E}[\tau_\delta] \quad \text{s.t.} \quad \mathbb{P}(\widehat{I} \neq I^*) \leq \delta,$$

where $\widehat{I} = (\widehat{i}_1, \cdots, \widehat{i}_M)$ is the recommendation at the stopping time.

Lower Bound

- Policy: π

Lower Bound

- Policy: π
- Arm Pulling Strategy: $A_t \in \sigma\big(\{A_s, X_{A_s,1}, \ldots, X_{A_s,M}\}_{s=1}^{t-1}\big)$;

Lower Bound

- Policy: π
- Arm Pulling Strategy: $A_t \in \sigma\big(\{A_s, X_{A_s,1}, \ldots, X_{A_s,M}\}_{s=1}^{t-1}\big)$;
- Error Probability: $\delta \in (0, 1)$;

Lower Bound

- Policy: π
- Arm Pulling Strategy: $A_t \in \sigma\big(\{A_s, X_{A_s,1}, \ldots, X_{A_s,M}\}_{s=1}^{t-1}\big)$;
- Error Probability: $\delta \in (0,1)$;
- Stopping Time: τ_δ;

Lower Bound

- Policy: π
- Arm Pulling Strategy: $A_t \in \sigma\big(\{A_s, X_{A_s,1}, \ldots, X_{A_s,M}\}_{s=1}^{t-1}\big)$;
- Error Probability: $\delta \in (0, 1)$;
- Stopping Time: τ_δ;
- Final Recommendation: $\widehat{I}_\delta \in [K]^M$.

Lower Bound

- Policy: π
- Arm Pulling Strategy: $A_t \in \sigma\left(\{A_s, X_{A_s,1}, \ldots, X_{A_s,M}\}_{s=1}^{t-1}\right)$;
- Error Probability: $\delta \in (0,1)$;
- Stopping Time: τ_δ;
- Final Recommendation: $\widehat{I}_\delta \in [K]^M$.

Definition

A policy π is δ-PAC if it returns the vector of best arms w.p. $\geq 1-\delta$ in finite time, i.e., for all instances v,

$$\mathbb{P}_v^\pi(\tau_\delta < +\infty) = 1 \qquad \text{and} \qquad \mathbb{P}_v^\pi(\widehat{I}_\delta = I^\star(v)) \geq 1-\delta.$$

Lower Bound

- Policy: π
- Arm Pulling Strategy: $A_t \in \sigma\big(\{A_s, X_{A_s,1}, \ldots, X_{A_s,M}\}_{s=1}^{t-1}\big)$;
- Error Probability: $\delta \in (0,1)$;
- Stopping Time: τ_δ;
- Final Recommendation: $\widehat{I}_\delta \in [K]^M$.

Definition

A policy π is δ-PAC if it returns the vector of best arms w.p. $\geq 1-\delta$ in finite time, i.e., for all instances v,

$$\mathbb{P}_v^\pi(\tau_\delta < +\infty) = 1 \qquad \text{and} \qquad \mathbb{P}_v^\pi(\widehat{I}_\delta = I^\star(v)) \geq 1-\delta.$$

Definition

Given instance v, the gap of arm $i \in [K]$ under objective $m \in [M]$ is

$$\Delta_{i,m}(v) = \mu_{i_m^*,m} - \mu_{i,m}.$$

Lower Bound

Information-Theoretic Lower Bound

For any sequence of δ-PAC policies $\{\pi_\delta\}_{\delta\in(0,1)}$,

$$\liminf_{\delta\to 0^+} \frac{\mathbb{E}_v^{\pi}[\tau_\delta]}{\log(\frac{1}{\delta})} \geq c^*(v) \qquad \forall \text{ instances } v,$$

where $c^*(v)$ is given by

$$c^*(v)^{-1} := \sup_{\omega\in\Gamma} \min_{m\in[M]} \min_{i\in[K]\setminus i_m^*(v)} \frac{\omega_i\, \omega_{i_m^*(v)}\, \Delta_{i,m}^2(v)}{2(\omega_i + \omega_{i_m^*(v)})}. \tag{1}$$

Lower Bound

Information-Theoretic Lower Bound

For any sequence of δ-PAC policies $\{\pi_\delta\}_{\delta\in(0,1)}$,

$$\liminf_{\delta\to 0^+} \frac{\mathbb{E}_v^\pi[\tau_\delta]}{\log(\frac{1}{\delta})} \geq c^*(v) \qquad \forall \text{ instances } v,$$

where $c^*(v)$ is given by

$$c^*(v)^{-1} := \sup_{\omega\in\Gamma} \min_{m\in[M]} \min_{i\in[K]\setminus i_m^*(v)} \frac{\omega_i\, \omega_{i_m^*(v)}\; \Delta_{i,m}^2(v)}{2(\omega_i + \omega_{i_m^*(v)})}. \tag{1}$$

- Unknown gaps $\Delta_{i,m}(v)$.

Lower Bound

Information-Theoretic Lower Bound

For any sequence of δ-PAC policies $\{\pi_\delta\}_{\delta \in (0,1)}$,

$$\liminf_{\delta \to 0^+} \frac{\mathbb{E}_v^\pi[\tau_\delta]}{\log(\frac{1}{\delta})} \geq c^*(v) \qquad \forall \text{ instances } v,$$

where $c^*(v)$ is given by

$$c^*(v)^{-1} := \sup_{\omega \in \Gamma} \min_{m \in [M]} \min_{i \in [K] \setminus i_m^*(v)} \frac{\omega_i \, \omega_{i_m^*(v)} \, \Delta_{i,m}^2(v)}{2(\omega_i + \omega_{i_m^*(v)})}. \tag{1}$$

- Unknown gaps $\Delta_{i,m}(v)$.
- In (1), Γ denotes the set of probability distributions on $[K]$.

Lower Bound

Information-Theoretic Lower Bound

For any sequence of δ-PAC policies $\{\pi_\delta\}_{\delta\in(0,1)}$,

$$\liminf_{\delta\to 0^+} \frac{\mathbb{E}_v^{\pi}[\tau_\delta]}{\log(\frac{1}{\delta})} \geq c^*(v) \qquad \forall \text{ instances } v,$$

where $c^*(v)$ is given by

$$c^*(v)^{-1} := \sup_{\omega\in\Gamma} \min_{m\in[M]} \min_{i\in[K]\setminus i_m^*(v)} \frac{\omega_i\, \omega_{i_m^*(v)}\, \Delta_{i,m}^2(v)}{2(\omega_i + \omega_{i_m^*(v)})}. \tag{1}$$

- Unknown gaps $\Delta_{i,m}(v)$.
- In (1), Γ denotes the set of probability distributions on $[K]$.
- Let $\omega^* \in \Gamma$ attain the maximum of "sup" in (1).

Lower Bound

Information-Theoretic Lower Bound

For any sequence of δ-PAC policies $\{\pi_\delta\}_{\delta\in(0,1)}$,

$$\liminf_{\delta\to 0^+} \frac{\mathbb{E}_v^\pi[\tau_\delta]}{\log(\frac{1}{\delta})} \geq c^*(v) \qquad \forall \text{ instances } v,$$

where $c^*(v)$ is given by

$$c^*(v)^{-1} := \sup_{\boxed{\omega \in \Gamma}} \min_{m\in[M]} \min_{i\in[K]\setminus i_m^*(v)} \frac{\omega_i\, \omega_{i_m^*(v)}\ \boxed{\Delta_{i,m}^2(v)}}{2(\omega_i + \omega_{i_m^*(v)})}. \tag{1}$$

- Unknown gaps $\boxed{\Delta_{i,m}(v)}$.
- In (1), $\boxed{\Gamma}$ denotes the set of probability distributions on $[K]$.
- Let $\boxed{\omega^* \in \Gamma}$ attain the maximum of "sup" in (1).
- Then, $\boxed{\omega^*}$ represents the optimal proportion of arm pulls!

Methodology: Overview

- To derive an (asymptotically) optimal algorithm, calculate:

$$\omega^* = \arg\max_{\omega \in \Gamma} \min_{m \in [M]} \min_{i \in [K] \setminus i_m^*(v)} \frac{\omega_i \, \omega_{i_m^*(v)} \, \Delta_{i,m}^2(v)}{2(\omega_i + \omega_{i_m^*(v)})}$$

Then pull arms according to ω^*.

Methodology: Overview

- To derive an (asymptotically) optimal algorithm, calculate:

$$\omega^* = \arg\max_{\omega \in \Gamma} \min_{m \in [M]} \min_{i \in [K] \setminus i_m^*(v)} \frac{\omega_i \, \omega_{i_m^*(v)} \, \Delta_{i,m}^2(v)}{2(\omega_i + \omega_{i_m^*(v)})}$$

  Then pull arms according to ω^*.

- Difficulty: Difficult to obtain a closed-form solution for ω^*.

Methodology: Overview

- To derive an (asymptotically) optimal algorithm, calculate:

$$\omega^* = \arg\max_{\omega\in\Gamma} \min_{m\in[M]} \min_{i\in[K]\setminus i_m^*(v)} \frac{\omega_i\,\omega_{i_m^*(v)}\,\Delta_{i,m}^2(v)}{2(\omega_i+\omega_{i_m^*(v)})}$$

  Then pull arms according to ω^*.

- Difficulty: Difficult to obtain a closed-form solution for ω^*.

- Possible Solution: Iterative numerical method to compute ω^*.

Methodology: Overview

- To derive an (asymptotically) optimal algorithm, calculate:

$$\omega^* = \arg\max_{\omega \in \Gamma} \min_{m \in [M]} \min_{i \in [K] \setminus i_m^*(v)} \frac{\omega_i \, \omega_{i_m^*(v)} \, \Delta_{i,m}^2(v)}{2(\omega_i + \omega_{i_m^*(v)})}$$

  Then pull arms according to ω^*.

- Difficulty: Difficult to obtain a closed-form solution for ω^*.
- Possible Solution: Iterative numerical method to compute ω^*.
- Problem: May not be provably optimal if we run the method finitely many iterations.

Methodology: MO-BAI Policy

Recall that

$$c^*(v)^{-1} = \sup_{\omega \in \Gamma} \min_{m \in [M]} \min_{i \in [K] \setminus i_m^*(v)} \frac{\omega_i \, \omega_{i_m^*(v)} \, \Delta_{i,m}^2(v)}{2(\omega_i + \omega_{i_m^*(v)})}.$$

Methodology: MO-BAI Policy

Recall that

$$c^*(v)^{-1} = \sup_{\omega \in \Gamma} \min_{m \in [M]} \min_{i \in [K] \setminus i_m^*(v)} \underbrace{\frac{\omega_i \, \omega_{i_m^*(v)} \, \Delta_{i,m}^2(v)}{2(\omega_i + \omega_{i_m^*(v)})}}_{g_v^{(i,m)}(\omega)}.$$

- Define first-order approximation for each arm and objective $g_v^{(i,m)}(\omega)$:

$$g_v^{(i,m)}(\omega) + \langle \nabla_\omega g_v^{(i,m)}(\omega), \, z - \omega \rangle.$$

Methodology: MO-BAI Policy

Recall that

$$c^*(v)^{-1} = \sup_{\omega\in\Gamma} \underbrace{\min_{m\in[M]} \min_{i\in[K]\setminus i_m^*(v)} \underbrace{\frac{\omega_i\, \omega_{i_m^*(v)}\, \Delta_{i,m}^2(v)}{2(\omega_i + \omega_{i_m^*(v)})}}_{g_v^{(i,m)}(\omega)}}_{g_v(\omega)}.$$

- Define first-order approximation for each arm and objective $g_v^{(i,m)}(\omega)$:

$$g_v^{(i,m)}(\omega) + \langle \nabla_\omega g_v^{(i,m)}(\omega),\, \mathrm{z} - \omega\rangle.$$

- Define overall gradient-related function:

$$h_v(\omega, \mathrm{z}) := \min_{m\in[M]} \min_{i\in[K]\setminus i_m^*(v)} \left\{ g_v^{(i,m)}(\omega) + \langle \nabla_\omega g_v^{(i,m)}(\omega),\, \mathrm{z} - \omega\rangle \right\}.$$

Methodology: MO-BAI Policy

- Gradient-related function

$$h_v(\omega, z) := \min_{m\in[M]} \min_{i\in[K]\setminus i_m^*(v)} \left\{ g_v^{(i,m)}(\omega) + \langle \nabla_\omega g_v^{(i,m)}(\omega),\, z - \omega \rangle \right\}.$$

Methodology: MO-BAI Policy

- Gradient-related function

$$h_v(\omega, z) := \min_{m \in [M]} \min_{i \in [K] \setminus i_m^*(v)} \left\{ g_v^{(i,m)}(\omega) + \langle \nabla_\omega g_v^{(i,m)}(\omega),\, z - \omega \rangle \right\}.$$

- $h_v(\omega, z)$ is designed to approximate the overall objective $g_v(\omega)$.

Methodology: $\mathrm{MO\text{-}BAI}$ Policy

- Gradient-related function

$$h_v(\omega, z) := \min_{m \in [M]} \min_{i \in [K] \setminus i_m^*(v)} \left\{ g_v^{(i,m)}(\omega) + \langle \nabla_\omega g_v^{(i,m)}(\omega), z - \omega \rangle \right\}.$$

- $h_v(\omega, z)$ is designed to approximate the overall objective $g_v(\omega)$.
- But $h_v(\omega, z)$ is not a "linear approximation" of $g_v(\omega)$.

Methodology: MO-BAI Policy

- Gradient-related function

$$h_v(\omega, \mathrm{z}) := \min_{m\in[M]} \min_{i\in[K]\setminus i_m^*(v)} \left\{ \boxed{g_v^{(i,m)}(\omega) + \langle \nabla_\omega g_v^{(i,m)}(\omega),\, \mathrm{z} - \omega\rangle} \right\}.$$

- $h_v(\omega, z)$ is designed to approximate the overall objective $g_v(\omega)$.
- But $h_v(\omega, z)$ is not a "linear approximation" of $g_v(\omega)$.
- We take linear approximations of the inner terms

$$\boxed{g_v^{(i,m)}(\omega) + \langle \nabla_\omega g_v^{(i,m)}(\omega),\, \mathrm{z} - \omega\rangle.}$$

Methodology: $\mathrm{MO\text{-}BAI}$ Policy

- Gradient-related function

$$h_v(\omega, \mathrm{z}) := \min_{m \in [M]} \min_{i \in [K] \setminus i_m^*(v)} \left\{ g_v^{(i,m)}(\omega) + \langle \nabla_\omega g_v^{(i,m)}(\omega), \, \mathrm{z} - \omega \rangle \right\}.$$

- $h_v(\omega, z)$ is designed to approximate the overall objective $g_v(\omega)$.
- But $h_v(\omega, z)$ is not a "linear approximation" of $g_v(\omega)$.
- We take linear approximations of the inner terms

$$g_v^{(i,m)}(\omega) + \langle \nabla_\omega g_v^{(i,m)}(\omega), \, \mathrm{z} - \omega \rangle.$$

- Guide the agent to pull arms in the "direction of the gradient".

- Gradient-related function

$$h_v(\omega, \mathrm{z}) := \min_{m \in [M]} \min_{i \in [K] \setminus i_m^*(v)} \left\{ \boxed{g_v^{(i,m)}(\omega) + \langle \nabla_\omega g_v^{(i,m)}(\omega),\, \mathrm{z} - \omega \rangle} \right\}.$$

- $h_v(\omega, z)$ is designed to approximate the overall objective $g_v(\omega)$.
- But $h_v(\omega, z)$ is not a "linear approximation" of $g_v(\omega)$.
- We take linear approximations of the inner terms

$$\boxed{g_v^{(i,m)}(\omega) + \langle \nabla_\omega g_v^{(i,m)}(\omega),\, \mathrm{z} - \omega \rangle.}$$

- Guide the agent to pull arms in the "direction of the gradient".
- Adapting algorithm in Wang et al. (2021) to our setting

Methodology: MO-BAI Policy

- Gradient-related function

$$h_v(\omega, z) := \min_{m\in[M]} \min_{i\in[K]\setminus i_m^*(v)} \left\{ g_v^{(i,m)}(\omega) + \langle \nabla_\omega g_v^{(i,m)}(\omega),\, z - \omega \rangle \right\}.$$

- $h_v(\omega, z)$ is designed to approximate the overall objective $g_v(\omega)$.
- But $h_v(\omega, z)$ is not a "linear approximation" of $g_v(\omega)$.
- We take linear approximations of the inner terms

$$g_v^{(i,m)}(\omega) + \langle \nabla_\omega g_v^{(i,m)}(\omega),\, z - \omega \rangle.$$

- Guide the agent to pull arms in the "direction of the gradient".
- Adapting algorithm in Wang et al. (2021) to our setting
- Maintaining computational tractability and considering the K^M tuples of possible best arms

Methodology: MO-BAI Policy

Surrogate proportion at time step t:

$$\mathrm{s}_t := \arg\max_{\mathrm{s}\in\Gamma^{(\eta)}} h_{\widehat{v}_{I_t}}(\widehat{\omega}_{\cdot,t-1}, \mathrm{s}), \qquad \text{(a Linear Program)}$$

where

Methodology: MO-BAI Policy

Surrogate proportion at time step t:

$$\mathsf{s}_t := \arg\max_{\mathsf{s} \in \Gamma^{(\eta)}} h_{\widehat{\mathsf{v}}_{I_t}}(\widehat{\omega}_{\cdot,t-1}, \mathsf{s}) \qquad \text{(a Linear Program)}$$

where

- Average allocation up to time $t-1$

$$\widehat{\omega}_{\cdot,t-1} := \sum_{i=1}^{t-1} \frac{\mathsf{s}_i}{t-1}.$$

Methodology: MO-BAI Policy

Surrogate proportion at time step t:

$$\mathsf{s}_t := \arg\max_{\mathsf{s} \in \Gamma^{(\eta)}} h_{\widehat{v}_{l_t}}\left(\widehat{\omega}_{\cdot, t-1}, \mathsf{s}\right), \qquad \text{(a Linear Program)}$$

where

- Average allocation up to time $t-1$

$$\widehat{\omega}_{\cdot, t-1} := \sum_{i=1}^{t-1} \frac{\mathsf{s}_i}{t-1}.$$

- Empirical instances at time t is $\widehat{v}_t$

Methodology: MO-BAI Policy

Surrogate proportion at time step t:

$$\mathrm{s}_t := \arg\max_{\mathrm{s}\in\Gamma^{(\eta)}} h_{\widehat{v}_{l_t}}\left(\widehat{\omega}_{\cdot,t-1}, \mathrm{s}\right), \qquad \text{(a Linear Program)}$$

where

- Average allocation up to time $t-1$

$$\widehat{\omega}_{\cdot,t-1} := \sum_{i=1}^{t-1} \frac{\mathrm{s}_i}{t-1}.$$

- Empirical instances at time t is $\widehat{v}_t$
- $l_t := \max_{k\in\mathbb{N}:2^k\le t} 2^k$ is to prevent the instance $\widehat{v}_{l_t}$ from changing too frequently.

Methodology: $\mathrm{MO\text{-}BAI}$ Policy

Sampling Rule:

$$A_t \in \underset{i \in [K]}{\arg\max}\, [\mathsf{B}_{\cdot,t-1} + \mathsf{s}_t]_i,$$

where $\mathsf{B}_{\cdot,t}$ is the buffer defined as

$$\mathsf{B}_{\cdot,0} = \underline{0} \quad \text{and} \quad \mathsf{B}_{\cdot,t} = \mathsf{B}_{\cdot,t-1} - \mathsf{e}_{A_t} + \mathsf{s}_t.$$

Methodology: MO-BAI Policy

Sampling Rule:

$$A_t \in \arg\max_{i \in [K]} [\mathsf{B}_{\cdot,t-1} + \mathsf{s}_t]_i,$$

where $\mathsf{B}_{\cdot,t}$ is the buffer defined as

$$\mathsf{B}_{\cdot,0} = \underline{0} \quad \text{and} \quad \mathsf{B}_{\cdot,t} = \mathsf{B}_{\cdot,t-1} - \mathsf{e}_{A_t} + \mathsf{s}_t.$$

Example: $K = 2$.

Methodology: MO-BAI Policy

Sampling Rule:

$$A_t \in \arg\max_{i \in [K]} [\mathsf{B}_{\cdot,t-1} + \mathsf{s}_t]_i,$$

where $\mathsf{B}_{\cdot,t}$ is the buffer defined as

$$\mathsf{B}_{\cdot,0} = \underline{0} \quad \text{and} \quad \mathsf{B}_{\cdot,t} = \mathsf{B}_{\cdot,t-1} - \mathsf{e}_{A_t} + \mathsf{s}_t.$$

Example: $K = 2$. At time $t = 1$, suppose

$$\mathsf{s}_1 = \begin{bmatrix} 0.1 \\ 0.9 \end{bmatrix} \implies \text{pull arm 2} \implies \mathsf{B}_{\cdot,1} = \begin{bmatrix} 0.1 \\ -0.1 \end{bmatrix}$$

Methodology: MO-BAI Policy

Sampling Rule:

$$A_t \in \arg\max_{i\in[K]} [\mathsf{B}_{\cdot,t-1} + \mathsf{s}_t]_i,$$

where $\mathsf{B}_{\cdot,t}$ is the buffer defined as

$$\mathsf{B}_{\cdot,0} = \underline{0} \quad \text{and} \quad \mathsf{B}_{\cdot,t} = \mathsf{B}_{\cdot,t-1} - \mathsf{e}_{A_t} + \mathsf{s}_t.$$

<u>Example</u>: $K = 2$. At time $t = 1$, suppose

$$\mathsf{s}_1 = \begin{bmatrix} 0.1 \\ 0.9 \end{bmatrix} \quad \Longrightarrow \quad \text{pull arm 2} \quad \Longrightarrow \quad \mathsf{B}_{\cdot,1} = \begin{bmatrix} 0.1 \\ -0.1 \end{bmatrix}$$

At time $t = 2$, suppose

$$\mathsf{s}_2 = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \quad \mathsf{B}_{\cdot,1} + \mathsf{s}_2 = \begin{bmatrix} 0.6 \\ 0.4 \end{bmatrix} \quad \Longrightarrow \quad \text{pull arm 1} \quad \Longrightarrow \quad \mathsf{B}_{\cdot,2} = \begin{bmatrix} 0.4 \\ -0.4 \end{bmatrix}$$

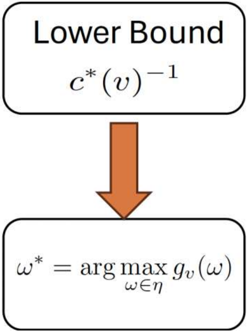
Sampling Rule Pipeline

Lower Bound

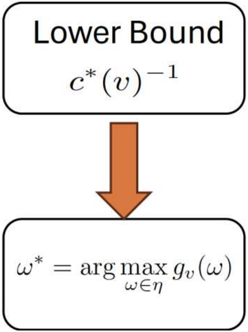
$c^*(v)^{-1}$

Methodology: MO-BAI Policy

Sampling Rule Pipeline



Sampling Rule Pipeline



$\approx$

Methodology: MO-BAI Policy

Sampling Rule Pipeline

Lower Bound

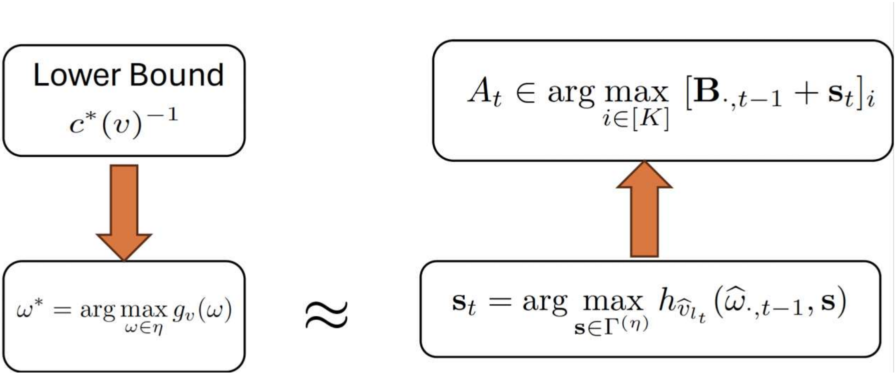
$$c^*(v)^{-1}$$

$$\omega^* = \arg\max_{\omega \in \eta} g_v(\omega)$$

$$\approx$$

$$\mathbf{s}_t = \arg\max_{\mathbf{s} \in \Gamma^{(\eta)}} h_{\widehat{v}_{l_t}}(\widehat{\omega}_{\cdot,t-1}, \mathbf{s})$$

Sampling Rule Pipeline



Stopping Rule:

Stopping Rule:

- Chernoff's stopping rule (Kaufmann et al., 2016) inspired by Chen et al. (2023).

Stopping Rule:

- Chernoff's stopping rule (Kaufmann et al., 2016) inspired by Chen et al. (2023).
- Let

$$Z(t) := \min_{m\in[M]} \min_{i\in[K]\setminus\widehat{i}_m(t)} \underbrace{\frac{N_{i,t}\, N_{\widehat{i}_m(t),t}\, \widehat{\Delta}^2_{i,m}(t)}{2(N_{i,t}+N_{\widehat{i}_m(t),t})}}_{\text{approx of } g_v^{(i,m)}(\omega)}$$

Stopping Rule:

- Chernoff's stopping rule (Kaufmann et al., 2016) inspired by Chen et al. (2023).
- Let

$$Z(t) := \min_{m\in[M]} \min_{i\in[K]\setminus\widehat{i}_m(t)} \underbrace{\frac{N_{i,t}\, N_{\widehat{i}_m(t),t}\, \widehat{\Delta}^2_{i,m}(t)}{2(N_{i,t}+N_{\widehat{i}_m(t),t})}}_{\text{approx of } g_v^{(i,m)}(\omega)}$$

- The stopping time of MO-BAI is

$$\tau_\delta = \min\{t \geq K : Z(t) > \beta(t,\delta)\},$$

where $\beta(t,\delta)$ is a carefully chosen threshold.

Theoretical Results

Proposition: δ-PACness

Fix $\delta \in (0,1)$. Then, $\mathrm{MO\text{-}BAI}$ is δ-PAC, i.e., for all instances v,

$$\mathbb{P}_v^{\mathrm{MO\text{-}BAI}}(\tau_\delta < +\infty) = 1 \quad \text{and}$$
$$\mathbb{P}_v^{\mathrm{MO\text{-}BAI}}\big(\widehat{I_\delta} = I^*(v)\big) \geq 1 - \delta.$$

Theoretical Results

Proposition: δ-PACness

Fix $\delta \in (0,1)$. Then, $\mathrm{MO\text{-}BAI}$ is δ-PAC, i.e., for all instances v,

$$\mathbb{P}_v^{\mathrm{MO\text{-}BAI}}(\tau_\delta < +\infty) = 1 \quad \text{and}$$
$$\mathbb{P}_v^{\mathrm{MO\text{-}BAI}}\big(\widehat{I_\delta} = I^*(v)\big) \geq 1 - \delta.$$

Theorem: Asymptotic Optimality

Under $\mathrm{MO\text{-}BAI}$, for all instances v,

$$\limsup_{\delta \to 0^+} \frac{\mathbb{E}_v^{\mathrm{MO\text{-}BAI}}[\tau_\delta]}{\log(\frac{1}{\delta})} \leq c^*(v) \quad \text{and}$$

$$\mathbb{P}_v^{\mathrm{MO\text{-}BAI}}\left(\limsup_{\delta \to 0^+} \frac{\tau_\delta}{\log(\frac{1}{\delta})} \leq c^*(v)\right) = 1.$$

Numerical Study on Synthetic Dataset

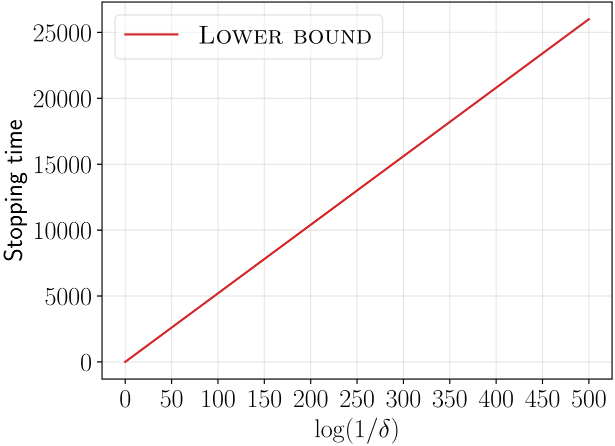


Figure 1: Average τ_δ of $\mathrm{MO\text{-}BAI}$ and Multi-Objective adaptation of D-Tracking

Numerical Study on Synthetic Dataset

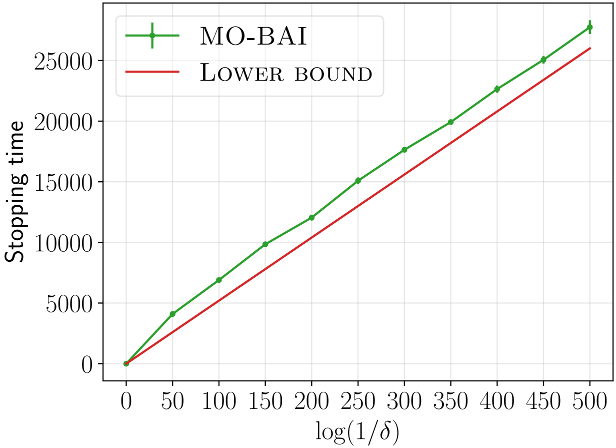


Figure 1: Average τ_δ of $\mathrm{MO\text{-}BAI}$ and Multi-Objective adaptation of D-Tracking

Numerical Study on Synthetic Dataset

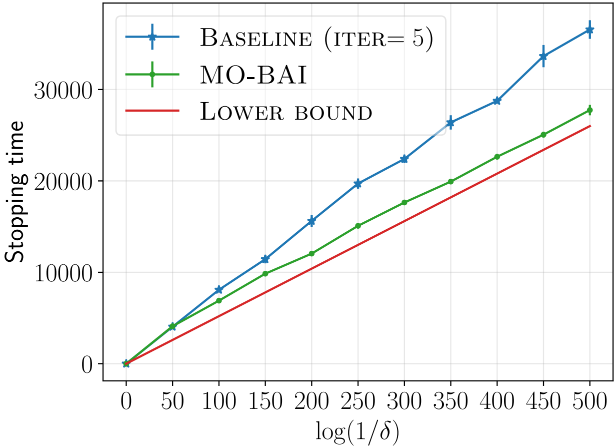


Figure 1: Average τ_δ of MO-BAI and Multi-Objective adaptation of D-Tracking

Numerical Study on Synthetic Dataset

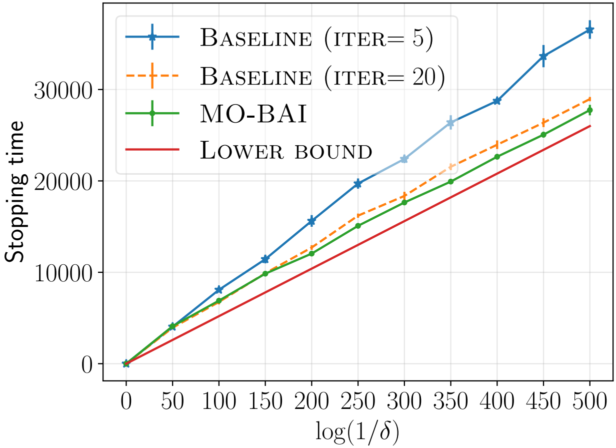


Figure 1: Average τ_δ of MO-BAI and Multi-Objective adaptation of D-Tracking

Numerical Study on the SNW Dataset

	$\delta = 0.1$	$\delta = 0.05$
MO-BAI	968.82 ± 58.21	$1,023.77 \pm 67.42$
Baseline	$4,485.98 \pm 124.92$	$6,168.29 \pm 132.01$
Baseline-Non-Unif	$3,841.05 \pm 136.44$	$4,320.55 \pm 128.26$
MO-SE	$2,322.39 \pm 461.54$	$2,411.16 \pm 421.88$

Table 1: Average stopping times obtained by running 100 independent trials with $\delta \in \{0.1, 0.05\}$ for the SNW dataset. In Baseline and Baseline-Non-Unif, we set Iter $= 20$.

Conclusion

- Multi-Objective Best Arm Identification problem with fixed-confidence

Conclusion

- Multi-Objective Best Arm Identification problem with fixed-confidence

$M = 2,\ K = 3$

$\boxed{i_1^* = 1}$, $\boxed{i_2^* = 3}$

Conclusion

- Multi-Objective Best Arm Identification problem with fixed-confidence

$M = 2,\ K = 3$

	$\boxed{0.8}$	0.1	0.3
	0.1	0.2	$\boxed{0.9}$

$\boxed{i_1^* = 1}$, $\boxed{i_2^* = 3}$

- Pulling arm A_t yields a vector of rewards

$$X_{A_t,m}(t) \sim \mathcal{N}(\mu_{A_t,m}, 1) \qquad \forall\, m \in [M].$$

Conclusion

- Multi-Objective Best Arm Identification problem with fixed-confidence

$M = 2,\ K = 3$

	0.8	0.1	0.3
	0.1	0.2	0.9

$i_1^* = 1$, $i_2^* = 3$

- Pulling arm A_t yields a vector of rewards

$$X_{A_t,m}(t) \sim \mathcal{N}(\mu_{A_t,m}, 1) \qquad \forall\, m \in [M].$$

- Derived an asymptotically optimal and efficient algorithm

$$\boxed{c^*(v)} \le \liminf_{\delta \to 0^+} \frac{\mathbb{E}_v^{\pi}[\tau_\delta]}{\log(\frac{1}{\delta})} \le \limsup_{\delta \to 0^+} \frac{\mathbb{E}_v^{\text{MO-BAI}}[\tau_\delta]}{\log(\frac{1}{\delta})} \le \boxed{c^*(v)}.$$