Best Arm Identification with Fixed Confidence: Multi-Objectives and Applications in Wireless Communications

Vincent Y. F. Tan

National University of Singapore

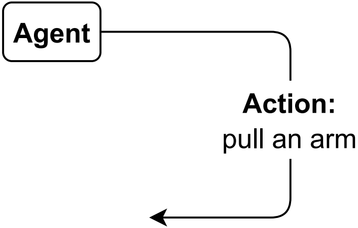
NCC 2025

2025 NATIONAL CONFERENCE ON COMMUNICATIONS

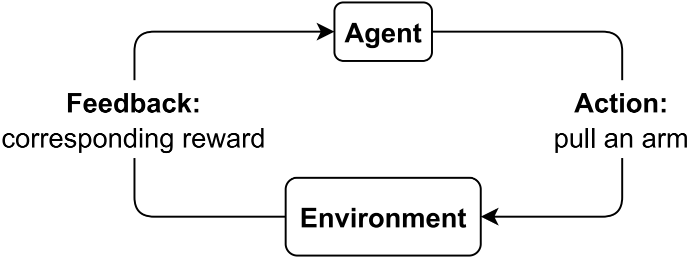
Mar 2025

Multi-Armed Bandits (MAB)

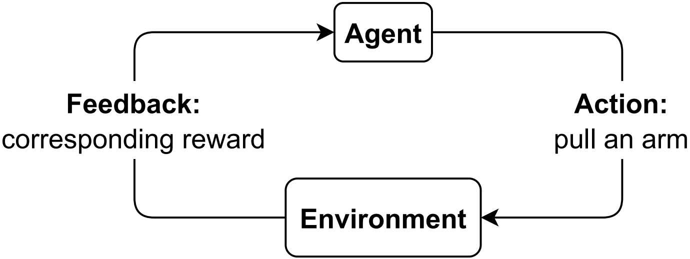
Multi-Armed Bandits (MAB)



Multi-Armed Bandits (MAB)



Multi-Armed Bandits (MAB)



Objectives

1. Maximize the cumulative reward over a fixed horizon $\Rightarrow$ Exploration-Exploitation tradeoff.
2. Our focus: Find the best arm or arms (largest expected reward(s))

Multi-Armed Bandits with Multiple Objectives

Zhirui Chen (NUS)

P. N. Karthik (IIT Hyderabad)

Yeow Meng Chee (NUS)

Multi-Armed Bandits with Multiple Objectives

Zhirui Chen (NUS)

P. N. Karthik (IIT Hyderabad)

Yeow Meng Chee (NUS)

Optimal Multi-Objective Best Arm Identification with Fixed Confidence

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

To appear in AISTATS, 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.

$M = 2,\ K = 3$

	0.8	0.1	0.3

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.

$M = 2, \ K = 3$

			
	0.8	0.1	0.3
	0.1	0.2	0.9

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

Background: Multi-Objective Optimization

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

$M = 2,\ K = 3$

	0.8	0.1	0.3
	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

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

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

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.
- After stopping, agent recommends the empirically best arms $\hat{i}_m$.

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

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

- Based on the history of arm pulls and rewards up to time t, agent can decide whether to stop at the time step t.
- After stopping, agent recommends the empirically best arms $\hat{i}_m$.
- Objective:

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

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

Lower Bound

- Policy and Error Probability: $\pi = \{\pi_t\}_{t=1}^{\infty}$ and δ

Lower Bound

- Policy and Error Probability: $\pi = \{\pi_t\}_{t=1}^{\infty}$ and δ
- Arm Pulling Strategy at time t: $A_t \in \sigma\big(\{A_s, X_{A_s,1}, \ldots, X_{A_s,M}\}_{s=1}^{t-1}\big)$;

Lower Bound

- Policy and Error Probability: $\pi = \{\pi_t\}_{t=1}^{\infty}$ and δ
- Arm Pulling Strategy at time t: $A_t \in \sigma\big(\{A_s, X_{A_s,1}, \ldots, X_{A_s,M}\}_{s=1}^{t-1}\big)$;
- Stopping Time: τ_δ;

Lower Bound

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

Lower Bound

- Policy and Error Probability: $\pi = \{\pi_t\}_{t=1}^{\infty}$ and δ
- Arm Pulling Strategy at time t: $A_t \in \sigma\big(\{A_s, X_{A_s,1}, \ldots, X_{A_s,M}\}_{s=1}^{t-1}\big)$;
- 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 and Error Probability: $\pi = \{\pi_t\}_{t=1}^{\infty}$ and δ
- Arm Pulling Strategy at time t: $A_t \in \sigma\big(\{A_s, X_{A_s,1}, \ldots, X_{A_s,M}\}_{s=1}^{t-1}\big)$;
- 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

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

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

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

- 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

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

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

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

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

Methodology: A Possible Solution

Calculate

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

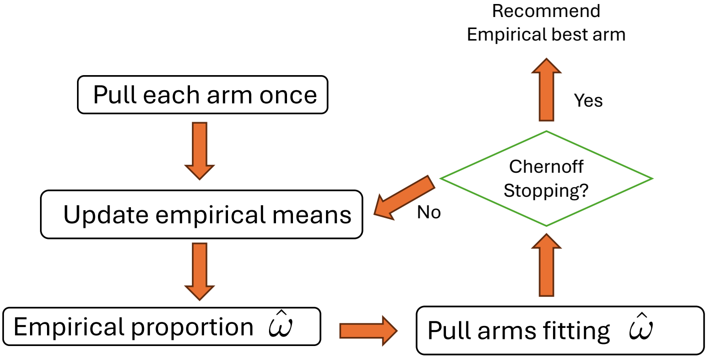
Then, a natural algorithm is to pull arms by empirical values of $\boldsymbol{\omega}^*$.

Methodology: A Possible Solution

Calculate

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

Then, a natural algorithm is to pull arms by empirical values of $\boldsymbol{\omega}^*$.



Methodology: Difficulties

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

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

Then pull arms according to the proportions in the probability vector $\boldsymbol{\omega}^*$.

Methodology: Difficulties

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

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

  Then pull arms according to the proportions in the probability vector $\boldsymbol{\omega}^*$.

- Difficulty: Difficult to obtain a closed-form solution for $\boldsymbol{\omega}^*$.

Methodology: Difficulties

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

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

  Then pull arms according to the proportions in the probability vector $\boldsymbol{\omega}^*$.

- Difficulty: Difficult to obtain a closed-form solution for $\boldsymbol{\omega}^*$.

- Possible Solution: Iterative numerical method to compute $\boldsymbol{\omega}^*$.

Methodology: Difficulties

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

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

  Then pull arms according to the proportions in the probability vector $\boldsymbol{\omega}^*$.

- Difficulty: Difficult to obtain a closed-form solution for $\boldsymbol{\omega}^*$.

- Possible Solution: Iterative numerical method to compute $\boldsymbol{\omega}^*$.

- 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

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

- 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

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

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

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

- Define overall gradient-related function:

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

Methodology: MO-BAI Policy

- Gradient-related function

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

Methodology: MO-BAI Policy

- Gradient-related function

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

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

Methodology: MO-BAI Policy

- Gradient-related function

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

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

Methodology: MO-BAI Policy

- Gradient-related function

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

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

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

Methodology: MO-BAI Policy

- Gradient-related function

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

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

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

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

Methodology: MO-BAI Policy

- Gradient-related function

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

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

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

- 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

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

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

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

- 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:

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

where

Methodology: MO-BAI Policy

Surrogate proportion at time step t:

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

where

- Average allocation up to time $t-1$

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

Methodology: MO-BAI Policy

Surrogate proportion at time step t:

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

where

- Average allocation up to time $t-1$

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

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

Methodology: MO-BAI Policy

Surrogate proportion at time step t:

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

where

- Average allocation up to time $t-1$

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

- 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: MO-BAI Policy

Sampling Rule:

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

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

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

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

Sampling Rule:

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

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

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

Example: $K = 2$.

Methodology: MO-BAI Policy

Sampling Rule:

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

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

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

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

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

Methodology: MO-BAI Policy

Sampling Rule:

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

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

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

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

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

At time $t = 2$, suppose

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

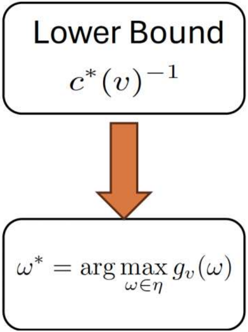
Methodology: MO-BAI Policy

Sampling Rule Pipeline

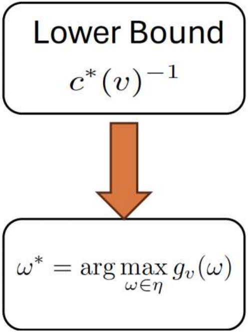
Lower Bound

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

Sampling Rule Pipeline



Sampling Rule Pipeline



$\approx$

Methodology: MO-BAI Policy

Sampling Rule Pipeline

Lower Bound

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

$$\Downarrow$$

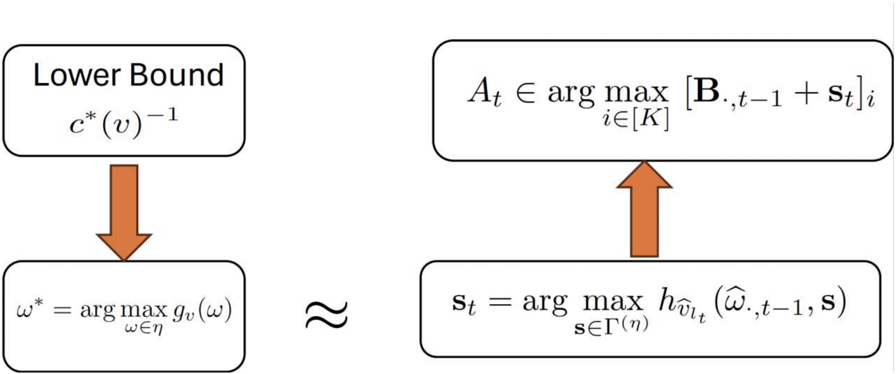
$$\omega^* = \arg\max_{\omega \in \eta} g_v(\omega)$$

$$\approx$$

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

Methodology: MO-BAI Policy

Sampling Rule Pipeline



Stopping Rule:

Stopping Rule:

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

Stopping Rule:

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

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

Stopping Rule:

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

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

- 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 tuned 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,

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

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,

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

Numerical Study on Synthetic Dataset

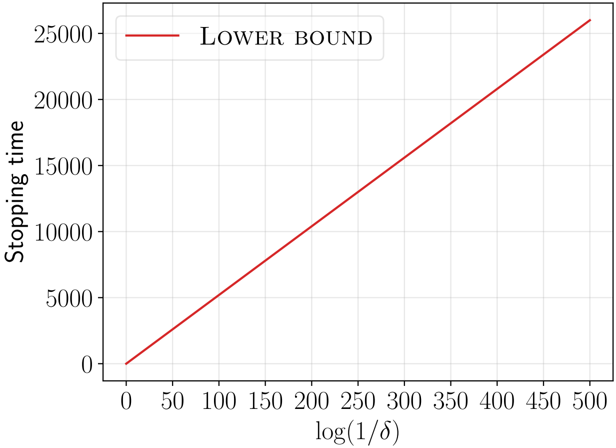


Figure 1: Average τ_δ of $\mathrm{MO\text{-}BAI}$ and Multi-Objective adaptation of D-Tracking

Numerical Study on Synthetic Dataset

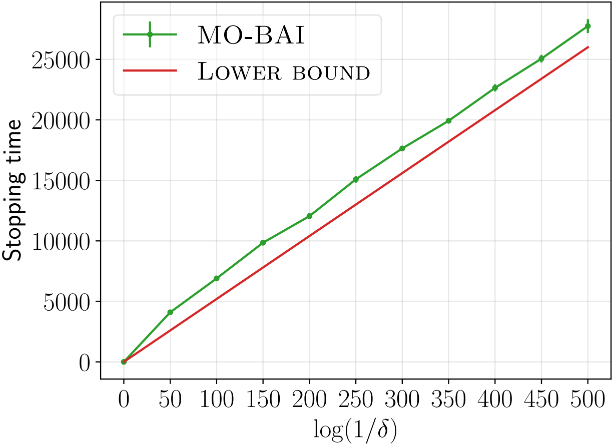


Figure 1: Average τ_δ of MO-BAI and Multi-Objective adaptation of D-Tracking

Numerical Study on Synthetic Dataset

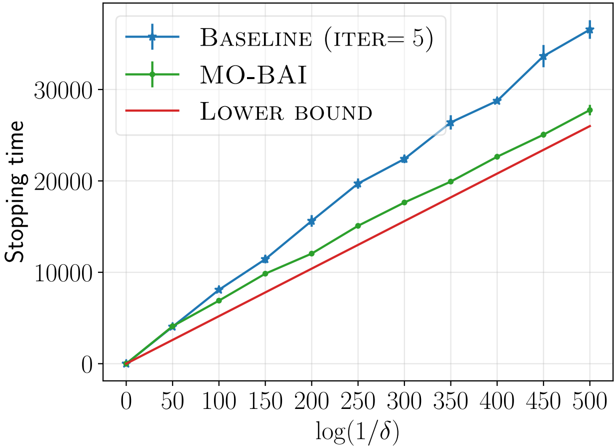


Figure 1: Average τ_δ of MO-BAI and Multi-Objective adaptation of D-Tracking

Numerical Study on Synthetic Dataset

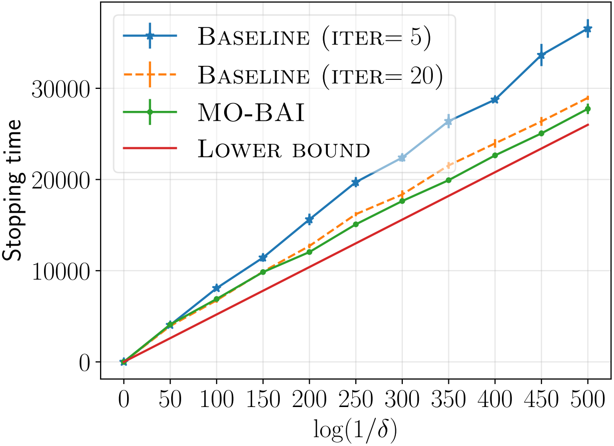


Figure 1: Average τ_δ of $\mathrm{MO\text{-}BAI}$ and Multi-Objective adaptation of D-Tracking

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 for MO-BAI

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

Conclusion for MO-BAI

- 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

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

Conclusion for MO-BAI

- 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

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

Conclusion for MO-BAI

- 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

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

How can we apply the theory to real-world wireless communication systems?

From Theory to Applications

How can we apply the theory to real-world wireless communication systems?

3264 IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, VOL. 22, NO. 5, MAY 2023

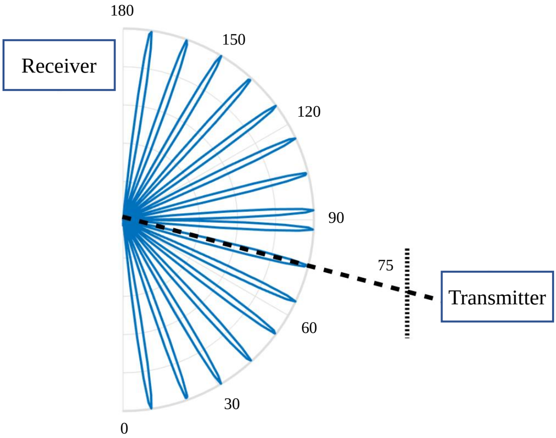
Fast Beam Alignment via Pure Exploration in Multi-Armed Bandits

Yi Wei, Zixin Zhong, and Vincent Y. F. Tan, *Senior Member, IEEE*

Zhejiang University

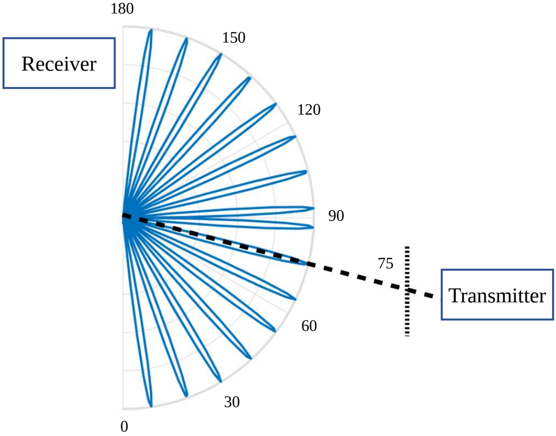
HKUST (Guangzhou)

Beam Alignment



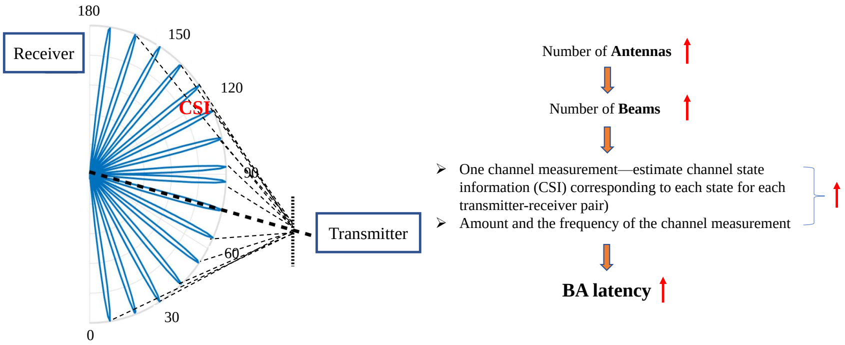
- Beams at Tx and Rx are narrow directional.

Beam Alignment



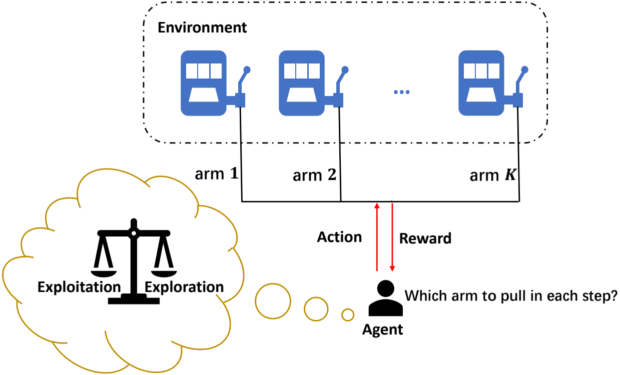
- Beams at Tx and Rx are narrow directional.
- Beam Alignment ensures Tx and Rx beams are accurately aligned to establish a reliable communication link.

Fundamental challenges

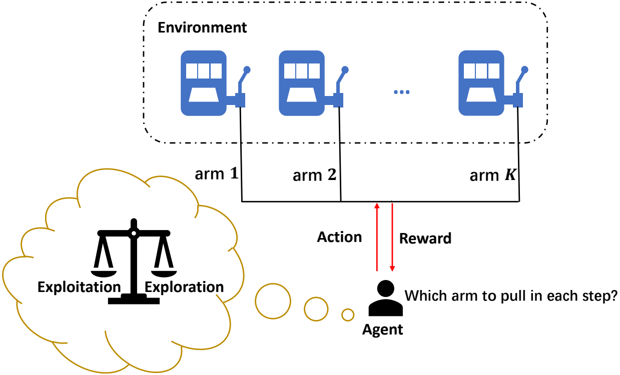


- Channel state information for each Tx-Rx pair is measured.
- Frequency of measurement is high due to mobility.
- Results in beam alignment latency which increases with the number of antennas at the Rx and Tx.

Beam Alignment as Multi-Armed Bandits

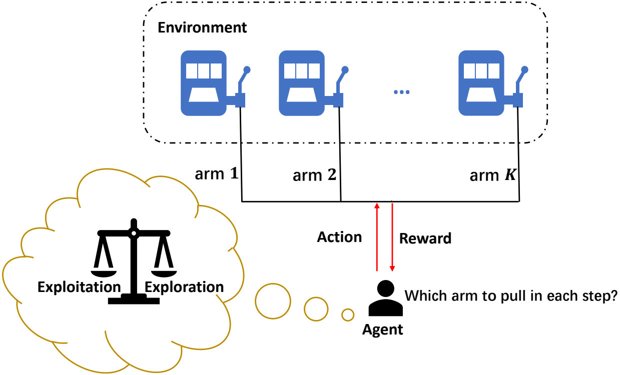


Beam Alignment as Multi-Armed Bandits



Pure Exploration: Identify the arm with the largest mean using as few samples as possible.

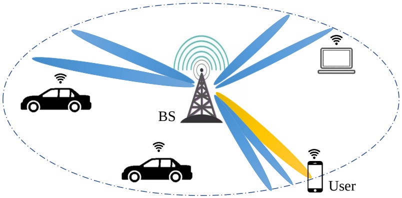
Beam Alignment as Multi-Armed Bandits



Pure Exploration: Identify the arm with the largest mean using as few samples as possible.

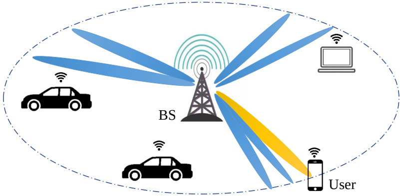
Idea: Formulate the beam alignment problem as a pure exploration problem with the objective of minimizing the required time steps in the fixed-confidence setting.

System Model: A mmWave massive MISO system



- **Massive mmWave MISO system:** a base station (BS) equipped with N transmit antennas serves a single-antenna user.

System Model: A mmWave massive MISO system

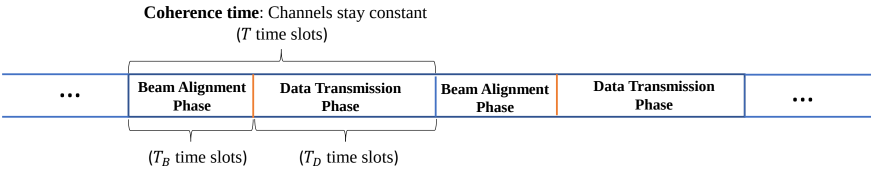


- **Massive mmWave MISO system:** a base station (BS) equipped with N transmit antennas serves a single-antenna user.
- Saleh–Valenzuela channel model (limited propagation path in mmWave channel)

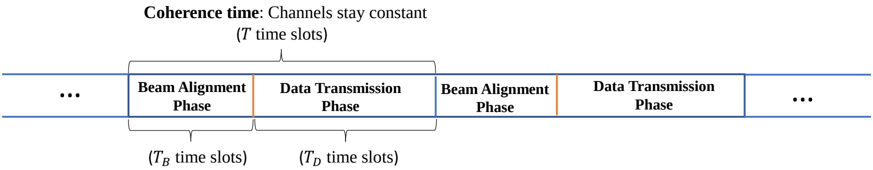
$$\mathbf{h} = \beta^{(1)}\mathbf{a}\left(\theta^{(1)}\right) + \sum_{l=2}^{L} \beta^{(l)}\mathbf{a}\left(\theta^{(l)}\right)$$

1 line-of-sight (LoS) path **Amplitude** $\geq$ L − 1 non-LoS (NLoS) paths

Transmission Scheme



Transmission Scheme



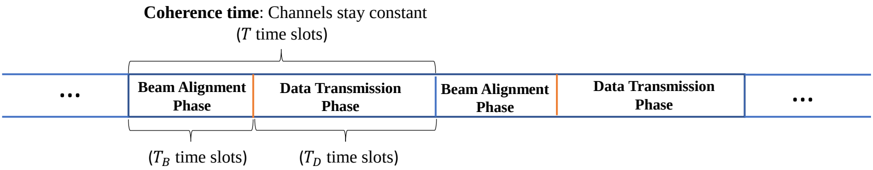
- **Beam alignment phase:** Finds the optimal beam from the codebook

$$\mathcal{C} = \{\boldsymbol{f}_k = \boldsymbol{a}(-1 + 2k/K) : k = 0, 1, \ldots, K-1\}$$

where the array response vector is

$$\boldsymbol{a}(x) = \frac{1}{\sqrt{N}} \left[1, e^{j\frac{2\pi}{\lambda}dx}, e^{j\frac{2\pi}{\lambda}2dx}, \ldots, e^{j\frac{2\pi}{\lambda}(N-1)dx}\right] \in \mathbb{C}^N.$$

Transmission Scheme



- **Beam alignment phase:** Finds the optimal beam from the codebook

$$\mathcal{C} = \{\boldsymbol{f}_k = \boldsymbol{a}(-1 + 2k/K) : k = 0, 1, \ldots, K-1\}$$

  where the array response vector is

$$\boldsymbol{a}(x) = \frac{1}{\sqrt{N}}\left[1, e^{j\frac{2\pi}{\lambda}dx}, e^{j\frac{2\pi}{\lambda}2dx}, \ldots, e^{j\frac{2\pi}{\lambda}(N-1)dx}\right] \in \mathbb{C}^N.$$

- **Data transmission phase:** Base station transmits the data using the selected $\boldsymbol{f}^* \in \mathcal{C}$. Received signal at the user in time slot t:

$$y_t = \sqrt{p}\ \boldsymbol{h}^{\mathrm{H}}\boldsymbol{f}^* s_t + n_t \qquad t \in \mathbb{N}.$$

Beam Alignment Phase

- **System Throughput Performance:** Effective achievable rate

$$R_{\mathrm{eff}} \triangleq \left(1 - \frac{T_{\mathrm{B}}}{T_{\mathrm{D}}}\right) \log\left(1 + \frac{p|\boldsymbol{h}^{\mathrm{H}}\boldsymbol{f}^*|^2}{\sigma^2}\right)$$

T_{B} should be minimized to maximize R_{eff}.

Beam Alignment Phase

- **System Throughput Performance:** Effective achievable rate

$$R_{\text{eff}} \triangleq \left(1 - \frac{T_{\mathrm{B}}}{T_{\mathrm{D}}}\right) \log\left(1 + \frac{p|\boldsymbol{h}^{\mathrm{H}}\boldsymbol{f}^*|^2}{\sigma^2}\right)$$

T_{B} should be minimized to maximize R_{eff}.

- **Measurement:** Received signal power if $\boldsymbol{f}_k$ is chosen:

$$R(\mathbf{f}_k) = |\sqrt{p}\mathbf{h}^{\mathrm{H}}\mathbf{f}_k + n|^2 = p|\mathbf{h}^{\mathrm{H}}\mathbf{f}_k|^2 + 2\sqrt{p}\Re(\mathbf{h}^{\mathrm{H}}\mathbf{f}_k n^*) + |n|^2$$

Heteroscedastic Gaussian Variable
$\mathcal{N}(p|\mathbf{h}^{\mathrm{H}}\mathbf{f}_k|^2, 2p|\mathbf{h}^{\mathrm{H}}\mathbf{f}_k|^2\sigma^2)$

Gamma Variable
$\Gamma(1, 1/\sigma^2)$

Approximate (Because: noise power $\ll$ transmit power)

$$r_k = p|\mathbf{h}^{\mathrm{H}}\mathbf{f}_k|^2 + 2\sqrt{p}\Re(\mathbf{h}^{\mathrm{H}}\mathbf{f}_k n^*)$$

Properties of/Assumptions on Beam Alignment Problem

Beam Alignment	→	Pure Exploration in MAB
Find the optimal beam as soon as possible	→	Find the optimal beam as soon as possible
beams	→	base arms
received signal power	→	rewards

Properties of/Assumptions on Beam Alignment Problem

Beam Alignment	→	Pure Exploration in MAB
Find the optimal beam as soon as possible	→	Find the optimal beam as soon as possible
beams	→	base arms
received signal power	→	rewards

Properties: Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$, and let $\mu_{(1)} \geq \mu_{(2)} \geq \ldots \geq \mu_{(K)}$.

Properties of/Assumptions on Beam Alignment Problem

Beam Alignment $\Rightarrow$ Pure Exploration in MAB

Find the optimal beam as soon as possible $\rightarrow$ Find the optimal beam as soon as possible

beams $\rightarrow$ base arms

received signal power $\rightarrow$ rewards

Properties: Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$, and let $\mu_{(1)} \geq \mu_{(2)} \geq \ldots \geq \mu_{(K)}$.

1. The means of the reward associated with arms k and i, where $|i-k| \leq J/2$, are close.

Properties of/Assumptions on Beam Alignment Problem

Beam Alignment	→	Pure Exploration in MAB
Find the optimal beam as soon as possible	→	Find the optimal beam as soon as possible
beams	→	base arms
received signal power	→	rewards

Properties: Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$, and let $\mu_{(1)} \geq \mu_{(2)} \geq \ldots \geq \mu_{(K)}$.

1. The means of the reward associated with arms k and i, where $|i - k| \leq J/2$, are close.
2. There are $K - LJ$ arms that have approximately mean zero rewards, i.e., $\mu_{(LJ+1)} \approx \mu_{(LJ+2)} \approx \ldots \approx \mu_{(K)} \approx 0$.

Properties of/Assumptions on Beam Alignment Problem

Beam Alignment	→	Pure Exploration in MAB
Find the optimal beam as soon as possible	→	Find the optimal beam as soon as possible
beams	→	base arms
received signal power	→	rewards

Properties: Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$, and let $\mu_{(1)} \geq \mu_{(2)} \geq \ldots \geq \mu_{(K)}$.

1. The means of the reward associated with arms k and i, where $|i - k| \leq J/2$, are close.
2. There are $K - LJ$ arms that have approximately mean zero rewards, i.e., $\mu_{(LJ+1)} \approx \mu_{(LJ+2)} \approx \ldots \approx \mu_{(K)} \approx 0$.
3. The variance each arm is related to its mean as follows: $\sigma_k^2 = 2\mu_k\sigma^2$.

Group Property

$\frac{1}{J}$-resolution beam codebook

- Constructed by grouping the nearby beams in the codebook $\mathcal{C}$

$$\mathcal{C}_{(J)} \triangleq \left\{ \boldsymbol{b}_g = \sum_{k=J(g-1)+1}^{Jg} \boldsymbol{f}_k \,\middle|\, g = 0, 1, \ldots, G-1 \right\}$$

Group Property

$\frac{1}{J}$-resolution beam codebook

- Constructed by grouping the nearby beams in the codebook $\mathcal{C}$

$$\mathcal{C}_{(J)} \triangleq \left\{ \boldsymbol{b}_g = \sum_{k=J(g-1)+1}^{Jg} \boldsymbol{f}_k \;\middle|\; g = 0, 1, \ldots, G-1 \right\}$$

- Received power for beam $\boldsymbol{b}_g$ (a super arm)

$$R_g = p|\boldsymbol{h}^{\mathrm{H}}\boldsymbol{b}_g|^2 + 2\sqrt{p}\Re(\boldsymbol{h}^{\mathrm{H}}\boldsymbol{b}_g n^*),$$

follows a **heteroscedastic Gaussian distribution**.

Group Property

$\frac{1}{J}$-resolution beam codebook

- Constructed by grouping the nearby beams in the codebook $\mathcal{C}$

$$\mathcal{C}_{(J)} \triangleq \left\{ \boldsymbol{b}_g = \sum_{k=J(g-1)+1}^{Jg} \boldsymbol{f}_k \,\middle|\, g = 0, 1, \ldots, G-1 \right\}$$

- Received power for beam $\boldsymbol{b}_g$ (a super arm)

$$R_g = p|\boldsymbol{h}^{\mathrm{H}}\boldsymbol{b}_g|^2 + 2\sqrt{p}\Re(\boldsymbol{h}^{\mathrm{H}}\boldsymbol{b}_g n^*),$$

  follows a **heteroscedastic Gaussian distribution**.

- Information of a set of beams can be obtained at each time step.

Bandit Beam Alignment Problem Setup

Bandit Beam Alignment Problem

Bandit Beam Alignment Problem Setup

Bandit Beam Alignment Problem

- K base arms $[K] \triangleq \{1, \ldots, K\}$: each associated with the beam $\boldsymbol{f}_k$;

Bandit Beam Alignment Problem Setup

Bandit Beam Alignment Problem

- K base arms $[K] \triangleq \{1, \ldots, K\}$: each associated with the beam $\boldsymbol{f}_k$;
- $\{[K], J\}$: set of all non-empty consecutive tuples of length $\leq J$
  - Example: $\{[6], 2\} =$ $\{\{1\}, \{1,2\}, \{2\}, \{2,3\}, \{3\}, \{3,4\}, \{4\}, \{4,5\}, \{5\}, \{5,6\}, \{6\}\}$

Bandit Beam Alignment Problem Setup

Bandit Beam Alignment Problem

- K base arms $[K] \triangleq \{1, \ldots, K\}$: each associated with the beam $\boldsymbol{f}_k$;
- $\{[K], J\}$: set of all non-empty consecutive tuples of length $\leq J$
  - Example: $\{[6], 2\} =$ $\{\{1\}, \{1,2\}, \{2\}, \{2,3\}, \{3\}, \{3,4\}, \{4\}, \{4,5\}, \{5\}, \{5,6\}, \{6\}\}$
- (K, J)-super arm: Each tuple in $\{[K], J\}$ is associated with

$$\boldsymbol{b}_g = \sum_{k=J(g-1)+1}^{Jg} \boldsymbol{f}_k \in \mathcal{C}_{(J)}.$$

Bandit Beam Alignment Problem Setup

At time step t

Bandit Beam Alignment Problem Setup

At time step t

- Choose an action (or a (K, J)-super arm) $A(t) \in \{[K], J\}$.

Bandit Beam Alignment Problem Setup

At time step t

- Choose an action (or a (K, J)-super arm) $A(t) \in \{[K], J\}$.
- Observe the reward

$$R(A(t)) = \mathcal{F}\left(\sum_{k \in A(t)} \boldsymbol{f}_k, p, \boldsymbol{h}, n_t \right)$$

where

$$\mathcal{F}(\boldsymbol{f}, p, \boldsymbol{h}, n) = p|\boldsymbol{h}^{\mathrm{H}} \boldsymbol{f}|^2 + 2\sqrt{p}\Re(\boldsymbol{h}^{\mathrm{H}} \boldsymbol{f} n^*)$$

Bandit Beam Alignment Problem Setup

At time step t

- Choose an action (or a (K, J)-super arm) $A(t) \in \{[K], J\}$.
- Observe the reward

$$R(A(t)) = \mathcal{F}\left(\sum_{k \in A(t)} \boldsymbol{f}_k, p, \boldsymbol{h}, n_t \right)$$

where

$$\mathcal{F}(\boldsymbol{f}, p, \boldsymbol{h}, n) = p|\boldsymbol{h}^{\mathrm{H}} \boldsymbol{f}|^2 + 2\sqrt{p}\Re(\boldsymbol{h}^{\mathrm{H}} \boldsymbol{f} n^*)$$

Note that for a given superarm $A \in \{[K], J\}$, the reward $R(A)$ is

$$R(A) \sim \mathcal{N}(\mu_A, 2\mu_A\sigma^2) \qquad \text{and} \qquad \mu_A = p\left| \boldsymbol{h}^{\mathrm{H}} \sum_{k \in A} \boldsymbol{f}_k \right|^2,$$

which is a heteroscedastic Gaussian distribution.

Bandit Beam Alignment Problem Setup

At time step t

- Choose an action (or a (K, J)-super arm) $A(t) \in \{[K], J\}$.
- Observe the reward

$$R(A(t)) = \mathcal{F}\left(\sum_{k \in A(t)} \boldsymbol{f}_k, p, \boldsymbol{h}, n_t \right)$$

where

$$\mathcal{F}(\boldsymbol{f}, p, \boldsymbol{h}, n) = p|\boldsymbol{h}^{\mathrm{H}} \boldsymbol{f}|^2 + 2\sqrt{p}\,\Re(\boldsymbol{h}^{\mathrm{H}} \boldsymbol{f}\, n^*)$$

Note that for a given superarm $A \in \{[K], J\}$, the reward $R(A)$ is

$$R(A) \sim \mathcal{N}(\mu_A, 2\mu_A \sigma^2) \qquad \text{and} \qquad \mu_A = p\left| \boldsymbol{h}^{\mathrm{H}} \sum_{k \in A} \boldsymbol{f}_k \right|^2,$$

which is a heteroscedastic Gaussian distribution.

Bandit Beam Alignment Problem Setup

Algorithm: $\pi := \{(\pi_t)_t, \tau^\pi, \psi^\pi, J\}$

Bandit Beam Alignment Problem Setup

Algorithm: $\pi := \{(\pi_t)_t, \tau^\pi, \psi^\pi, J\}$

- **Sampling rule** π_t: determines the (K, J)-super arm $A(t)$ to pull at time step t based on the observation history and the arm history

$$\mathcal{H}_{t-1} = \{A(1), R(1), A(2), R(2), \cdots, A(t-1), R(t-1)\}.$$

Bandit Beam Alignment Problem Setup

Algorithm: $\pi := \{(\pi_t)_t, \tau^\pi, \psi^\pi, J\}$

- **Sampling rule** π_t: determines the (K, J)-super arm $A(t)$ to pull at time step t based on the observation history and the arm history

$$\mathcal{H}_{t-1} = \{A(1), R(1), A(2), R(2), \cdots, A(t-1), R(t-1)\}.$$

- **Stopping rule**: leads to a stopping time τ^π satisfying

$$\mathbb{P}(\tau^\pi < +\infty) = 1.$$

Bandit Beam Alignment Problem Setup

Algorithm: $\pi := \{(\pi_t)_t, \tau^\pi, \psi^\pi, J\}$

- **Sampling rule** π_t: determines the (K, J)-super arm $A(t)$ to pull at time step t based on the observation history and the arm history

$$\mathcal{H}_{t-1} = \{A(1), R(1), A(2), R(2), \cdots, A(t-1), R(t-1)\}.$$

- **Stopping rule**: leads to a stopping time τ^π satisfying

$$\mathbb{P}(\tau^\pi < +\infty) = 1.$$

- **Recommendation rule** ψ^π: outputs a base arm $k^\pi \in [K]$.

Bandit Beam Alignment Problem Setup

Algorithm: $\pi := \{(\pi_t)_t, \tau^\pi, \psi^\pi, J\}$

- **Sampling rule** π_t: determines the (K, J)-super arm $A(t)$ to pull at time step t based on the observation history and the arm history

$$\mathcal{H}_{t-1} = \{A(1), R(1), A(2), R(2), \cdots, A(t-1), R(t-1)\}.$$

- **Stopping rule**: leads to a stopping time τ^π satisfying

$$\mathbb{P}(\tau^\pi < +\infty) = 1.$$

- **Recommendation rule** ψ^π: outputs a base arm $k^\pi \in [K]$.

Aim: Use as few samples as possible to output an arm that is optimal with probability at least $1 - \delta$.

Information-Theoretic Lower Bound

- Heteroscedastic Gaussian bandit instance:

$$\nu = \big(\mathcal{N}(\mu_1^\nu, 2\mu_1^\nu\sigma^2), \cdots, \mathcal{N}(\mu_K^\nu, 2\mu_K^\nu\sigma^2)\big).$$

- Optimal arm $A^*(\nu) = \arg\max_{k\in[K]} \mu_k^\nu$.

Information-Theoretic Lower Bound

- Heteroscedastic Gaussian bandit instance:

$$\nu = \big(\mathcal{N}(\mu_1^\nu, 2\mu_1^\nu\sigma^2), \cdots, \mathcal{N}(\mu_K^\nu, 2\mu_K^\nu\sigma^2)\big).$$

- Optimal arm $A^*(\nu) = \arg\max_{k\in[K]} \mu_k^\nu$.

Theorem (Lower Bound)

For any (δ, J)-PAC algorithm,

$$\mathbb{E}_\pi[\tau_\delta] \geq c^*(\nu) \log\left(\frac{1}{4\delta}\right),$$

where

$$c^*(\nu)^{-1} = \sup_{\boldsymbol{w}\in\Gamma} \inf_{\mathrm{u}\in\mathrm{Alt}(\nu)} \Big(\sum_{k=1}^K w_k D_{\mathrm{HG}}(\mu_k^\nu, \mu_k^{\mathrm{u}})\Big),$$

where D_{HG} is the KL-divergence between two heteroscedastic Gaussians.

Two-Phase Track & Stop (2PHT&S) Algorithm

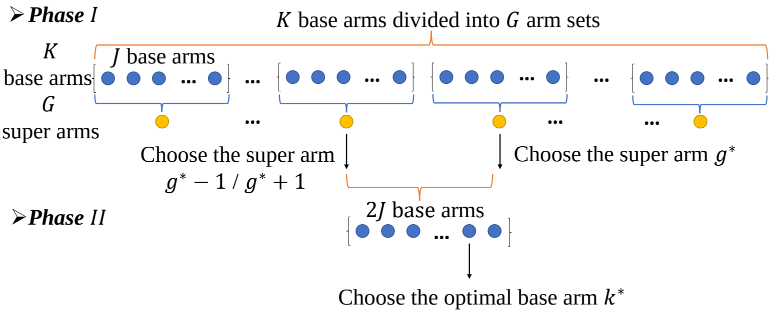
Main Idea: Exploit **prior knowledge**:

- Correlation
- Heteroscedasticity
- Group property

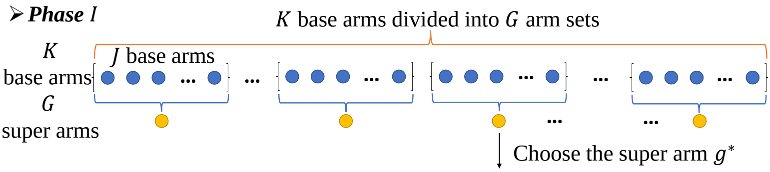
Two-Phase Track & Stop (2PHT&S) Algorithm

Main Idea: Exploit **prior knowledge**:

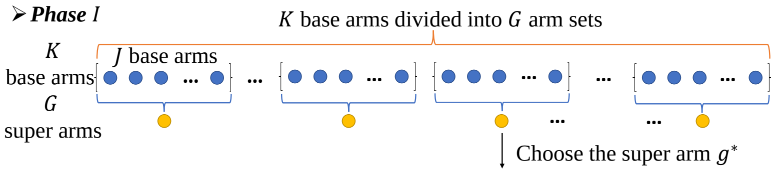
- Correlation
- Heteroscedasticity
- Group property



Two-Phase Track & Stop (2PHT&S): Phase I

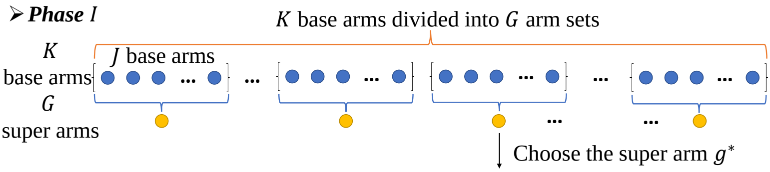


Two-Phase Track & Stop (2PHT&S): Phase I



Phase I: Search for the optimal super arm with probability $\geq 1 - \delta_1$

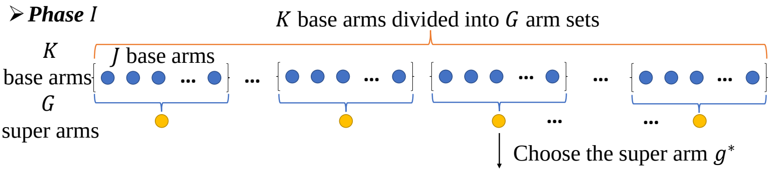
Two-Phase Track & Stop (2PHT&S): Phase I



Phase I: Search for the optimal super arm with probability $\geq 1-\delta_1$

- Group K base arms into G arm sets to reduce the search space

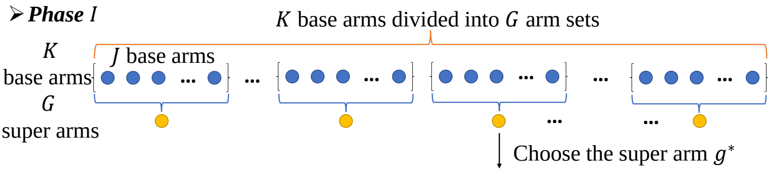
Two-Phase Track & Stop (2PHT&S): Phase I



Phase I: Search for the optimal super arm with probability $\geq 1 - \delta_1$

- Group K base arms into G arm sets to reduce the search space
- Choose one super arm (beam group) by the sampling rule of HT&S

Two-Phase Track & Stop (2PHT&S): Phase I

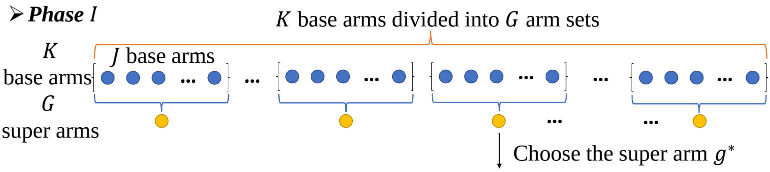


Phase I: Search for the optimal super arm with probability $\geq 1 - \delta_1$

- Group K base arms into G arm sets to reduce the search space
- Choose one super arm (beam group) by the sampling rule of HT&S
- Use the grouped beam to transmit the pilot symbols and observe

$$R_g(t) = \mathcal{F}\left(\sum_{k \in S_g} \boldsymbol{f}_k, p, \boldsymbol{h}, n_t\right).$$

Two-Phase Track & Stop (2PHT&S): Phase I



Phase I: Search for the optimal super arm with probability $\geq 1 - \delta_1$

- Group K base arms into G arm sets to reduce the search space
- Choose one super arm (beam group) by the sampling rule of HT&S
- Use the grouped beam to transmit the pilot symbols and observe

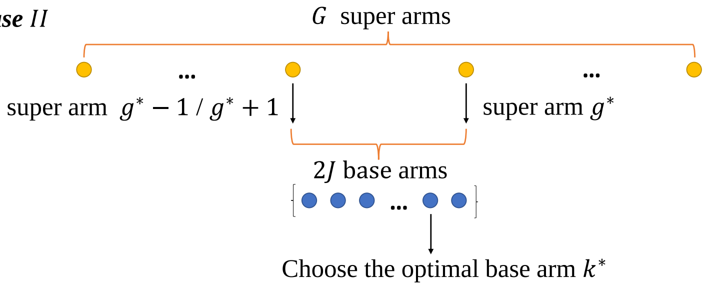
$$R_g(t) = \mathcal{F}\left(\sum_{k \in S_g} \boldsymbol{f}_k, p, \boldsymbol{h}, n_t \right).$$

- Select the optimal super arm

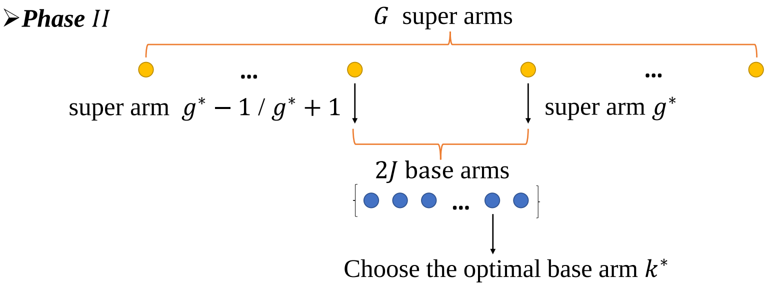
$$g^* = \arg\max_{g \in [G]} \mathbb{E}[R_g(t)].$$

Two-Phase Track & Stop (2PHT&S): Phase II

➢ ***Phase II***



Two-Phase Track & Stop (2PHT&S): Phase II



Phase II: Search for the optimal base arm with probability $\geq 1 - \delta_2$

- Construct a **base arm set**, including the optimal super arm and its neighboring super arm
- Search the optimal base arm in the **base arm set** using the HT&S algorithm

HT&S Algorithm: An improved T&S Algorithm

- **Sampling Rule:** Estimate the number of times each arm should be sampled

$$Q(t) = \begin{cases} \arg\min_{i\in[K]} T_i(t-1), & \min_{i\in[K]} T_i(t-1) \le \sqrt{t}, \\ \arg\max_{i\in[K]} t\hat{w}_i^*(t-1) - T_i(t-1), & \text{otherwise.} \end{cases}$$

HT&S Algorithm: An improved T&S Algorithm

- **Sampling Rule:** Estimate the number of times each arm should be sampled

$$Q(t) = \begin{cases} \arg\min_{i \in [K]} T_i(t-1), & \min_{i \in [K]} T_i(t-1) \le \sqrt{t}, \\ \arg\max_{i \in [K]} t\hat{w}_i^*(t-1) - T_i(t-1), & \text{otherwise.} \end{cases}$$

- **Stopping Rule:** Stop when the numbers of times all arms are pulled satisfy

$$\tau_\delta = \min\Big\{ t \in \mathbb{N} : Z(t) \ge \beta(t, \delta, \alpha) \Big\}.$$

HT&S Algorithm: An improved T&S Algorithm

- **Sampling Rule:** Estimate the number of times each arm should be sampled

$$Q(t) = \begin{cases} \arg\min_{i\in[K]} T_i(t-1), & \min_{i\in[K]} T_i(t-1) \le \sqrt{t}, \\ \arg\max_{i\in[K]} t\hat{w}_i^*(t-1) - T_i(t-1), & \text{otherwise.} \end{cases}$$

- **Stopping Rule:** Stop when the numbers of times all arms are pulled satisfy

$$\tau_\delta = \min\Big\{t \in \mathbb{N} : Z(t) \ge \beta(t,\delta,\alpha)\Big\}.$$

- **Heteroscedasticity:** Considered in $\hat{w}_i^*(t-1)$ and $Z(t)$.

Sample Complexity Analysis of 2PHT&S

Sample Complexity Analysis of 2PHT&S

Theorem (Performance of 2PHT&S)

Let

$$\mathrm{s} = \left(\mathcal{N}(\mu_1^{\mathrm{s}}, 2\mu_1^{\mathrm{s}}\sigma^2), \ldots, \mathcal{N}(\mu_G^{\mathrm{s}}, 2\mu_G^{\mathrm{s}}\sigma^2)\right) \quad \text{and}$$

$$\mathrm{b} = \left(\mathcal{N}(\mu_{\mathcal{S}_f(1)}^{\mathrm{b}}, 2\mu_{\mathcal{S}_f(1)}^{\mathrm{b}}\sigma^2), \ldots, \mathcal{N}(\mu_{\mathcal{S}_f(2J)}^{\mathrm{b}}, 2\mu_{\mathcal{S}_f(2J)}^{\mathrm{b}}\sigma^2)\right)$$

be the super arm and base arm heteroscedastic Gaussian bandits in Phase I and Phase II, where

$$\mu_g^{\mathrm{s}} = p\left|\boldsymbol{h}^{\mathrm{H}}\Big(\sum_{k\in\mathcal{S}_g} \boldsymbol{f}_k\Big)\right|^2.$$

Sample Complexity Analysis of 2PHT&S

Theorem (Performance of 2PHT&S)

Let

$$\mathrm{s} = \big(\mathcal{N}(\mu_1^{\mathrm{s}}, 2\mu_1^{\mathrm{s}}\sigma^2), \ldots, \mathcal{N}(\mu_G^{\mathrm{s}}, 2\mu_G^{\mathrm{s}}\sigma^2)\big) \quad \textit{and}$$

$$\mathrm{b} = \big(\mathcal{N}(\mu_{\mathcal{S}_f(1)}^{\mathrm{b}}, 2\mu_{\mathcal{S}_f(1)}^{\mathrm{b}}\sigma^2), \ldots, \mathcal{N}(\mu_{\mathcal{S}_f(2J)}^{\mathrm{b}}, 2\mu_{\mathcal{S}_f(2J)}^{\mathrm{b}}\sigma^2)\big)$$

be the super arm and base arm heteroscedastic Gaussian bandits in Phase I and Phase II, where

$$\mu_g^{\mathrm{s}} = p\left|\boldsymbol{h}^{\mathrm{H}}\Big(\sum_{k\in\mathcal{S}_g} \boldsymbol{f}_k\Big)\right|^2.$$

Using 2PHT&S, we obtain

$$\limsup_{\delta\to 0} \frac{\mathbb{E}[\tau^{\mathit{2PHT\&S}}]}{\log(1/\delta)} \le C_{\mathrm{s}}^{-1} + C_{\mathrm{b}}^{-1},$$

where C_{s} and C_{b} are hardness parameters of Phase I and Phase II resp.

Simulation Results

Experiment Setup

- Massive mmWave MISO system;
- Base station equipped with $N = 64$ transmit antennas serving a single-antenna user;
- Size of codebook is set as $K = 128$.
- Correlation Length $J = 2\left\lceil \frac{K}{N} \right\rceil - 1 = 3$.

Simulation Results

Experiment Setup

- Massive mmWave MISO system;
- Base station equipped with $N = 64$ transmit antennas serving a single-antenna user;
- Size of codebook is set as $K = 128$.
- Correlation Length $J = 2\left\lceil \frac{K}{N} \right\rceil - 1 = 3$.

Baseline Algorithms

- Original Track-and-Stop (T&S) algorithm (Garivier and Kaufmann, 2016);
- HT&S algorithm;
- Two-phase Track-and-Stop (2PT&S) algorithm.

Simulated Scenario for $\delta = 0.1$ and $\delta_1 = \delta_2 = \frac{\delta}{2}$

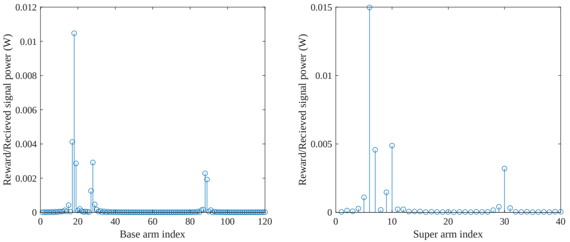


Figure 2: Mean of the reward of each base arm and super arm in ($p = 10$dBm).

Simulated Scenario for $\delta = 0.1$ and $\delta_1 = \delta_2 = \frac{\delta}{2}$

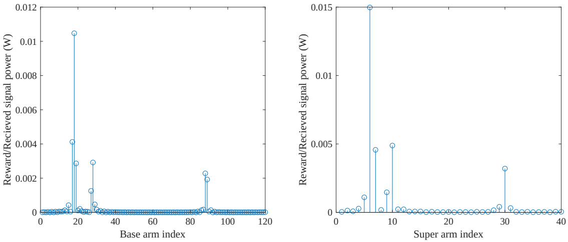
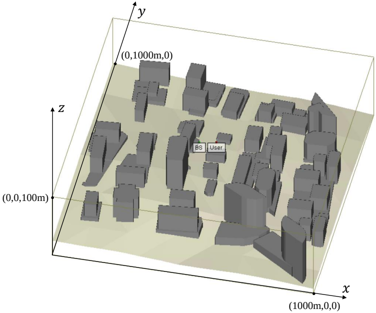


Figure 2: Mean of the reward of each base arm and super arm in ($p = 10$dBm).

Table 2: Average sample complexities for $\delta = 0.1$, averaged over 100 experiments.

Power	4	6	8	10	12
T&S	1154.3 $\pm$338.7	654.6 $\pm$212.1	382.5 $\pm$129.6	209.4$\pm$68.6	133.7 $\pm$8.9
HT&S	473.2 $\pm$275.5	271.4 $\pm$143.4	175.6 $\pm$69.2	133.2 $\pm$24.1	123.9 $\pm$6.5
2PT&S	206.2 $\pm$60.4	120.2 $\pm$35.0	68.4 $\pm$19.4	49.1 $\pm$4.6	45.2 $\pm$1.1
2HPT&S	84.3 $\pm$41.5	58.0 $\pm$19.6	48.4 $\pm$6.3	45.5 $\pm$1.6	45 $\pm$0

Practical Scenario: Generated using Wireless InSite



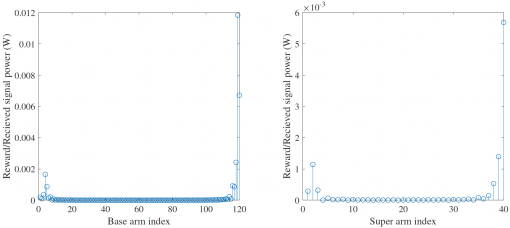


Figure 3: (Left) Practical beam alignment in a city; (Right) Means of the rewards of each base and super arm. Sample complexities for $\delta = 0.1$ shown below.

Practical Scenario: Generated using Wireless InSite

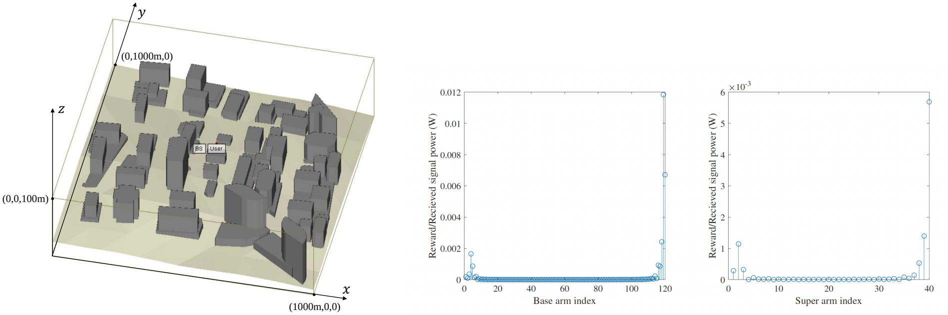


Figure 3: (Left) Practical beam alignment in a city; (Right) Means of the rewards of each base and super arm. Sample complexities for $\delta = 0.1$ shown below.

Power	4	6	8	10	12
T&S	840.6 $\pm$331.1	540.5.9 $\pm$190.9	339.1 $\pm$138.8	231.1 $\pm$95.8	162.7 $\pm$59.6
HT&S	515.5 $\pm$305.1	345.2 $\pm$186.4	253.9 $\pm$122.6	176.1 $\pm$71.1	141.3 $\pm$45.0
2PT&S	189.9 $\pm$43.2	119.1 $\pm$29.8	138.8 $\pm$82.8	55.8 $\pm$18.4	45.4 $\pm$3.9
2PHT&S	74.4 $\pm$33.9	57.6 $\pm$20.6	50.7 $\pm$14.9	45.8 $\pm$5.5	45 $\pm$0

Conclusions

- Adapted multi-armed bandit framework to beam alignment.
- Exploited structure to get improved results over naïve techniques.

Conclusions

- Adapted multi-armed bandit framework to beam alignment.
- Exploited structure to get improved results over naïve techniques.