Probe-then-Commit Multi-Objective Bandits: Theoretical Benefits of Limited Multi-Arm Feedback

Ming Shi

Department of Electrical Engineering, University at Buffalo, Buffalo, NY

Abstract—We study an online resource-selection problem motivated by multi-radio access selection and mobile edge computing offloading. In each round, an agent chooses among K candidate links/servers (arms) whose performance is a stochastic d-dimensional vector (e.g., throughput, latency, energy, reliability). The key interaction is *probe-then-commit (PtC)*: the agent may probe up to $q > 1$ candidates via control-plane measurements to observe their vector outcomes, but must execute exactly one candidate in the data plane. This limited multi-arm feedback regime strictly interpolates between classical bandits ($q = 1$) and full-information experts ($q = K$), yet existing multi-objective learning theory largely focuses on these extremes. We develop PTC-P-UCB, an optimistic probe-then-commit algorithm whose technical core is frontier-aware probing under uncertainty in a Pareto mode, e.g., it selects the q probes by approximately maximizing a hypervolume-inspired frontier-coverage potential and commits by marginal hypervolume gain to directly expand the attained Pareto region. We prove a dominated-hypervolume frontier error of $\tilde{O}(K_P d/\sqrt{qT})$, where K_P is the Pareto-frontier size and T is the horizon, and scalarized regret $\tilde{O}(L_\phi d\sqrt{(K/q)T})$, where ϕ is the scalarizer. These quantify a transparent $1/\sqrt{q}$ acceleration from limited probing. We further extend to *multi-modal probing*: each probe returns M modalities (e.g., CSI, queue, compute telemetry), and uncertainty fusion yields variance-adaptive versions of the above bounds via an effective noise scale.

Index Terms—multi-objective bandits, probe-then-commit (PtC), limited multi-arm feedback, Pareto frontier, scalarization, hypervolume, multi-modal feedback, online resource selection

I. INTRODUCTION

Next-generation wireless and edge systems increasingly rely on online selection among multiple candidate network resources while meeting heterogeneous service requirements [1]. For example, in multi-radio access technology (multi-RAT) selection, a user equipment (UE) chooses among candidate links (e.g., 5G New Radio, WiFi, unmanned aerial vehicle relay) whose instantaneous channel quality, contention, and scheduling delay fluctuate across slots [2]. For another example, in mobile edge computing (MEC) offloading, a device selects an edge server whose queueing delay, compute load, and radio access conditions jointly determine end-to-end latency and success probability [3]. These decisions are inherently *multi-objective*. Optimizing a single scalar metric can yield operating points that violate service-level objectives (SLOs), or systematically sacrifice a "weak" KPI (key performance indicators) to gain another (e.g., sacrifice reliability to gain throughput).

Existing online-learning abstractions tend to focus on two extremes. Multi-objective bandits (MOB) observe only the vector outcome of the *single* executed arm per slot [4]–[7],

while full-information online optimization and learning (and vector-payoff approachability) observes outcomes of *all K* arms each round [8], [9]. However, wireless or edge systems often operate in a distinct intermediate regime enabled by the control plane. A UE can *probe* a small set of candidates using channel state information reference signals (CSI-RS) and beam sweeping, beacon frames or round-trip time (RTT) pings, or queue/CPU telemetry, but it can execute only one link/server due to data-plane constraints (one transmission/offload per slot). This yields *limited multi-arm feedback*, which is richer than standard MOB, but much cheaper than full information.

We formalize this interaction as a per-round *Probe-then-Commit* (PtC) protocol. At each slot t, the learner selects a probe set S_t with $|S_t| = q$, observes vector outcomes $\{\mathbf{r}_t(k)\}_{k \in S_t}$ (control-plane feedback), and then commits to one executed resource $k_t \in S_t$ whose outcome incurs realized system performance (data-plane execution). This model exposes an explicit *optimization-feedback tradeoff*: increasing q improves information and should accelerate learning, while probing consumes measurement and signaling budget. This PtC regime bridges MOB and full-information experts, but requires new algorithmic and analytical tools to handle vector objectives and frontier criteria under probe-limited feedback.

Wireless/edge systems often need to operate across multiple modes. For example, one needs to consider energy-saving against latency-critical. Thus, a learner should discover the Pareto frontier rather than only optimize one fixed operating point. Accordingly, we evaluate preference-free learning by a hypervolume-based frontier coverage metric, and we evaluate preference-based learning by scalarized regret for monotone concave utilities, e.g., fairness-sensitive aggregations.

PtC multi-objective learning couples probe design and execution in ways that do not arise in classical models. (i) *Uncertainty-aware frontier exploration:* probe-set design must lift multiple KPIs while diversifying across frontier regions; (ii) *Frontier accuracy under partial feedback:* coverage guarantees require controlling a set-valued error (hypervolume gap); (iii) *Quantifying the value of limited probing:* theory should expose how q sharpens rates beyond the bandit regime; (iv) *Multi-modal sensing:* probing often returns multiple modalities (CSI, queue length, CPU load) with heterogeneous noise, necessitating fusion that preserves valid confidence bounds [10].

Our main contributions are summarized as follows.

- **PtC multi-objective multi-feedback model (Sec. II).** We introduce a stochastic multi-objective MAB under the PtC protocol with probe budget q. We formalize two complemen-

tary evaluation metrics, preference-free frontier learning via a dominated-hypervolume coverage gap and preference-based learning via scalarized regret for monotone concave utilities.

- **Algorithmic ideas: uncertainty-aware frontier coverage from probed samples (Sec. IV).** We develop PTC-P-UCB (Algorithm 2), which elects the q probes by approximately maximizing a hypervolume-inspired frontier-coverage potential and commits by marginal hypervolume gain to directly expand the attained Pareto region. The design is compatible with both frontier-coverage evaluation (hypervolume) and preference-based operation when a scalarizer is specified.
- **Theory: explicit value of limited multi-arm feedback (Sec. V).** We prove a dominated-hypervolume frontier coverage gap that vanishes at rate $\tilde{O}(K_P d/\sqrt{qT})$ (with frontier size K_P, d objectives, and T rounds), and a scalarized regret bound of order $\tilde{O}(d\sqrt{(K/q)T})$ for monotone L_ϕ-Lipschitz concave scalarizers, where $\tilde{O}$ hides constants and logarithmic terms. These results quantify a clean $1/\sqrt{q}$ *improvement* from limited multi-arm probing, interpolating between the bandit limit ($q = 1$) and the full-information limit ($q = K$).
- **Multi-modal extension with variance-adaptive gains (Sec. VI).** We extend the framework to bundled multi-modal probing with M modalities. We develop MM-PTC-P-UCB (Algorithm 3) and show that fusion tightens confidence bounds through an effective noise scale, yielding variance-adaptive improvements for both frontier coverage and regret.
- **Empirical validation.** Simulations on multi-RAT- and MEC-inspired instances corroborate our theory. Modest probing budgets (e.g., $q \in \{2, 4\}$) significantly accelerate learning and improve Pareto coverage with moderate overhead, and multi-modal fusion provides an additional orthogonal gain.

A. Related Work

1) Multi-objective bandits with Pareto criteria and scalarizations: Multi-objective bandits study vector-valued rewards and compare actions via Pareto optimality or preference scalarizations. Upper confidence bound (UCB)-style approaches for vector rewards and Pareto efficiency appear in early MOB work (e.g., [4]), while preference-based learning uses monotone utilities, e.g., inequality-averse aggregations such as generalized Gini [6], to encode fairness-sensitive tradeoffs and enable regret analysis [5], [11]. More recent work develops Pareto-oriented regret notions that avoid fixing a scalarizer [7], [12]. Our focus is complementary. We study an intermediate *probe-limited multi-feedback* regime induced by wireless probing protocols, and provide guarantees for both frontier coverage (hypervolume) and preference-based regret under PtC feedback.

2) Multiple-play, semi-bandits, and side-observation models: Observing multiple actions per round relates to power-of-2-arms [13], multiple-play bandits [14], combinatorial/semi-bandit models [15], and learning with structured side observations such as feedback graphs [16]. These works are predominantly *single-objective* and typically assume additive reward/loss decompositions, whereas PtC enforces a single executed action with vector outcomes and Pareto criteria. This changes both the algorithmic goal (probe-set design for

frontier coverage) and the analysis (simultaneously controlling frontier estimation error and preference-based regret under probe-limited sampling).

3) Vector-payoff online learning and approachability: Full-information vector-payoff learning connects to Blackwell approachability [8] and its equivalence to no-regret learning [17]. This regime also includes standard online convex optimization with function information [9], [18]. Our PtC model in this paper can be viewed as a probe-limited, partial-information counterpart tailored to wireless/edge measurement pipelines.

4) Hypervolume as a Pareto-set quality measure: Dominated hypervolume is a standard quality measure for Pareto sets in multi-objective optimization and evolutionary computation [19], [20]. We adopt a hypervolume gap as a principled metric for frontier discovery over time under PtC feedback.

5) Multi-modal sensing and fusion in wireless/edge decision making: Next-generation wireless and edge platforms expose heterogeneous *modalities* correlated with service quality, including radio measurements, active probes, and system telemetry (e.g., queue/compute load). Recent wireless research also emphasizes multi-modal learning at the network level, including foundation-model perspectives for 6G systems [10]. To our knowledge, we take the first effort to model multi-modal probing as *multiple noisy views* of the same underlying multi-objective outcome vector, and design confidence-bound-driven learning rules whose uncertainty tightens via an effective variance under fusion. This yields variance-adaptive improvements in learning performance, while preserving the explicit $1/\sqrt{q}$ benefit of limited multi-arm probing under the PtC protocol.

Notation: For an integer n, $[n] = \{1, 2, \ldots, n\}$. For a scalar x, $[x]^+ \triangleq \max\{0, x\}$. For $u, v \in \mathbb{R}^d$, $u \succeq v$ denotes component-wise inequality and $\|u\|_\infty = \max_j |u_j|$. We write $\Delta_d = \{w \in \mathbb{R}_+^d : \sum_{j=1}^d w_j = 1\}$ for the probability simplex.

II. PROBLEM FORMULATION

We study a multi-objective online esource-selection problem motivated by wireless access and edge computing systems, where each decision must balance heterogeneous KPIs under limited probing and measurement opportunities. The key interaction is *probe-then-commit* (PtC): in each round, the agent may probe multiple candidates via control-plane feedback, but ultimately executes only one due to data-plane constraints.

A. System Model and PtC Feedback Protocol

There are K candidate resources (arms), indexed by $k \in [K]$ (e.g., access links, edge servers), and d objectives indexed by $j \in [d]$ (e.g., throughput, latency, energy, reliability). Time is slotted with horizon T. At each round $t \in [T]$, each arm k is associated with a random vector-valued reward outcome

$$\mathbf{r}_t(k) = \left(r_t^{(1)}(k), \ldots, r_t^{(d)}(k)\right) \in [0, 1]^d, \quad (1)$$

with *unknown* mean $\boldsymbol{\mu}(k) = \mathbb{E}[\mathbf{r}_t(k)] \in [0, 1]^d$. The d coordinates represent heterogeneous KPIs. If a KPI is naturally minimized (e.g., delay/energy), we convert it to a maximization objective by negating and normalizing it to $[0, 1]$.

Algorithm 1 Probe-then-Commit (PtC) interaction at round t

Require: Probe budget $q \in [K]$.
 1: **Probe selection:** choose a probe set $S_t \subseteq [K]$ with $|S_t| = q$.
 2: **Measurement:** observe vector outcomes $\{\mathbf{r}_t(k)\}_{k \in S_t}$.
 3: **Commit:** select one executed arm $k_t \in S_t$.
 4: **Realization:** incur the round performance $\mathbf{r}_t(k_t)$ and proceed to next round $t+1$.

In many practical scenarios (e.g., multi-RAT selection and MEC offloading), a device can probe and measure multiple candidates (e.g., channel probing, active RTT pings, queue/CPU reports), but can use only one for actual transmission/offloading in that round. We model this by PtC (see Algorithm 1).

Probe outcomes for all $k \in S_t$ are observed (control-plane measurement), but only the executed arm k_t contributes to realized system performance in that round (data-plane execution). Probing may incur overhead (time/energy/signaling). We treat q as a fixed per-round budget in the main theory and discuss explicit probing-cost models in Sec. II-C.

B. Pareto Structure and Preference Scalarizers

The vector nature of $\boldsymbol{\mu}(k)$ induces a partial order, i.e., an arm may be better in one KPI but worse in another. We therefore formalize preference-free efficiency through Pareto dominance and, when the system operates under a fixed preference, use scalarizers to select a single operating point.

1) Pareto dominance and frontier: For $u, v \in \mathbb{R}^d$, we say that u dominates v (denoted $u \succ v$) if $u_j \geq v_j$ for all j and $u_{j'} > v_{j'}$ for some j'. Given mean vectors $\{\boldsymbol{\mu}(k)\}_{k=1}^K$, define the Pareto frontier $\mathcal{P}^* = \mathrm{Pareto}\big(\{\boldsymbol{\mu}(k)\}_{k=1}^K\big)$, i.e., the set of mean vectors not dominated by any other. Moreover, we let $K_P \triangleq |\mathcal{P}^*|$ denote the size of the true frontier $\mathcal{P}^*$.

2) Scalarizers with system preferences: A deployed system may need a particular tradeoff based on operator policy or user preference. We model such preferences by scalarizers $\phi : \mathbb{R}^d \to \mathbb{R}$ that are *monotone* (improving any objective cannot decrease utility), L_ϕ-*Lipschitz* w.r.t. ℓ_∞ (i.e., $|\phi(u) - \phi(v)| \leq L_\phi \|u - v\|_\infty$), and *concave*. These conditions enable stable aggregation and regret analysis. Standard examples include: (i) weighted sum $\phi_w(u) = \sum_{j=1}^d w_j u_j$ with $w \in \Delta_d$; (ii) Chebyshev scalarizer $\phi_w^{\min}(u) = \min_{j \in [d]} w_j u_j$; and (iii) generalized Gini $\phi_\gamma(u) = \sum_{i=1}^d \gamma_i u_{(i)}$, where $u_{(1)} \leq \cdots \leq u_{(d)}$ are the sorted components and $\gamma_1 \geq \cdots \geq \gamma_d \geq 0$ [4], [6].

C. Performance Metrics and Probe Overhead

We evaluate algorithms using two complementary criteria. *Hypervolume coverage* quantifies preference-free frontier learning performance, while *scalarized regret* measures learning and operational loss under a specified system preference.

1) Hypervolume-based Pareto coverage: To quantify preference-free frontier learning, we use dominated hypervolume (HV) with respect to a fixed reference point $z_{\mathrm{ref}} \in \mathbb{R}^d$

that is component-wise worse than all attainable performance vectors. For compact set $\mathcal{S} \subseteq \mathbb{R}^d$, define the dominated region

$$\mathcal{D}(\mathcal{S}) \triangleq \big\{ y \in \mathbb{R}^d : \exists u \in \mathcal{S}, \text{ s.t. } z_{\mathrm{ref}} \preceq y \preceq u \big\}, \quad (2)$$

and let $\mathcal{H}(\mathcal{S})$ be the Lebesgue measure of $\mathcal{D}(\mathcal{S})$. Intuitively, larger $\mathcal{H}(\mathcal{S})$ means that $\mathcal{S}$ contains operating points that jointly perform well across objectives and spans a broader range of Pareto tradeoffs. Then, we let $\mathcal{Y}_T \triangleq \{\mathbf{r}_t(k_t)\}_{t=1}^T$ be the archive of executed outcome vectors and define the attained set

$$\mathcal{A}_T \triangleq \mathrm{conv}(\mathcal{Y}_T), \quad (3)$$

where the convex hull captures time-sharing among operating points. Since the same time-sharing interpretation applies to the Pareto benchmark, define the convexified Pareto set $\mathcal{C}^* \triangleq \mathrm{conv}(\mathcal{P}^*)$. We measure the remaining uncovered dominated volume by the attained-set hypervolume gap [19], [20]

$$\mathcal{L}_T^{\mathrm{HV}} \triangleq [\mathcal{H}(\mathcal{C}^*) - \mathcal{H}(\mathcal{A}_T)]^+. \quad (4)$$

By construction, $\mathcal{L}_T^{\mathrm{HV}} \geq 0$. Smaller $\mathcal{L}_T^{\mathrm{HV}}$ indicates that the executed decisions achieve tradeoffs whose dominated region approaches that of the (time-shareable) Pareto benchmark.

2) Scalarized regret: Fix a scalarizer ϕ and define the best arm in hindsight $k^* \triangleq \arg\max_{k \in [K]} \phi(\boldsymbol{\mu}(k))$. Then, the scalarized regret is defined as follows,

$$R_T^\phi = \sum_{t=1}^T \big(\phi(\boldsymbol{\mu}(k^*)) - \phi(\mathbf{r}_t(k_t))\big). \quad (5)$$

3) Probe overhead: Probing consumes control-plane resources (time and energy). A simple model assigns a per-probe cost $\tau > 0$ and penalizes each round by $-\tau|S_t| = -\tau q$. Equivalently, one may impose a hard constraint $q \leq M_{\max}$ or a long-term budget. Our main results treat q as fixed and quantify how increased probing improves learning rates.

III. MOTIVATING EXAMPLES

We highlight two wireless/edge scenarios that directly match the PtC protocol in Algorithm 1: a device can obtain control-plane measurements from up to q candidates within a slot, but can execute only one candidate on the data plane.

1) Multi-RAT link selection: A UE probes up to q candidate access points or gNBs using lightweight control-plane signals (e.g., pilot measurements, beacons, short RTT probes, or queue indicators), obtaining a KPI vector such as throughput, delay, energy, and reliability. It then commits to one link for data transmission, so only the executed link's outcome is realized.

2) MEC offloading: A device queries up to q MEC servers for multi-modal telemetry (e.g., radio quality, queue status, CPU load), forming an outcome vector that captures end-to-end latency, energy consumption, and SLO satisfaction/reliability. It then offloads to a single server, again matching PtC.

IV. ALGORITHM DESIGN

This section presents PTC-P-UCB (see Algorithm 2), a probe-then-commit algorithm for multi-objective learning under PtC feedback. The algorithm maintains arm-wise confidence bounds from probed samples, selects a probe set to accelerate learning, and then commits to one probed arm for execution.

Algorithm 2 PTC-P-UCB: Probe-then-Commit Pareto-UCB
___
Require: Probe budget q, weights $w \in \Delta_d$, reference point z_{ref}, confidence parameter $\{\beta_t\}$. Commit mode: SCALAR (use given ϕ) or HV (use marginal hypervolume gain)

1: Initialize $N_1(k) \leftarrow 0$, $\widehat{\mu}_1^{(j)}(k) \leftarrow 0$ for all $k \in [K]$ and $j \in [d]$, active set $\mathcal{K}_1 \leftarrow [K]$ and executed archive $\mathcal{Y}_0 \leftarrow \emptyset$
2: **for** $t = 1$ to T **do**
    Confidence bounds (from probed samples):
3:     **for** each $k \in \mathcal{K}_t$ and $j \in [d]$ **do**
4:         Calculate $b_t^{(j)}(k)$ using (7)
5:         Calculate $\text{UCB}_t^{(j)}(k)$ and $\text{LCB}_t^{(j)}(k)$ using (8)
6:     **end for**
7:     Define $u_t(k)$ and $\ell_t(k)$ for all $k \in \mathcal{K}_t$
    Pruning:
8:     $\widetilde{\mathcal{K}}_t \leftarrow \{k \in \mathcal{K}_t : \nexists k' \in \mathcal{K}_t \text{ s.t. } \ell_t(k') \succ u_t(k)\}$
9:     $\mathcal{K}_t \leftarrow \widetilde{\mathcal{K}}_t$
    Probe selection:
10:     **if** HV **then**   ▷ HV
11:         Select $S_t \subseteq \mathcal{K}_t$ by maximization of $F_t(S)$ in (10)
12:     **else**   ▷ SCALAR
13:         Select $S_t \subseteq \mathcal{K}_t$ by the surrogate top-q rule (11)
14:     **end if**
15:     Probe and observe $\{\mathbf{r}_t(k)\}_{k \in S_t}$
    Commit:
16:     **if** HV **then**   ▷ HV
17:         $k_t \leftarrow \arg\max_{k \in S_t} \Delta_t^{\mathcal{H}}(k)$ using (12)
18:     **else**   ▷ SCALAR
19:         $k_t \leftarrow \arg\max_{k \in S_t} \phi(\mathbf{r}_t(k))$
20:     **end if**
21:     Execute k_t and incur $\mathbf{r}_t(k_t)$
22:     **if** HV **then**
23:         Update archive $\mathcal{Y}_t \leftarrow \mathcal{Y}_{t-1} \cup \{\mathbf{r}_t(k_t)\}$
24:     **end if**
    Updates (for all probed arms):
25:     **for** each $k \in S_t$ **do**
26:         $N_{t+1}(k) \leftarrow N_t(k) + 1$
27:         **for** each $j \in [d]$ **do**
28:             $\widehat{\mu}_{t+1}^{(j)}(k) \leftarrow \widehat{\mu}_t^{(j)}(k) + \frac{r_t^{(j)}(k) - \widehat{\mu}_t^{(j)}(k)}{N_{t+1}(k)}$
29:         **end for**
30:     **end for**
31:     For $k \notin S_t$, $N_{t+1}(k) \leftarrow N_t(k)$, $\widehat{\mu}_{t+1}^{(j)}(k) \leftarrow \widehat{\mu}_t^{(j)}(k)$.
32:     Set $\mathcal{K}_{t+1} \leftarrow \mathcal{K}_t$.
33: **end for**
___

The PtC protocol requires executing exactly one arm per round. Our hypervolume coverage metric in (4) is defined on the executed attained set $\mathcal{A}_T$, and therefore depends on which arm is committed each round. To obtain preference-free frontier coverage for $\mathcal{A}_T$, we use a coverage-aware commit rule based on marginal hypervolume gain. However, if the system is operated under a known preference, the commit step can maximize a scalarizer ϕ to minimize scalarized regret.

Our design follows three main principles as follows.

1) *Probe selection should increase frontier coverage under uncertainty.* Since probing provides the side information (q observations per round), the probe set should **include** arms that are plausibly Pareto-efficient under optimism and **diversify** across frontier regions. We achieve this by approximately maximizing a hypervolume-based coverage potential over optimistic vectors, with a score-based surrogate.

2) *Multi-objective optimism with component-wise confidence.* Maintain per-objective confidence intervals and form optimistic vectors to guide probing and safe elimination.

3) *Commit must match the metric.* For preference-free frontier learning under (4), we commit using marginal hypervolume gain of the executed archive. For preference-based operation we commit using ϕ to minimize scalarized regret.

A. Preliminaries for Confidence Bounds

Since the learner observes outcomes for all probed arms, we index learning progress by the number of times an arm has been probed, $N_t(k) \triangleq \sum_{\tau=1}^{t-1} \mathbb{I}\{k \in S_\tau\}$. For each objective $j \in [d]$, maintain the empirical mean based on probed samples,

$$\widehat{\mu}_t^{(j)}(k) \triangleq \frac{1}{\max\{1, N_t(k)\}} \sum_{\tau=1}^{t-1} \mathbb{I}\{k \in S_\tau\} r_\tau^{(j)}(k). \quad (6)$$

Choose a confidence parameter $\beta_t = 2\log(2Kdt^2/\delta)$ for Hoeffding-style bounds. Define the bonus term for each (k, j),

$$b_t^{(j)}(k) \triangleq \sqrt{\beta_t / \max\{1, N_t(k)\}}. \quad (7)$$

Then, we form clipped upper/lower bounds

$$\text{UCB}_t^{(j)}(k) = \min\left\{1, \widehat{\mu}_t^{(j)}(k) + b_t^{(j)}(k)\right\},$$
$$\text{LCB}_t^{(j)}(k) = \max\left\{0, \widehat{\mu}_t^{(j)}(k) - b_t^{(j)}(k)\right\}. \quad (8)$$

We define the optimistic and pessimistic vectors

$$u_t(k) \triangleq \left(\text{UCB}_t^{(1)}(k), \ldots, \text{UCB}_t^{(d)}(k)\right),$$
$$\ell_t(k) \triangleq \left(\text{LCB}_t^{(1)}(k), \ldots, \text{LCB}_t^{(d)}(k)\right). \quad (9)$$

With high probability, $\ell_t(k) \preceq \mu(k) \preceq u_t(k)$ component-wise for all k, t, enabling safe pruning and optimistic probe selection.

B. Probe Selection via Frontier-Coverage Potential

A key decision is how to choose the probe set S_t of size q. Selecting the top-q arms by a single scalar score may overly concentrate probing around one region of the frontier. To encourage *diverse* coverage, we define a set-based potential.

1) Coverage potential (set function): Let z_{ref} be the hypervolume reference point (as in Section II-C). Given optimistic vectors $\{u_t(k)\}_{k=1}^K$, define the potential of a probe set S as

$$F_t(S) \triangleq \mathcal{H}\left(\text{conv}\{u_t(k)\}_{k \in S}\right), \quad (10)$$

i.e., the dominated hypervolume of the convexified optimistic set. This potential rewards probe sets whose optimistic vectors jointly dominate a large region, which aligns with minimizing the Pareto coverage gap (up to optimism and estimation error).

2) Greedy probe selection: Maximizing $F_t(S)$ over all $|S| = q$ is combinatorial. We therefore use a greedy approximation. Starting from $S = \emptyset$, iteratively add the arm with the largest marginal gain in F_t. When $F_t(\cdot)$ is (approximately) monotone submodular, greedy achieves a constant-factor approximation.

3) Fast surrogate (general scalarizer): When scalarized regret is the metric, we use a modular surrogate obtained by applying a preference scalarizer to the optimistic vector, i.e.,

$$\text{score}_t^\phi(k) \triangleq \phi\big(u_t(k)\big). \tag{11}$$

We then set S_t to be the q arms in $\mathcal{K}_t$ with largest $\text{score}_t^\phi(k)$.

C. Commit Rule for Execution

After probing, the learner observes $\{\mathbf{r}_t(k)\}_{k \in S_t}$ and must execute one arm $k_t \in S_t$. We provide two commit rules that correspond to the two evaluation objectives.

1) Coverage-based commit (for hypervolume coverage on $\mathcal{A}_T$): Let $\mathcal{Y}_{t-1}$ denote the archive of executed outcomes up to round $t-1$ (so that $\mathcal{A}_{t-1} = \text{conv}(\mathcal{Y}_{t-1})$). For each candidate $k \in S_t$, define the marginal hypervolume gain

$$\Delta_t^{\mathcal{H}}(k) \triangleq \mathcal{H}\left(\text{conv}\left(\mathcal{Y}_{t-1} \cup \{\mathbf{r}_t(k)\}\right)\right) - \mathcal{H}\left(\text{conv}(\mathcal{Y}_{t-1})\right). \tag{12}$$

We then commit to the arm that optimizes this gain, i.e.,

$$k_t^{\mathcal{H}} \in \arg\max_{k \in S_t} \Delta_t^{\mathcal{H}}(k). \tag{13}$$

Intuitively, this chooses the probed arm that most expands the dominated region of the attained set, directly targeting $\mathcal{L}_T$.

2) Preference-based commit (for scalarized regret): Given a preference scalarizer ϕ, we commit to the best observed arm

$$k_t^\phi \in \arg\max_{k \in S_t} \phi(\mathbf{r}_t(k)). \tag{14}$$

This extracts immediate operational value from probing.

D. Frontier Pruning

To concentrate probing on plausible Pareto-frontier arms, we maintain an active set $\mathcal{K}_t \subseteq [K]$. An arm k can be safely discarded if it is certifiably dominated, i.e., $\exists k' \in \mathcal{K}_t$, s.t. $\ell_t(k') \succ u_t(k)$. On the high-probability event $\ell_t(\cdot) \preceq \mu(\cdot) \preceq u_t(\cdot)$, this rule never removes a true Pareto arm, but can dramatically reduce computation and improve the dependence on the frontier size K_P in the coverage analysis. Moreover, since $\mu(k') \succeq \mu(k)$ implies $\phi(\mu(k')) \geq \phi(\mu(k))$ for monotone scalarizer ϕ, the rule also never removes a ϕ-optimal arm.

E. Complexity and Probe Overhead

Per round, computing confidence bounds costs $O(|\mathcal{K}_t|d)$. Probe selection costs $O(q|\mathcal{K}_t|C_{\mathcal{H}})$ under greedy hypervolume or $O(K \log K)$ under the score-based surrogate, where $C_{\mathcal{H}}$ is the cost of a hypervolume marginal computation (small for $d \leq 4$). The probing overhead scales linearly with q in control-plane signaling. A per-probe cost τ can be incorporated by constraining $q \leq M_{\max}$ or subtracting τ from the scalar utility in deployment, without changing the learning machinery.

V. THEORETICAL RESULTS

This section provides performance guarantees for PTC-P-UCB (Algorithm 2) under the stochastic PtC model for the two evaluation metrics in Sec. II-C. We provide the proof sketches for main theorems, while complete proofs and supporting lemmas are provided in our technical report [21].

A. Assumption and Key Bookkeeping

Let $\mathcal{F}_t$ be the filtration generated by past probe sets and observed probes up to the end of round t. A central identity is

$$\sum\nolimits_{k=1}^{K} N_{T+1}(k) = qT, \tag{15}$$

i.e., each round yields q vector samples. Compared with traditional bandits, the learner observes q times more arm outcomes per round thanks to probes, which shrinks estimation error faster and translates into improved learning rates.

Assumption 1 (Conditionally sub-Gaussian noise)**.** *For each $k \in [K]$ and $j \in [d]$, the noise $r_t^{(j)}(k) - \mu^{(j)}(k)$ is conditionally σ-sub-Gaussian given $\mathcal{F}_{t-1}$ and independent over round t.*

B. Preference-Free Pareto Frontier Learning

We first analyze frontier learning under the HV mode of Algorithm 2, where the probe set is chosen to increase a hypervolume-based coverage potential and the commit step selects the probed arm with the largest marginal hypervolume gain. Bounding $\mathcal{L}_T^{\text{HV}}$ (defined in (4)) requires controlling two distinct effects. (i) *Learning error:* whether the algorithm probes enough to discover near-frontier arms (this is where q helps via (15)). (ii) *Execution sampling error:* the attained set uses instantaneous outcomes $\mathbf{r}_t(k_t)$ rather than mean vector $\mu(k_t)$, which introduces an additional statistical fluctuation.

Theorem 1 (Attained-set hypervolume gap of PTC-P-UCB (HV mode))**.** *Under Assumption 1, run Algorithm 2 and the marginal-gain commit rule* (12)–(13)*. Then, we have*

$$\mathbb{E}\left[\mathcal{L}_T^{\text{HV}}\right] = \tilde{O}\left(K_P d/\sqrt{qT} + d/\sqrt{T}\right). \tag{16}$$

The first term in (16) is the *frontier-learning* term. First, probing indeed accelerates estimation of Pareto-relevant arms. This yields a $1/\sqrt{q}$ improvement in the performance. Second, the factor K_P captures frontier complexity. The dependence on K_P suggests that the dominant learning burden is to resolve and cover the K_P Pareto-relevant mean vectors, while dominated arms do not directly affect hypervolume. The second term in (16) reflects the fact that the attained set is formed from instantaneous executed outcomes rather than means. It vanishes as T grows and is unavoidable for an execution-based metric.

Proof sketch. We define the denoised archive $\tilde{\mathcal{Y}}_T \triangleq \{\hat{\boldsymbol{\mu}}_T(k_t)\}_{t=1}^T$, with $\tilde{\mathcal{A}}_T \triangleq \text{conv}(\tilde{\mathcal{Y}}_T)$. We upper bound the hypervolume gap by inserting $\text{conv}(\{\boldsymbol{\mu}(k_t)\}_{t=1}^T)$, i.e.,

$$\mathcal{L}_T^{\text{HV}} \leq \underbrace{\left[\mathcal{H}(\mathcal{C}^*) - \mathcal{H}(\text{conv}(\{\boldsymbol{\mu}(k_t)\}_{t=1}^T))\right]^+}_{\text{learning and coverage error}}$$
$$+ \underbrace{\left|\mathcal{H}(\text{conv}(\{\boldsymbol{\mu}(k_t)\}_{t=1}^T)) - \mathcal{H}(\tilde{\mathcal{A}}_T)\right|}_{\text{estimation error}}. \tag{17}$$

Step 1 (learning and coverage error). On the event that all coordinate-wise confidence intervals (CIs) hold, the probe selection uses optimistic vectors $u_t(k)$, so any Pareto-relevant arm with large uncertainty induces a large marginal gain in the optimistic coverage potential. A standard potential argument then shows that each Pareto arm is probed enough that its CI shrinks to $\tilde{O}(1/\sqrt{qT})$. Combining this with hypervolume stability over K_P frontier points yields $\mathcal{H}(\mathcal{C}^*) - \mathcal{H}(\text{conv}(\{\boldsymbol{\mu}(k_t)\})) = \tilde{O}(K_P d/\sqrt{qT})$ in expectation.

Step 2 (estimation error). A hypervolume stability lemma for sets in $[0,1]^d$ gives $\left| \mathcal{H}(\text{conv}(U)) - \mathcal{H}(\text{conv}(V)) \right| \leq L_{\mathcal{H}} \cdot d_H(\text{conv}(U), \text{conv}(V))$, and $d_H(\text{conv}(U), \text{conv}(V)) \leq \max_{t \leq T} \|u_t - v_t\|_\infty$. Taking $u_t = \boldsymbol{\mu}(k_t)$ and $v_t = \widehat{\boldsymbol{\mu}}_T(k_t)$ yields $\left| \mathcal{H}(\text{conv}(\{\boldsymbol{\mu}(k_t)\})) - \mathcal{H}(\tilde{\mathcal{A}}_T) \right| \leq L_{\mathcal{H}} \cdot \max_{t \leq T} \|\widehat{\boldsymbol{\mu}}_T(k_t) - \boldsymbol{\mu}(k_t)\|_\infty$. Uniform concentration under adaptive probing implies $\max_{k \in [K]} \|\widehat{\boldsymbol{\mu}}_T(k) - \boldsymbol{\mu}(k)\|_\infty = \tilde{O}(1/\sqrt{qT})$ in expectation (using $\sum_k N_{T+1}(k) = qT$), hence the estimation term is $\tilde{O}(d/\sqrt{qT})$.

Combining Steps 1–2 gives the claimed rate. $\square$

C. Fixed-Confidence ϵ-Frontier Identification

We next translate confidence bounds into a fixed-confidence sample complexity guarantee for identifying an ϵ-accurate frontier approximation. An arm k is ϵ-*Pareto optimal* (in ℓ_∞) if there is no k' such that $\mu(k') \succeq \mu(k) + \epsilon\mathbf{1}$. Using the coordinate-wise confidence bounds, define the output set

$$\widehat{\mathcal{P}}_T^{(\epsilon)} \triangleq \Big\{ k \in [K] :$$
$$\not\exists k' \in [K], \text{ s.t. } \mathsf{LCB}_T(k') \succeq \mathsf{UCB}_T(k) + \epsilon\mathbf{1} \Big\}. \quad (18)$$

This rule is conservative, i.e., on the event that all confidence intervals are valid, it produces no ϵ-dominated false positives.

Theorem 2 (Sample complexity for ϵ-frontier identification)**.** *Fix $\epsilon, \delta \in (0, 1)$. Under Assumption 1, on the event that all coordinate-wise confidence intervals hold, $\widehat{\mathcal{P}}_T^{(\epsilon)}$ contains all truly Pareto-optimal arms and contains no arm that is ϵ-dominated. With probability at least $1 - \delta$, it suffices that*

$$qT \geq C \cdot K_P d \log\left(Kd/\delta\right)/\epsilon^2, \quad (19)$$

for a universal constant $C > 0$. Equivalently, the number of probed samples required is $N_\epsilon = \tilde{O}\left(\frac{K_P d}{\epsilon^2}\right)$ or $T = \tilde{O}\left(\frac{K_P d}{q\epsilon^2}\right)$.

For fixed confidence (ϵ, δ), increasing q reduces the required horizon linearly, since it increases the number of observed arm vectors per round. The dependence on K_P formalizes that only Pareto-relevant arms must be resolved to ϵ accuracy.

D. Scalarized Regret in SCALAR Mode

We finally turn to preference-based operation. In SCALAR mode, the learner probes using the optimistic scalar index $\text{score}_t^\phi(k) = \phi(u_t(k))$ and commits via (14).

Theorem 3 (Scalarized regret of PTC-P-UCB (SCALAR mode))**.** *Under Assumptions 1, run Algorithm 2. Then,*

$$\mathbb{E}[R_T^\phi] = \tilde{O}\left(L_\phi d\sqrt{KT/q}\right). \quad (20)$$

Moreover, with probability at least $1 - \delta$, $\delta \in (0,1)$, the same rate holds up to $\text{polylog}(K, d, T, 1/\delta)$ factors.

Relative to the standard bandit rate $\tilde{O}(\sqrt{KT})$, probing yields an effective sample-size increase by $\sqrt{1/q}$ because q arms are observed per round. The additional d factor arises from uniform control over coordinates and Lipschitz stability of ϕ.

Remark 1 (Boundary cases). When $q = 1$, (20) recovers the standard scalarized multi-objective bandit rate up to logs. When $q = K$, all arms are observed each round and the rate becomes $\tilde{O}(L_\phi d\sqrt{T})$, matching full-information scaling.

Proof sketch. Let $\Delta(k) \triangleq \phi(\mu(k^*)) - \phi(\mu(k))$ denote the gap. Let $\mathcal{E}$ be the high-probability event on which all coordinate-wise CIs hold uniformly. By a union bound and sub-Gaussian concentration, $\mathbb{P}(\mathcal{E}) \geq 1 - \delta$ for an appropriate β_t.

Step 1 (optimism implies that any selected arm must still be "uncertain enough"). On $\mathcal{E}$, monotonicity of ϕ gives $\phi(\mu(k)) \leq \phi(u_t(k))$ for all k. Define the optimistic score $U_t(k) \triangleq \phi(u_t(k))$ used for probe selection. Since S_t contains the top-q arms by $U_t(\cdot)$, whenever an arm k is probed, it must satisfy $U_t(k) \geq U_t(k^*) \geq \phi(\mu(k^*))$ on $\mathcal{E}$. Therefore, $\Delta(k) = \phi(\mu(k^*)) - \phi(\mu(k)) \leq \phi(u_t(k)) - \phi(\mu(k)) \leq L_\phi \|u_t(k) - \mu(k)\|_\infty \leq L_\phi \max_j b_t^{(j)}(k)$, where the last inequality uses $\mu^{(j)}(k) \leq u_t^{(j)}(k) \leq \mu^{(j)}(k) + b_t^{(j)}(k)$ on $\mathcal{E}$. Hence, if $\Delta(k) > \varepsilon$, then arm k can only remain in the top-q probe set while $\max_j b_t^{(j)}(k) \gtrsim \varepsilon/L_\phi$, which implies $N_t(k) \lesssim \beta_T(L_\phi/\varepsilon)^2$ (up to constants and d).

Step 2 ("parallel exploration" converts probe complexity into a $1/q$ reduction in time). Let $\mathcal{K}_\varepsilon = \{k : \Delta(k) > \varepsilon\}$. From Step 1, each $k \in \mathcal{K}_\varepsilon$ needs at most $O(\beta_T(L_\phi/\varepsilon)^2)$ probes before its optimistic score drops below $\phi(\mu(k^*))$ on $\mathcal{E}$, after which it cannot enter the top-q set again. Since each round allocates q probes, the total number of rounds in which any arm in $\mathcal{K}_\varepsilon$ can still be probed is at most $O\left(\frac{|\mathcal{K}_\varepsilon|}{q} \cdot \beta_T \frac{L_\phi^2}{\varepsilon^2}\right)$.

Step 3 (gap-free regret via a ε-decomposition). Decompose the regret as $\sum_{t=1}^T \Delta(k_t) \leq T\varepsilon + \sum_{t:\Delta(k_t)>\varepsilon} \Delta(k_t)$. The first term is $T\varepsilon$. For the second term, upper bound each $\Delta(k_t)$ by 1 and use Step 2 to bound the number of rounds where $\Delta(k_t) > \varepsilon$, yielding $\sum_{t:\Delta(k_t)>\varepsilon} \Delta(k_t) \leq O\left(\frac{K}{q} \cdot \beta_T \frac{L_\phi^2}{\varepsilon^2}\right)$. Choosing $\varepsilon \asymp L_\phi \sqrt{\frac{K\beta_T}{qT}}$ gives $\tilde{O}\left(L_\phi \sqrt{\frac{KT}{q}}\right)$. Applying a union bound over d coordinates in $\mathcal{E}$ introduces the stated d factor.

Step 4 (from high-probability to expectation and realized regret). On $\mathcal{E}^c$ we use the trivial bound $R_T^\phi \leq T$. Thus, $\mathbb{E}[R_T^\phi] \leq \tilde{O}(L_\phi d\sqrt{KT/q}) + T\delta$, and taking $\delta = 1/T$ yields the stated expectation bound. For the realized reward scalarizer $\phi(r_t(k_t))$, an additional term controlling the selection-dependent deviation $\sum_t (\phi(\mu(k_t)) - \phi(r_t(k_t)))$ is handled via sub-Gaussian maximal inequalities and Lipschitzness of ϕ, and does not change the leading $\tilde{O}(L_\phi d\sqrt{KT/q})$ order. $\square$

VI. EXTENSION: MULTI-MODAL FEEDBACK

In many wireless/edge systems, probing a candidate resource returns *multiple modalities* of side information, e.g., CSI

Algorithm 3 MM-PTC-P-UCB: a multi-modal extension

Require: Probe budget q, fusion weights $\alpha \in \Delta_M$, confidence parameter $\{\beta_t\}$.

1: Run PTC-P-UCB (Algorithm 2) but:
2:     (i) when $k \in S_t$, observe $\{\mathbf{z}_t^{(m)}(k)\}_{m=1}^M$ and form $\widetilde{\mathbf{r}}_t(k) = \sum_m \alpha_m \mathbf{z}_t^{(m)}(k)$;
3:     (ii) update $\widehat{\mu}_t(k)$ using $\widetilde{\mathbf{r}}_t(k)$ instead of $\mathbf{r}_t(k)$;
4:     (iii) use the effective-scale-based confidence radii $b_t^{(j)}(k) = \sigma_{\text{eff}}^{(j)}(k)\sqrt{\beta_t / \max\{1, N_t(k)\}}$.

---

measurements, queue-length reports, and CPU-load telemetry. When fused properly, these modalities can improve learning.

A. Multi-Modal Observation Model

We consider M modalities indexed by m. For each round t and arm k, there is a vector-valued outcome $\mathbf{r}_t(k) \in [0,1]^d$ with mean $\boldsymbol{\mu}(k)$. Under multi-modal probing, whenever $k \in S_t$ the learner observes *all* modality readings $\{\mathbf{z}_t^{(m)}(k)\}_{m=1}^M$, where

$$\mathbf{z}_t^{(m)}(k) = \mathbf{r}_t(k) + \boldsymbol{\eta}_t^{(m)}(k). \tag{21}$$

We assume $\boldsymbol{\eta}_t^{(m)}(k)$ is conditionally mean-zero given $\mathcal{F}_{t-1}$. For clarity, we state a diagonal (objective-wise) sub-Gaussian version: for each objective j, $\eta_t^{(m,j)}(k)$ is conditionally $\sigma_m^{(j)}(k)$-sub-Gaussian given $\mathcal{F}_{t-1}$ and independent over t for each fixed (k, m, j). Then, given fusion weights $\alpha = (\alpha_1, \ldots, \alpha_M) \in \Delta_M$, we define the fused observation as follows,

$$\widetilde{\mathbf{r}}_t(k) \triangleq \sum_{m=1}^M \alpha_m \mathbf{z}_t^{(m)}(k). \tag{22}$$

$\widetilde{\mathbf{r}}_t(k)$ is an unbiased noisy observation of $\mathbf{r}_t(k)$, and $\widetilde{r}_t^{(j)}(k) - r_t^{(j)}(k)$ is conditionally sub-Gaussian with effective scale $\left(\sigma_{\text{eff}}^{(j)}(k)\right)^2 = \sum_{m=1}^M \alpha_m^2 \left(\sigma_m^{(j)}(k)\right)^2$. Intuitively, multi-modality yields an *orthogonal* acceleration mechanism. Besides the m-fold sample increase from probing, fusion can reduce per-sample uncertainty through σ_{eff}.

B. Algorithm: MM-PTC-P-UCB

The multi-modal extension (Algorithm 3) involves a *local* change to the learning pipeline in Algorithm 2. We replace the single probed sample $\mathbf{r}_t(k)$ by the fused sample $\widetilde{\mathbf{r}}_t(k)$ in the mean updates, and replace the base noise scale by σ_{eff} in the confidence radii. Probe selection and the commit rule follow exactly the modes as Sec. IV. In particular, in the multi-modal model, the learner only observes modality feedback, not directly observing $\mathbf{r}_t(k)$. Thus, both in HV mode (commit via marginal hypervolume gain) and in SCALAR mode (commit via $\phi(\cdot)$), it is needed to compute the commit decision using the best available estimate of the probed outcome, i.e., $\widetilde{\mathbf{r}}_t(k)$.

C. Guarantees: Variance-Adaptive Improvement

With fixed fusion weights α, the analysis in Sec. V carries over by replacing the base noise scale by σ_{eff} in the concentration arguments. Intuitively, PtC provides q samples per round, while fusion reduces the noise per sample. Recall that in HV mode the learner maintains the executed archive

$\mathcal{Y}_t = \{\widetilde{\mathbf{r}}_s(k_s)\}_{s=1}^t$ and the attained set $\mathcal{A}_t = \text{conv}(\mathcal{Y}_t)$, where $\widetilde{\mathbf{r}}_t(k)$ is considered here and is the fused observation in (22).

Theorem 4 (Variance-adaptive attained-set hypervolume gap under fixed fusion). *Consider the multi-modal model* (21)–(22) *with fixed fusion weights* $\alpha \in \Delta_M$. *Let* $\sigma_{\text{eff}} \triangleq \max_{k \in [K], j \in [d]} \sigma_{\text{eff}}^{(j)}(k)$. *Run* PTC-P-UCB *in* HV *mode, using the coverage-based commit rule* (13). *Then, we have*

$$\mathbb{E}\left[\mathcal{L}_T^{\text{HV}}\right] = \tilde{O}\left(K_P d\sigma_{\text{eff}}/\sqrt{qT} + d/\sqrt{T}\right). \tag{23}$$

Eq. (23) makes the two acceleration mechanisms explicit. First, the PtC probe budget contributes the same $1/\sqrt{q}$ improvement via the identity $\sum_k N_{T+1}(k) = qT$. Second, multi-modal fusion improves the *per-sample* statistical accuracy by shrinking the effective noise scale from σ to σ_{eff}.

Proof sketch. The proof has three steps. (1) Under the sub-Gaussian assumption and the fused estimator (22), uniform coordinate-wise concentration yields, for all k and j, $|\widehat{\mu}_T^{(j)}(k) - \mu^{(j)}(k)| \lesssim \sigma_{\text{eff}}^{(j)}(k)\sqrt{\log(\cdot)/N_{T+1}(k)}$. (2) Using Cauchy–Schwarz together with $\sum_k N_{T+1}(k) = qT$ bounds the aggregate estimation error on Pareto-relevant arms by $\tilde{O}(\sigma_{\text{eff}}/\sqrt{qT})$. (3) A hypervolume stability lemma upper-bounds the perturbation of $\mathcal{H}(\cdot)$ over sets in $[0,1]^d$ by $O(K_P d)$ times the ℓ_∞ estimation error, which yields (23). $\square$

Theorem 5 (Variance-adaptive scalarized regret under fixed fusion). *Consider the bundled multi-modal model* (21)-(22) *with fixed* $\alpha \in \Delta_M$. *Let* $\sigma_{\text{eff}} \triangleq \max_{k \in [K], j \in [d]} \sigma_{\text{eff}}^{(j)}(k)$. *Run* PTC-P-UCB *with confidence radii scaled by* σ_{eff} *yields*

$$\mathbb{E}\left[R_T^\phi\right] = \tilde{O}\left(L_\phi d\sigma_{\text{eff}}\sqrt{KT/q}\right). \tag{24}$$

VII. NUMERICAL RESULTS

We evaluate PTC-P-UCB and MM-PTC-P-UCB under both metrics from Sec. II-C. We use synthetic instances motivated by wireless/edge tradeoffs and include near-dominated "confusers" that make dominance relations statistically fragile.

We simulate $K = 24$ arms and $d = 4$ objectives over horizon T. To induce a nontrivial frontier while retaining hard-to-separate alternatives, we generate means by a mixture construction, i.e., first sample a subset of "frontier" arms from clustered points on a tradeoff surface, and then generate remaining arms as dominated perturbations around these clusters. This yields realistic regimes where small estimation errors can flip pairwise dominance, slowing frontier learning.

We run PtC with probe budgets $q \in \{1, 2, 4, K\}$. For multi-modal experiments ($M = 3$), each probe returns a set $\mathbf{z}_t^{(m)}(k) = \mathbf{r}_t(k) + \boldsymbol{\eta}_t^{(m)}(k)$ with heterogeneous scales $(\sigma_m)_{m=1}^M = (0.08, 0.12, 0.20)$. Moreover, we fuse modalities using inverse-variance weights $\alpha_p \propto 1/(\bar{\sigma}_p^2)$.

A. Main Results and Findings

1) Frontier discovery improves with limited probing: Fig. 1 reports the frontier hypervolume gap $\mathcal{G}_T^{\text{HV}}$. Increasing the probe budget yields consistently faster decay. Moving from $q = 1$ to $q = 4$ substantially reduces the time needed to reach the same
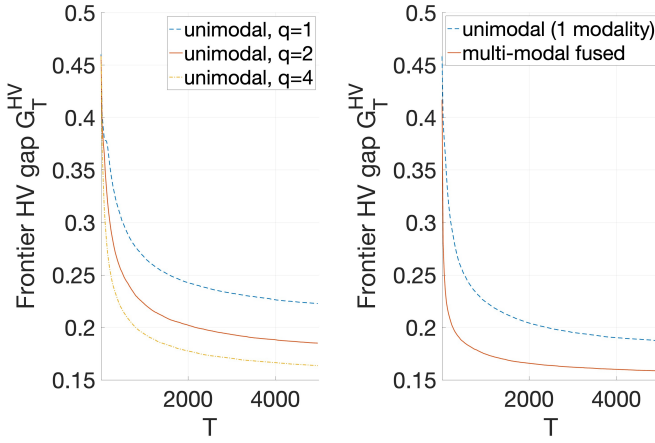
Fig. 1. Frontier hypervolume gap $\mathcal{G}_T^{\mathrm{HV}}$ versus T: effect of q and benefit of multi-modal fusion (set $q = 2$).
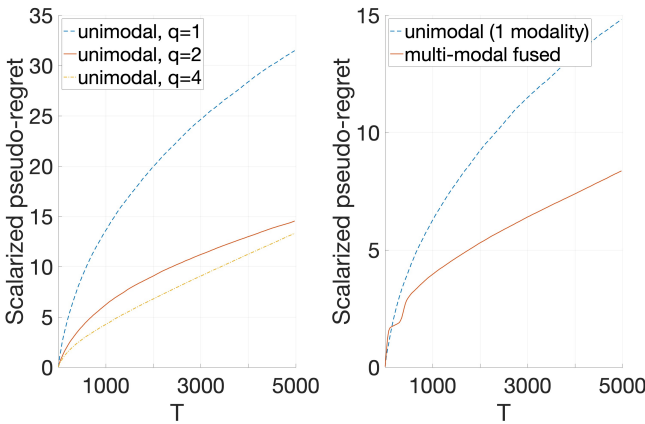

Fig. 2. Worst-case scalarized regret versus T: effect of q and benefit of multi-modal fusion (set $q = 2$).

coverage level. This matches the predicted $1/\sqrt{q}$ acceleration in the hypervolume guarantee (Theorem 1) and highlights that multi-feedback probing accelerates frontier learning, not merely exploitation under a fixed preference.

2) Scalarized performance accelerates at the same $1/\sqrt{q}$ rate: Fig. 2 shows worst-case scalarized regret. Across environments, the ordering $q = 4 < q = 2 < q = 1$ persists throughout the horizon. Moreover, the separations between curves are consistent with the theoretical scaling $\tilde{O}\big(\sqrt{K/(qT)}\big)$ from Theorem 3. This explicitly confirms that the q side observations translate into an effective sample-size gain.

3) Multi-modal fusion yields an orthogonal variance-reduction gain: With $M = 3$ modalities, inverse-variance fusion reduces σ_{eff}, tightening confidence bounds and improving both metrics at fixed q. Fig. 1 shows that fused feedback reaches the same hypervolume gap earlier than unimodal sensing, consistent with the variance-adaptive analysis in Sec. VI.

VIII. Conclusion

We introduced a *multi-objective multi-feedback* MAB capturing realistic probing in wireless/edge systems, designed PtC-

P-UCB, and proved regret and Pareto coverage bounds with explicit q-dependence. Our multi-modal extension is variance-adaptive and practically effective. Future work includes contextual/linear structure, delayed feedback, and nonstationarity.

References

[1] S. Lin, M. Shi, A. Arora, R. Bassily, E. Bertino, C. Caramanis, K. Chowdhury, E. Ekici, A. Eryilmaz, S. Ioannidis *et al.*, "Leveraging synergies between ai and networking to build next generation edge networks," in *2022 IEEE 8th International Conference on Collaboration and Internet Computing (CIC)*. IEEE, 2022, pp. 16–25.

[2] C. Li, J. Jiang, and J. Li, "Multi-radio access technology (multi-rat) diversity for ultra-reliable low-latency communication (urllc)," Nov. 10 2020, uS Patent 10,833,902.

[3] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE communications surveys & tutorials*, vol. 19, no. 3, pp. 1628–1656, 2017.

[4] M. M. Drugan and A. Nowe, "Designing multi-objective multi-armed bandits algorithms: A study," in *The 2013 international joint conference on neural networks (IJCNN)*. IEEE, 2013, pp. 1–8.

[5] L. Cao, M. Shi, and N. B. Shroff, "Provably efficient multi-objective bandit algorithms under preference-centric customization," *arXiv preprint arXiv:2502.13457*, 2025.

[6] J. A. Weymark, "Generalized gini inequality indices," *Mathematical social sciences*, vol. 1, no. 4, pp. 409–430, 1981.

[7] D. M. Roijers, P. Vamplew, S. Whiteson, and R. Dazeley, "A survey of multi-objective sequential decision-making," *Journal of Artificial Intelligence Research*, vol. 48, pp. 67–113, 2013.

[8] D. Blackwell, "An analog of the minimax theorem for vector payoffs." 1956.

[9] E. Hazan *et al.*, "Introduction to online convex optimization," *Foundations and Trends® in Optimization*, vol. 2, no. 3-4, pp. 157–325, 2016.

[10] J. Du, T. Lin, C. Jiang, Q. Yang, C. F. Bader, and Z. Han, "Distributed foundation models for multi-modal learning in 6g wireless networks," *IEEE Wireless Communications*, vol. 31, no. 3, pp. 20–30, 2024.

[11] M. Agarwal, V. Aggarwal, and T. Lan, "Multi-objective reinforcement learning with non-linear scalarization," in *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, 2022, pp. 9–17.

[12] A. Garivier, W. M. Koolen *et al.*, "Sequential learning of the pareto front for multi-objective bandits," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2024, pp. 3583–3591.

[13] M. Shi, X. Lin, and L. Jiao, "Power-of-2-arms for adversarial bandit learning with switching costs," *IEEE Transactions on Networking*, 2025.

[14] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: Iid rewards," *IEEE Transactions on Automatic Control*, vol. 32, no. 11, pp. 968–976, 2003.

[15] B. Kveton, Z. Wen, A. Ashkan, and C. Szepesvari, "Tight regret bounds for stochastic combinatorial semi-bandits," in *Artificial Intelligence and Statistics*. PMLR, 2015, pp. 535–543.

[16] N. Alon, N. Cesa-Bianchi, O. Dekel, and T. Koren, "Online learning with feedback graphs: Beyond bandits," in *Conference on Learning Theory*. PMLR, 2015, pp. 23–35.

[17] J. Abernethy, P. L. Bartlett, and E. Hazan, "Blackwell approachability and no-regret learning are equivalent," in *Proceedings of the 24th Annual Conference on Learning Theory*. JMLR Workshop and Conference Proceedings, 2011, pp. 27–46.

[18] M. Shi, X. Lin, and S. Fahmy, "Competitive online convex optimization with switching costs and ramp constraints," *IEEE/ACM Transactions on Networking*, vol. 29, no. 2, pp. 876–889, 2021.

[19] E. Zitzler and L. Thiele, "Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach," *IEEE transactions on Evolutionary Computation*, vol. 3, no. 4, pp. 257–271, 2002.

[20] A. Auger, J. Bader, D. Brockhoff, and E. Zitzler, "Theory of the hypervolume indicator: optimal μ-distributions and the choice of the reference point," in *Proceedings of the tenth ACM SIGEVO workshop on Foundations of genetic algorithms*, 2009, pp. 87–102.

[21] , "Probe-then-commit multi-objective bandits: Theoretical benefits of limited multi-arm feedback," https://mingshihomepage.com/CommunicationâĂŞCorruption%20Coupling%20and%20Verification%20in%20Cooperative%20MOB_ISIT2026.pdf, 2026.