Risk-Averse Best Arm Set Identification with Fixed Budget and Fixed Confidence

Shunta Nonaga NONAGA0811@EIS.HOKUDAI.AC.JP
Koji Tabata[*] KTABATA@ES.HOKUDAI.AC.JP
Yuta Mizuno MIZUNO@ES.HOKUDAI.AC.JP
Tamiki Komatsuzaki[*] TAMIKI@ES.HOKUDAI.AC.JP
Hokkaido University, Hokkaido, Japan

Editors: Hung-yi Lee and Tongliang Liu

Abstract

Decision making under uncertain environments in the maximization of expected reward while minimizing its risk is one of the ubiquitous problems in many subjects. Here, we introduce a novel problem setting in stochastic bandit optimization that jointly addresses two critical aspects of decision-making: maximizing expected reward and minimizing associated uncertainty, quantified via the *mean-variance*(MV) criterion. Unlike traditional bandit formulations that focus solely on expected returns, our objective is to efficiently and accurately identify the Pareto-optimal set of arms that strikes the best trade-off between expected performance and risk. We propose a unified meta-algorithmic framework capable of operating under both fixed-confidence and fixed-budget regimes, achieved through adaptive design of confidence intervals tailored to each scenario using the same sample exploration strategy. We provide theoretical guarantees on the correctness of the returned solutions in both settings. To complement this theoretical analysis, we conduct extensive empirical evaluations across synthetic benchmarks, demonstrating that our approach outperforms existing methods in terms of both accuracy and sample efficiency, highlighting its broad applicability to risk-aware decision-making tasks in uncertain environments.

Keywords: Stochastic multi-armed bandits; Multi-objective optimization; Pareto set identification

1. Introduction

Stochastic multi-armed bandit (MAB) problems Lattimore and Szepesvári (2020) have emerged as a fundamental framework for online decision making under uncertainty, with broad applications ranging from adaptive drug discovery to recommendation systems Madhukar et al. (2017); Qin et al. (2014); Li et al. (2010, 2011). The focus in MAB has been on the maximization of cumulative rewards by sequentially choosing from a set of options —referred to as "arm"— based on stochastic feedback. A conceptually distinct but equally fundamental variant within this framework is the best arm identification (BAI) problem, where the objective is not reward maximization over time, but rather the accurate identification of the optimal arm(s) using as few samples as possible. Because only the final decision matters, BAI operates under a pure exploration regime. This leads to unique algorithmic and theoretical challenges not encountered in classical reward maximization.

BAI problems have been studied primarily under two canonical settings: (i) the fixed-confidence setting, where the goal is to guarantee the correctness of the identified arm(s) with high probability (at least $1 - \delta$ for any $\delta \in (0, 1)$); and (ii) the fixed-budget setting, in which a learner is restricted to a fixed number of samples T and must maximize the probability of correct identification. Foundational work in the fixed-confidence setting introduced PAC-style guarantees Even-Dar et al. (2002), later refined through approaches such as LUCB Kalyanakrishnan et al. (2012) achieving tighter bounds on sample complexity by leveraging confidence intervals. In the fixed-budget setting, algorithms such as Successive Rejects (SR) Audibert and Bubeck (2010) and Sequential Halving (SH) Karnin et al. (2013) eliminate suboptimal arms based on empirical ranking procedures. Despite their algorithmic differences, both settings share underlying complexity measures, such as the problem-dependent hardness parameter, often defined as the sum of inverse squared gaps between the best and suboptimal arms.

Recognizing the algorithmic parallels between the fixed-confidence and fixed-budget settings, Gabillon et al. (2012) introduced the Unified Gap-based Exploration (UGapE) algorithm, which provides a single arm selection strategy applicable to both settings. This work laid the foundation for the unified algorithm design across different settings and emphasized the role of gap-based strategies in pure exploration.

More recently, increasing attention has been paid to *risk-aware* variants of BAI, motivated by applications in medical trials or finance, where expected reward alone is insufficient Huo and Fu (2017); Tamkin et al. (2019); Keramati et al. (2020); Du et al. (2021); Chen et al. (2022); Shen et al. (2022). In such variants, measure of variability, such as *variance*, *tail risk*, or *quantiles*, must be taken into account in the decision process. For example, Hou et al. (2022) proposed the Variance-Aware (VA)-LUCB algorithm, which aims to identify the arm with the highest mean subject to a strict upper bound on variance. Their approach introduces a variance-aware hardness measure and shows nearly optimal sample complexity. Other approaches have explored alternative risk criteria such as Conditional Value-at-Risk (CVaR) and quantiles David and Shimkin (2016).

Parallel to this, much attention has been gained to incorporation with multi-objective BAI problems, where arms are evaluated across multiple criteria. Under the *Pareto Set Identification* (PSI) framework, the goal is to identify the set of non-dominated arms (=Pareto-optimal arms) that are not outperformed across all objectives by any other. Early PSI algorithms such as the confidence-bound-based method proposed by Auer et al. (2016) were developed under fixed-confidence settings, using uniform sampling and acceptance-rejection schemes. More recently, the adaptive LUCB-like algorithm for PSI Kone et al. (2023) has improved the sample efficiency by exploiting gap information.

Despite this progress, PSI under fixed-budget constraints has remained comparatively underexplored until recently. Kone et al. (2024) introduced the *Empirical Gap Elimination (EGE)* framework, which generalizes SR and SH to the multi-objective setting. EGE estimates empirical gaps to eliminate arms and classifies them as Pareto-optimal or suboptimal. The EGE-based algorithms, EGE-SR and EGE-SH, were found to achieve exponential decay in error probability with respect to budget and are near-optimal according to an information-theoretic lower bound. Despite recent progresses, some limitations yet remain in the multi-objective BAI literature. Foremost among these is the limitation on both theoretical unification and practical applicability across both fixed-confidence, and fixed-budget canonical settings: most existing algorithms are tailored specifically to either of the two settings. In parallel, although recent efforts have been devoted in introducing risk-awareness into BAI —such as VA-LUCB under variance constraints— these approaches typically handle risk in isolation, without integrating it into multi-objective frameworks like Pareto Set Identification (PSI) Ulrich et al. (2008); Kone et al. (2025). Moreover, existing PSI algorithms either ignore uncertainty (risk) altogether or treat it as an independent objective, lacking a principled scheme to jointly evaluate utility and risk in arm selection.

To address these challenges, we propose a novel multi-objective optimization framework to take into account both mean and risk simultaneously that bridges the fixed-confidence and fixed-budget paradigms through a unified arm selection strategy, modulated only by confidence intervals and setting-specific stopping rules, we call RAMGapE (Risk-Averse Multi-objective Gap-based Exploration). Central to RAMGapE is a new gap-based criterion that incorporates both the expected reward and the associated risk, quantified through a mean-variance trade-off. This allows for efficient identification of ϵ-Pareto optimal arms while explicitly accounting for risk. Our theoretical analysis provides guarantees on correctness and sample complexity, while extensive experiments demonstrate that RAMGapE significantly outperforms existing methods in risk-sensitive settings—achieving superior decision quality with fewer samples. Specifically, the novelty of our approach lies in overcoming the limitations of prior studies through a new theoretical framework, summarized in three key aspects:

1 Adaptation of Gap-Based Analysis to Partial Orders: Unlike traditional methods that often assume a total order of arms (e.g., ranked by their mean rewards), our algorithm extends gap-based analysis to a partial-order setting defined by mean-risk Pareto dominance.

2 Handling of Variable-Size Pareto Sets: We remove the common assumption of a fixed number of optimal arms or a pre-defined boundary arm for comparison. RAMGapE is designed to identify the entire Pareto set whose size is unknown a priori and can vary, making it applicable to a wider range of real-world problems.

3 A Novel Exploration Rule Targeting Pareto Dominance: We introduce a new exploration strategy that explicitly targets the structure of Pareto dominance. Instead of focusing on a single best arm, the algorithm efficiently allocates samples to resolve uncertainties along the Pareto frontier, pruning provably suboptimal arms and identifying the set of non-dominated solutions.

To our knowledge, RAMGapE is the first algorithm to present provable guarantees, having the above features, for multi-objective, risk-averse PSI whose two objective variables are dependent to each other via the same reward distribution in both fixed-confidence and fixed-budget frameworks.

2. Problem Setting

In this section, we introduce the definitions and notation used throughout this paper. Let $[K] = \{1, 2, \ldots, K\}$ be the set of arms such that each arm $i \in [K]$ is characterized by a reward distribution ν_i bounded in $[0, 1]$ with mean μ_i and variance σ_i^2. Here we employ a risk criteria based on *mean-variance* (MV) Sani et al. (2012) as the second co-equal objective. The smaller the value of MV, the lower the risk of the arm. Here, ρ (≥ 0) is a hyperparameter to control the weight of its risk in the search, that is, when $\rho \to \infty$, the minimization of MV corresponds to finding arm(s) with a larger mean(s) without taking care of its variance, while it corresponds to finding arm(s) with smaller variance(s) when $\rho \to 0$ (ρ has the dimension of mean μ_i). We introduce a parameter α (> 0) to scale the MV measure as $\xi_i := \alpha \mathrm{MV}_i$ for each arm i, which preserves the relative position of the risk measure. Later, we set $\alpha = \frac{1}{3+\rho}$ to allow the construction of confidence intervals with equal widths for both mean and risk measures. In classical formulation, variance captures reward uncertainty. However, when identifying Pareto-optimal solutions over mean and variance as risk measure, arms with very low mean but low variance can still be deemed Pareto-optimal, even though they are of little practical importance. Using mean-variance (MV) criterion with adjusting the parameter ρ, we can design the identification of Pareto solutions that have both relatively high mean and low risk. However, due to its second-order moment property, MV accounts for the "risk" symmetrically. When the underlying reward distribution is skewed and higher-order moments are non-zero, MV may not adequately capture the severity of rare events (tail risk). Among other risk measures, for example, Conditional Value-at-Risk (CVaR) Rockafellar et al. (2000) is another possible choice of risk measure. CVaR quantifies the expected loss in the worst-case scenarios. The challenge of extending our gap-based framework to asymmetric, tail-focused risk measures like CVaR remains one of the forthcoming subjects to be resolved, as discussed in Section 5.

Next, we address Pareto optimality when the two stochastic variables μ_i and ξ_i are used as objective criteria.

We say that arm j *strictly dominates* arm i, denoted as $j \succ i$, if both $\mu_j > \mu_i$ and $\xi_j < \xi_i$ hold; that is, arm j has a strictly higher expected reward and strictly lower risk than those of arm i. An arm $i \in [K]$ is said to be *Pareto optimal* if there exists no other arm $j \in [K]$ such that $j \succ i$. In other words, arm i is not strictly dominated by any other arm(s). We denote by D^+ the set of all arms that satisfy this Pareto optimality condition. For each arm $i \in [K]$, according to Kone et al. (2024), we define a gap Δ_i as

$$\Delta_i := \begin{cases} \min\left\{ \min\limits_{j \in D^+ \setminus \{i\}} \Big(\min(M(i,j), M(j,i)) \Big), \min\limits_{j \notin D^+} \left(M(j,i)^+ + \Delta_j \right) \right\} & \text{if } i \in D^+; \\ \max\limits_{j \in D^+ \text{ s.t. } j \succ i} m(i,j) & \text{if } i \notin D^+, \end{cases} \tag{1}$$

where $m(i,j) := \min(\mu_j - \mu_i, \xi_i - \xi_j)$, $M(i,j) := \max(\mu_i - \mu_j, \xi_j - \xi_i)$, $M(j,i)^+ := \max(M(j,i), 0)$.

The definition of gap tells us that for $i \in D^+$, Δ_i properly quantifies how well arm i separates itself from other arms, capturing both the minimal margin from non-Pareto arms and the proximity to other Pareto-optimal arms, and for $i \notin D^+$, Δ_i does to what degree the non-Pareto optimal arm i is dominated by the other arms at most. We illustrate these quantities and explain the details in Appendix A. Given an allowance $\epsilon > 0$ defined by a user, a subset $S \subseteq [K]$ is called a ϵ-Pareto set if it satisfies the following condition,

$$\forall i \in S, \forall j \in [K], \mu_i > \mu_j - \epsilon \vee \xi_i < \xi_j + \epsilon,$$
$$\forall i \notin S, \exists j \in [K], \mu_i \le \mu_j - \epsilon \wedge \xi_i \ge \xi_j + \epsilon.$$

Hereinafter, we formulate the Risk-Averse Best Arm Set Identification Problem as the problem to find an ϵ-Pareto set for expected means and their risks. Note that ϵ can be regarded as a resolution associated with the observation in question and $\epsilon \to 0$ converges to the problem without any error in measurement.

The Risk-Averse Best Arm Set Identification Problem can be formalized as a process between a stochastic bandit environment and a forecaster. The reward distributions $\{\nu_i\}_{i=1}^K$ inherent to each arm are unknown a priori to the forecaster. At each round t, the forecaster pulls an arm $I(t) \in [K]$ and observes a sample independently drawn from the identical distribution $\nu_{I(t)}$. Let $T_i(t)$ be the number of times that arm i has been pulled up to the round t, the forecaster estimates the expected value of mean, variance, and risk of this arm by $\hat{\mu}_i(t) = \frac{1}{T_i(t)} \sum_{s=1}^{T_i(t)} X_i(s)$, $\hat{\sigma}_i^2(t) = \hat{\mu}_i^{(2)}(t) - \hat{\mu}_i^2(t)$, and $\hat{\xi}_i(t) = \alpha(\hat{\sigma}_i^2(t) - \rho\hat{\mu}_i(t))$, where $X_i(s)$ and $\hat{\mu}_i^{(2)}(t)$ are the s-th sample observed from ν_i and $\frac{1}{T_i(t)} \sum_{s=1}^{T_i(t)} X_i^2(s)$, respectively. For any set $S \subseteq [K]$ and arm $i \in [K]$, we introduce the notations of arm simple regret $r_i(S)$ for arm i as well as (set) simple regret r_S for set of arms S as follows:

$$r_i(S) = \begin{cases} \Delta_i & \text{if } i \in S \triangle D^+ \\ 0 & \text{otherwise} \end{cases}, \tag{2}$$

$$r_S = \max_{i \in [K]} r_i(S) \tag{3}$$

where $A \triangle B := (A \setminus B) \cup (B \setminus A)$ for any sets A and B.

We define a temporary set of Pareto arms with respect to the empirical values at round t as

$$\widehat{D}_t^+ := \{i \in [K] \mid \forall j \in [K],\ j \not\succ_t i\},$$

where the empirical (strict) dominance relation $\succ_t$ is defined as follows:

Definition 1 (Empirical Dominance Relation) *For any two arms $i, j \in [K]$ in round t, we say that arm j strictly dominates arm i at round t, denoted by $j \succ_t i$, if*

$$\hat{\mu}_j(t) > \hat{\mu}_i(t) \quad \text{and} \quad \hat{\xi}_j(t) < \hat{\xi}_i(t).$$

In other words, arm j is better than arm i in both mean and risk estimates at round t. We denote $j \not\succ_t i$ if this condition does not hold.

The simple regret in each round t can be written as $r_{\widehat{D}_t^+}$. Returning an ϵ-Pareto set is then equivalent to having $r_{\widehat{D}_t^+}$ smaller than ϵ. Given an allowance ϵ, we formalize the two settings of fixed budget and fixed confidence.

Fixed budget. The objective is to return the set of ϵ-Pareto arms with the highest possible confidence level using a fixed budget of n rounds. Formally, given a budget n, the performance of the forecaster is measured by the probability $\widetilde{\delta}$ of not satisfying the conditions of the set of ϵ-Pareto arms, i.e., $\widetilde{\delta} = \mathbb{P}\left[r_{\widehat{D}_n^+} \geq \epsilon\right]$, the smaller $\widetilde{\delta}$, the better the algorithm.

Fixed confidence. The objective is to design a forecaster that stops as soon as possible and returns the set of ϵ-Pareto arms with fixed confidence. Let $\widetilde{n}$ be the round at which the

Algorithm 1: PullArm

```
Input: $t$, $\{T_i(t)\}_{i=1}^K$, $\{\beta_i(t)\}_{i=1}^K$, $\{\hat{\mu}_i(t)\}_{i=1}^K$, $\{\hat{\xi}_i(t)\}_{i=1}^K$
if $\exists i$ such that $T_i(t) \leq 2$ then
    return $\arg\min_{i \in [K]} T_i(t)$
end
Compute $V_i(t)$ for each arm $i \in [K]$
Determine $m_t$ and $p_t$ using Eq. 6 and Eq. 7
return $\arg\max_{i \in \{m_t, p_t\}} \beta_i(t)$
```

algorithm stops, and let $\widehat{D}_{\widetilde{n}}^{+}$ be the set of arms returned. Given a fixed confidence δ, the forecaster must guarantee that $\mathbb{P}\left[r_{\widehat{D}_{\widetilde{n}}^{+}} \geq \epsilon\right] \leq \delta$. The forecaster performance is evaluated at the stopping round $\widetilde{n}$.

Although traditionally treated as distinct problems, in Section 3 we present a unified arm selection strategy that applies to both settings, differing only in the choice of stopping criterion.

3. Risk-Averse Multi-objective Gap-based Exploration Algorithm

In this section, we present the risk-averse gap-based exploration algorithm (RAMGapE) meta-algorithm, involving its implementation for fixed budget and fixed confidence settings, named RAMGapEb and RAMGapEc, respectively. The algorithm in each setting uses a common arm selection strategy, PullArm (Algorithm 1) (see also the pseudo-code Algorithm 2). RAMGapEb and RAMGapEc return an ϵ-Pareto set using the same definition of temporal Pareto set $\widehat{D}$. They only differ in the stopping rule. Given an allowance ϵ, both algorithms first suppose constant parameters such as the budget n and the hyperparameter a that controls the exploration rate RAMGapEb, the confidence level δ in RAMGapEc, respectively. RAMGapEb runs for n rounds and returns a set of arms $\widehat{D}_{n}^{+}$, whereas RAMGapEc runs until it achieves the required confidence level δ so that the probability of correctly extracting the Pareto optimal set is greater than $1 - \delta$ under the given allowance ϵ. The difference is caused by the different objectives of the two algorithms: RAMGapEb aims to maximize the quality of prediction under the fixed budget but RAMGapEc aims to minimize budget required to accomplish the given fixed confidence level.

To initialize variance estimation, each arm is first pulled twice before the adaptive exploration begins. This initialization step guarantees that variance estimates are properly defined when computing the risk-based criteria used throughout the algorithm. In PullArm (Algorithm 1), at each round t and for each arm $i \in [K]$, RAMGapE first uses the information observed up to the round t and computes quantities $V_i(t), V(t), m_t, p_t$, and $I(t)$ that

Algorithm 2: RAMGapE

```
Input: $K$, $a$, $n$, $\epsilon$, $\rho$
Initialize $T_i(1) \leftarrow 0$, $\beta_i(1) \leftarrow 0$, $\hat{\mu}_i(1) \leftarrow 0$, $\hat{\xi}_i(1) \leftarrow 0$ for $i = 1, 2, \dots, K$
Set $t \leftarrow 1$
while $t \leq n$ do
    $I(t) \leftarrow \mathrm{PullArm}\Big(t, \{T_i(t)\}, \{\beta_i(t)\}, \{\hat{\mu}_i(t)\}, \{\hat{\xi}_i(t)\}\Big)$
    Observe $X_{I(t)}(T_{I(t)}(t) + 1) \sim \nu_{I(t)}$
    $t \leftarrow t + 1$
    Update $\hat{\mu}_{I(t)}(t)$, $\hat{\xi}_{I(t)}(t)$, $\beta_{I(t)}(t)$, and $T_{I(t)}(t)$
    ; // (RAMGapEb)
    if $t > n$ then
        break
    end
    ; // (RAMGapEc)
    if $t > 2K \wedge V(t) < \epsilon$ then
        break
    end
end
return $\widehat{D}_n^+$
```

are defined by

$$V_i(t) := \begin{cases} \max\limits_{j \neq i} \min\Big(\overline{\mu}_j(t) - \underline{\mu}_i(t), \overline{\xi}_i(t) - \underline{\xi}_j(t)\Big) & \text{if } i \in \widehat{D}_t^+ \\ \min\limits_{j \in \widehat{D}_t^+ \text{ s.t. } j \succ_t i} \max\Big(\overline{\mu}_i(t) - \underline{\mu}_j(t), \overline{\xi}_j(t) - \underline{\xi}_i(t)\Big) & \text{if } i \notin \widehat{D}_t^+ \end{cases}, \tag{4}$$

$$V(t) := \max_{i \in [K]} V_i(t), \tag{5}$$

$$m_t := \operatorname*{argmax}_{i \in [K]} V_i(t) \tag{6}$$

$$p_t := \begin{cases} \operatorname*{argmax}\limits_{j \neq m_t} \min\Big(\overline{\mu}_j(t) - \underline{\mu}_{m_t}(t), \overline{\xi}_{m_t}(t) - \underline{\xi}_j(t)\Big) & \text{if } m_t \in \widehat{D}_t^+ \\ \operatorname*{argmin}\limits_{j \in \widehat{D}_t^+ \text{ s.t. } j \succ_t m_t} \max\Big(\overline{\mu}_{m_t}(t) - \underline{\mu}_j(t), \overline{\xi}_j(t) - \underline{\xi}_{m_t}(t)\Big) & \text{if } m_t \notin \widehat{D}_t^+ \end{cases}. \tag{7}$$

Here $\overline{\mu}_i(t), \underline{\mu}_i(t), \overline{\xi}_i(t)$ and $\underline{\xi}_i(t)$ represents the upper and lower bounds of the mean (μ_i) and risk (ξ_i) of the arm i after t rounds, respectively. In brief, $V_i(t)$ estimates the maximum gap of arm i from the rest by comparing the pessimistic predictions of μ_i and ξ_i and the optimistic predictions of those of the other arms when i belongs to the temporal Pareto set $\widehat{D}_t^+$ defined by the sample mean and sample risk at round t. Likewise, when i does not belong to $\widehat{D}_t^+$, it estimates the minimum gap between (μ_i, ξ_i) of the arm i and those of the arms belonging to $\widehat{D}_t^+$ to dominate the arm i. These quantities are defined by

$$\forall i \in [K], \forall t, \begin{cases} \overline{\mu}_i(t) := \hat{\mu}_i(t) + \beta_i(t) \\ \underline{\mu}_i(t) := \hat{\mu}_i(t) - \beta_i(t) \end{cases}, \begin{cases} \overline{\xi}_i(t) := \hat{\xi}_i(t) + \beta_i(t) \\ \underline{\xi}_i(t) := \hat{\xi}_i(t) - \beta_i(t) \end{cases}, \tag{8}$$

where $\beta_i(t)$ denotes their confidence intervals and a parameter denoted by a was employed in the definition of β_i, whose shape strictly depends on the concentration bound used by the algorithm. For example, we can derive β_i from the Hoeffding-Azuma inequality Azuma (1967); Tropp (2012) as

$$\begin{aligned} \text{RAMGapEb: } \beta_i(t) &= \sqrt{\tfrac{a}{T_i(t)}}, \\ \text{RAMGapEc: } \beta_i(t) &= \sqrt{\tfrac{4}{T_i(t)} \ln \tfrac{8K(\log_2 T_i(t))^2}{\delta}}. \end{aligned} \tag{9}$$

We introduce a quantity $V_S(t)$ for a set S as $V_S(t) := \max_{i \in S} V_i(t)$. After computing the quantities for all arms, RAMGapE selects two key arms: m_t, the arm with the largest $V_i(t)$, and p_t, the most relevant comparison arm to m_t based on the dominance relation. Depending on whether m_t and p_t are included in the currently estimated temporal Pareto set $\widehat{D}_t^+$, these arms may represent potentially optimal or suboptimal candidates. If both are in $\widehat{D}_t^+$, they are regarded as highly uncertain arms that could be Pareto optimal. RAMGapE then selects the arm with a fewer number of pulls between m_t and p_t, thereby prioritizing exploration toward the arm with greater uncertainty. The algorithm pulls the selected arm, observes its reward, and updates its empirical mean $\hat{\mu}_i(t)$, risk estimate $\hat{\xi}_i(t)$, and pull count $T_i(t)$.

The core mechanism of RAMGapE is distribution-agnostic, requiring only valid confidence intervals, $\beta_i(t)$. While we employ Hoeffding-based bounds for Beta distributions, this choice can be adapted. For instance, bounds for sub-Gaussian variables are applicable for unbounded rewards; we validate this empirically in Appendix E.1. For heavy-tailed rewards, robust estimators like truncated means Bubeck et al. (2013) or Catoni's M-estimator Catoni (2012) can be readily integrated to ensure valid concentration bounds without altering the algorithm's main structure. This modularity confirms the broad applicability of RAMGapE.

Theoretical Guarantees

We provide theoretical guarantees for RAMGapE under both the fixed-confidence and fixed-budget settings. The core of our analysis relies on establishing a high-probability event $\mathcal{E}$ (see Eq. 10 in Appendix B) where all empirical estimates remain within their confidence intervals. Under this event, we can guarantee the algorithm's performance. The detailed proofs for the following theorems are given in Appendix B.

Theorem 2 *(Error Bound for Fixed-Budget) If we run RAMGapEb with parameter* $0 < a \leq \frac{n-2K}{16K}\epsilon^2$ *for rounds* n*, total number of arms* K *and allowance rate* ϵ*, its simple regret* $r_{\widehat{D}_n^+}$ *satisfies*

$$\widetilde{\delta} = \mathbb{P}\left[r_{\widehat{D}_n} \geq \epsilon\right] \leq 4Kn\exp(-2a),$$

and, in particular, this probability is minimized at $a = \frac{n-2K}{16K}\epsilon^2$.

Theorem 3 *(Correctness and Termination for Fixed-Confidence) The RAMGapEc algorithm stops after* $\widetilde{n}$ *rounds and returns an* ϵ*-Pareto set,* $\widehat{D}_{\widetilde{n}}^+$*, that satisfies*

$$\mathbb{P}\left[r_{\widehat{D}_{\widetilde{n}}^+} \leq \epsilon \wedge \widetilde{n} \leq N\right] \geq 1 - \delta,$$

where $N = 2K + \mathcal{O}\left(\frac{K}{\epsilon^2}\log\left(\frac{K\log_2^2(1/\epsilon)}{\delta}\right)\right)$ *with confidence level* δ (< 1).

4. Experiments

In this section, we evaluate the performance of the proposed algorithm RAMGapE under both the fixed-confidence and fixed-budget settings. We compare the performance with those of other algorithms, including the standard Round-Robin strategy and several previously proposed approaches for risk-averse and multi-objective bandit problems.

Fixed-Confidence Setting

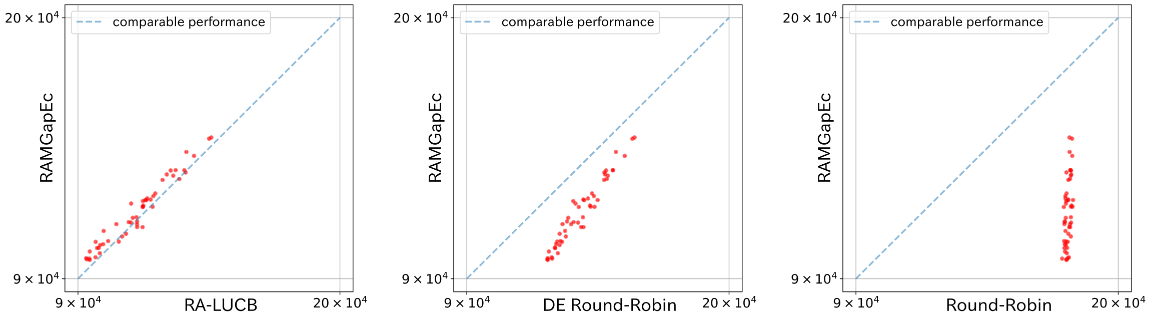
In the fixed-confidence setting, we compare RAMGapE with the following three representative algorithms. Round-Robin uniformly samples each arm and serves as a fundamental baseline. Dominated Elimination Round-Robin (DE Round-Robin) (also used as a baseline in Kone et al. (2024)) improves upon this by eliminating empirically dominated arms based on observed values. Risk-Averse LUCB (RA-LUCB) extends the classical LUCB algorithm Kalyanakrishnan et al. (2012) to risk-sensitive settings, where it selects and pulls two arms—denoted m_t and p_t—in each round (see also pseudo-codes in Appendix D). In contrast, RAMGapE differs from RA-LUCB in that it pulls only the less frequently sampled of the two arms m_t and p_t. This leads to improved sample efficiency while maintaining identification accuracy, and this selection rule forms the main distinction between the two algorithms.

Experiment 1 (Comparison of Stopping Time):
We compare the number of rounds required by each method to meet the stopping condition in 50 problem instances. The reward for each arm follows a Beta distribution, with means in $[0.4, 0.6]$ and variances in $[0.01, 0.2]$, and the number of arms is set to $K = 10$ (see Table 3). The algorithmic parameters are fixed at $(\delta, \epsilon, \rho) = (0.05, 0.1, 0.01)$. Please see also the similar experiment with $\epsilon = 0.05$ in Fig.7 (Appendix E.2).

Experiment 2 (Comparison of Confidence Intervals at Stopping Time): Using the same set of problem instances as in Experiment 1, we compare the width of confidence intervals at the stopping point for each algorithm. The tolerance parameter is set to $\epsilon = 0$, enabling us to assess how conservative or aggressive each method is in its stopping criterion. We consider two settings: **Experiment 2.1** corresponds to instances where the number of Pareto-optimal arms is small (about arms set, see Table 3, pattern 10), while **Experiment 2.2** targets instances where the number of Pareto-optimal arms is large (about arms set, see Table 3, pattern 46). This allows us to evaluate the behavior of the algorithms under different levels of Pareto set complexity.

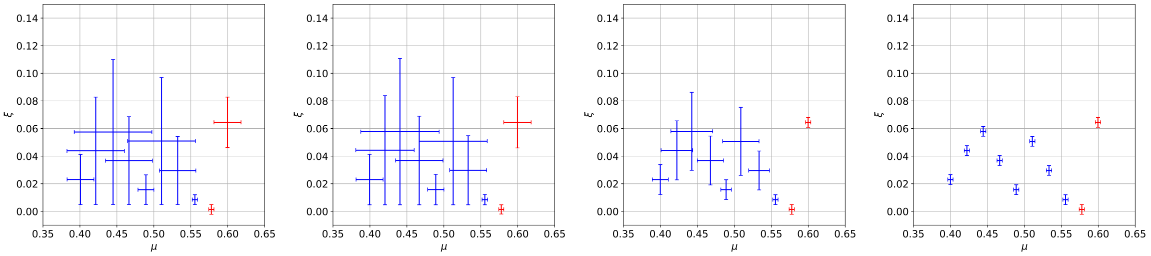
Fixed-Budget Setting

In the fixed-budget setting, we compare RAMGapE with several algorithms: the standard Round-Robin, Least-Important Elimination Round-Robin (LIE Round-Robin), RA-LUCB adapted for the fixed-budget case, the risk-sensitive ξ-LCB (used as a baseline in Sani et al. (2012)), hypervolume-based HVI-Pareto method (see e.g., Yang et al. (2019); Zitzler et al. (2007); Cao et al. (2015) for the definition and applications of hypervolume), and



(a) vs RA-LUCB (b) vs DE Round-Robin (c) vs Round-Robin

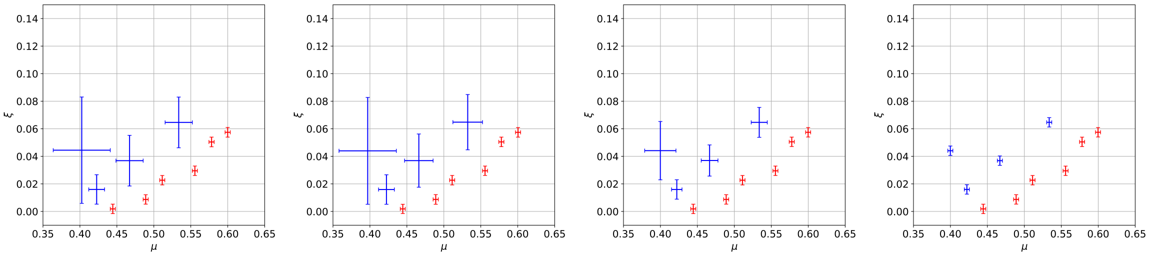
Figure 1: **Stopping Time Comparison of Experiment 1 with $\epsilon = 0.1$.** The blue dashed line corresponds to the identity line, i.e., the set of points where both methods terminate at the same time, indicating comparable performance. Points located below this line signify that the proposed method stops earlier than the baseline.



(a) RAMGapEc (b) RA-LUCB (c) DE Round-Robin (d) Round-Robin

Figure 2: **Visualization of confidence intervals at stopping time (Experiment 2.1).** Each panel shows the empirical mean (horizontal axis) and scaled risk (vertical axis: $\xi = \alpha(\sigma^2 - \rho\mu)$) of each arm at the termination round for different algorithms. The crosses represent confidence intervals of each arm; the longer the arms of the crossed interval, the fewer the samples allocated to that arm. Red points (=crosses) indicate arms included in the returned set $\widehat{D}_t^+$, while blue points indicate excluded arms. RAMGapEc and RA-LUCB not only avoid over-sampling non-Pareto arms (shown in blue), but also limit sampling for some arms included in $\widehat{D}_t^+$, particularly those located on the far right of the plot (i.e., arms with high mean but less impact on Pareto set boundaries). These arms exhibit wider confidence intervals, reflecting lower sample counts. This behavior highlights the algorithms' ability to allocate samples efficiently, gathering just enough information for confident identification without unnecessary exploration. The total sample counts of these examples are: RAMGapEc: 9,697,292; RA-LUCB: 9,728,010; DE Round-Robin: 15,283,296; Round-Robin: 43,548,822.

the Empirical Gap-based Pareto Set Exploration (EGP) (see the relevance in Kone et al. (2024)) (see the pseudo-codes of these comparison algorithms in Appendix D).



(*a*) RAMGapEc (*b*) RA-LUCB (*c*) DE Round-Robin (*d*) Round-Robin

Figure 3: **Visualization of confidence intervals at stopping time (Experiment 2.2).** The meanings of the crosses, and the colors are the same as in Fig. 2. The total number of samples used for these examples are: RAMGapEc: 28,261,200; RA-LUCB: 28,486,332; DE Round-Robin: 30,041,447; Round-Robin: 46,905,293.

It should be noted that Round-Robin-based evaluation strategies have traditionally been used in domains such as medicine, where repeated sampling and fair treatment allocation are a common strategy (see, for example, Pannee et al. (2016); Endris et al. (2016)).

Experiment 3 (Comparison of Average Simple Regret with a Small Number of Arms):
We evaluate average simple regret for each method under $K = 10$ arms (see Table 1). Each arm's reward distribution is a Beta distribution with randomly sampled means and variances from $[0.4, 0.6]$ and $[0.01, 0.2]$, respectively. The parameter a is set as $\frac{n-2K}{16K}\epsilon^2$ (see Appendix B.2). All algorithms are executed for 50 independent trials with $T = 10,000$ rounds. In order to reduce the influence of outliers, the lower and upper 25% of the simple regret values at each time round are excluded and the remaining middle 50% of the simple regret values are averaged and used for the evaluation of the algorithm performance.

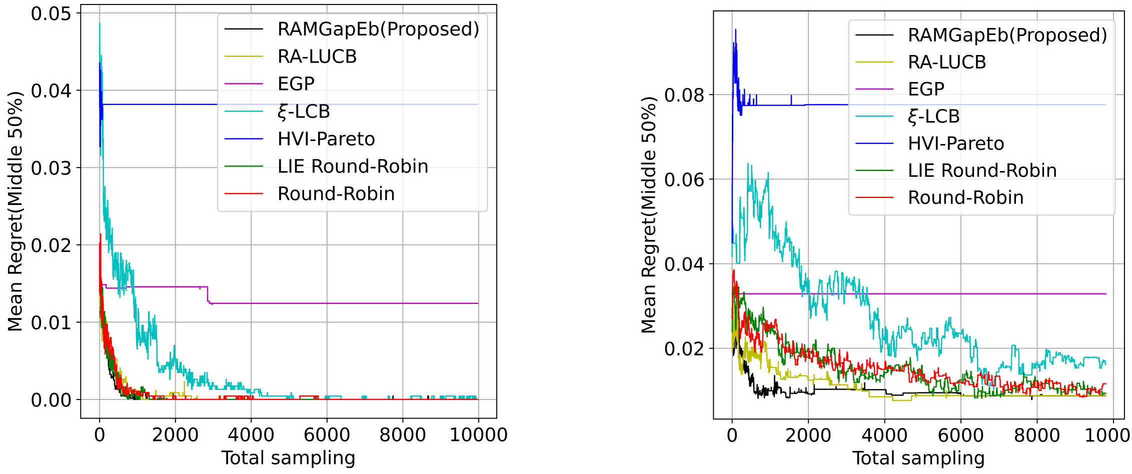
Experiment 4 (Comparison of Average Simple Regret with a Large Number of Arms):
To assess scalability, we increase the number of arms to $K = 100$ (see Table 2), while keeping the same settings as in Experiment 3.

In all experiments, we set the risk coefficient to $\alpha = \frac{1}{3+\rho}$, ensuring that the widths of confidence intervals for both the mean and the risk metric are balanced (see Appendix B.4.2). The use of Beta distributions allows us to model a variety of shapes—unimodal, U-shaped, monotonic, and uniform—making the evaluation more reflective of real-world scenarios. Further implementation details and algorithmic formulations are provided in Appendix D.

4.1. Results

Evaluation under the Fixed-Confidence Setting

We evaluated the performance of RAMGapE under the fixed-confidence setting by comparing with several baseline algorithms, including Round-Robin, Dominated Elimination



(a) Experiment 3 with $K = 10$ arms (b) Experiment 4 with $K = 100$ arms

Figure 4: **Comparison of average simple regret (middle 50%) over total number of samples.**

Round-Robin (DE Round-Robin), and Risk-Averse LUCB (RA-LUCB). As shown in Fig. 1, in Experiment 1, the stopping times of RAMGapEc significantly shorter than those of Round-Robin and DE Round-Robin, and those of RAMGapEc and RA-LUCB are comparable with marginal differences (see also Fig. 6 in Appendix E.1). The visualization of confidence intervals at stopping time (Experiments 2.1 and 2.2, Figs. 2-3) further demonstrates that RAMGapEc, as well as RA-LUCB, effectively avoids unnecessary exploration of non-Pareto arms, focusing sampling efforts on arms near the Pareto frontier.

Evaluation under the Fixed-Budget Setting

In the fixed-budget setting, as shown in Fig. 4, RAMGapE was compared with other existing approaches, including Round-Robin, LIE Round-Robin, RA-LUCB, ξ-LCB, EGP, and HVI-Pareto. The results of Experiments 3 ($K = 10$) and 4 ($K = 100$) show that RAMGapE exhibits a fast convergence of average simple regret both in small- and large-scale problems. Especially for the $K = 100$ problem, RAMGapE exhibits the fastest drop in the mean regret with respect to the total sampling much faster than the comparable algorithm RA-LUCB in Experiments 1 and 2 in this experiment. Note also that, except RAMGapE, the other algorithms either converge very slowly or converge to some mean regrets larger than that acquired by RAMGapE. This suggests that our RAMGapE not only unifies the fixed budget and fixed confidence settings with different stopping criterion but also outperforms or equally best performs among the comparison algorithms for both settings.

Moreover, an analysis of the pulling ratios reveals how RAMGapE efficiently allocates samples. As shown in Figs. 8 and 9 (Appendix E.3), rather than naively focusing only on empirically optimal arms, RAMGapE strategically balances exploration between Pareto and non-Pareto arms to precisely identify the boundary of the optimal set. Once suboptimal arms are identified with sufficient confidence, the algorithm adaptively shifts its focus, resulting in a higher proportion of samples being allocated to promising, Pareto-optimal arms in later stages. This efficient exploration strategy contributes directly to its steady reduction of regret and the accurate identification of the Pareto set, as shown in Fig. 4.

Overall Assessment

Overall, RAMGapE demonstrated stable performance across both fixed-confidence and fixed-budget settings, efficiently balancing exploration and exploitation. The results highlight its suitability for risk-averse decision-making in stochastic environments where identifying multiple viable solutions is required.

5. Conclusion

We presented RAMGapE (Risk-Averse Multi-objective Gap-based Exploration), a unified algorithmic framework for Risk-Averse Best Arm Set Identification problem that jointly optimizes both expected reward and risk via the *mean-variance* (MV) criterion. Unlike conventional approaches that treat risk as an isolated objective, RAMGapE integrates risk directly into the multi-objective formulation, enabling principled identification of Pareto-optimal solutions that simultaneously balance utility and uncertainty. We provided theoretical guarantees, including correctness and sample complexity bounds, and demonstrated that RAMGapE achieves efficient sampling and accurate identification of Pareto-optimal solutions. Our results show that RAMGapE adaptively concentrates sampling on uncertain regions near the Pareto frontier, while efficiently pruning non-Pareto arms far from Pareto fronts. This targeted exploration yields robust performance across both small- and large-scale problem instances. A key strength of RAMGapE lies in its ability to flexibly allocate sampling resources toward high-uncertainty regions near the Pareto frontier, making it well-suited for real-world risk-sensitive applications such as medical trials or portfolio optimization, where both performance and risk must be jointly optimized.

Future work includes several directions. A key avenue is to extend the present framework to incorporate richer and more complex risk measures beyond mean-variance measure. For instance, adapting RAMGapE to spectral risk measures like CVaR or EVaR Ahmadi-Javid (2012) is a non-trivial but important challenge. This would require redefining the gap quantities in a way that accommodates these tail-focused risk measures and deriving new concentration bounds for their empirical estimators. The partial-order structure induced by such measures may also differ significantly from the one in the mean-variance space, demanding a careful redesign of the exploration strategy. Further theoretical challenges include addressing non-stationary environments and refining the analysis to derive tighter sample complexity bounds. Overall, RAMGapE advances, with its unified formulation and significant performance, the state of risk-aware multi-objective bandit problem, providing a solid foundation for tackling complex, real-world decision-making problems under uncertainty.

Acknowledgments

The authors express their sincere gratitude to Professors Atsuyoshi Nakamura and Hiroshi Teramoto for their invaluable guidance and insightful discussions throughout this research. This work was supported by JST/CREST Innovative Measurement and Analysis (Grant

Number JPMJCR2333 to TK), JSPS Program for Forming Japan's Peak Research Universities (J-PEAKS) Grant Number JPJS00420230001, JSPS Grant in Aid for Scientific Research (A) (General) (Grant Number 24H00685 to KT) and JST SPRING (Grant Number JPMJSP2119 to SN).

References

Amir Ahmadi-Javid. Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*, 155(3):1105–1123, 2012.

Jean-Yves Audibert and Sébastien Bubeck. Best arm identification in multi-armed bandits. In *COLT-23th Conference on learning theory-2010*, pages 41–53, 2010.

Peter Auer, Chao-Kai Chiang, Ronald Ortner, and Madalina Drugan. Pareto front identification from stochastic bandit feedback. In *Artificial intelligence and statistics*, pages 939–947. PMLR, 2016.

Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

Yongtao Cao, Byran J Smucker, and Timothy J Robinson. On using the hypervolume indicator to compare pareto fronts: Applications to multi-criteria optimal experimental design. *Journal of Statistical Planning and Inference*, 160:60–74, 2015.

Olivier Catoni. Challenging the empirical mean and empirical variance: a deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 1148–1185, 2012.

Tianrui Chen, Aditya Gangrade, and Venkatesh Saligrama. Strategies for safe multi-armed bandits with logarithmic regret and risk. In *International Conference on Machine Learning*, pages 3123–3148. PMLR, 2022.

Yahel David and Nahum Shimkin. Pure exploration for max-quantile bandits. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 556–571. Springer, 2016.

Yihan Du, Siwei Wang, Zhixuan Fang, and Longbo Huang. Continuous mean-covariance bandits. *Advances in Neural Information Processing Systems*, 34:875–886, 2021.

Volker Endris, Albrecht Stenzinger, Nicole Pfarr, Roland Penzel, Markus Möbs, Dido Lenze, Silvia Darb-Esfahani, Michael Hummel, Andreas Jung, Ulrich Lehmann, et al. Ngs-based brca1/2 mutation testing of high-grade serous ovarian cancer tissue: results and conclusions of the first international round robin trial. *Virchows Archiv*, 468:697–705, 2016.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Pac bounds for multi-armed bandit and markov decision processes. In *Computational Learning Theory: 15th Annual Conference on Computational Learning Theory, COLT 2002 Sydney, Australia, July 8–10, 2002 Proceedings 15*, pages 255–270. Springer, 2002.

Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. volume 25, 2012.

Jean Honorio and Tommi Jaakkola. Tight bounds for the expected risk of linear classifiers and pac-bayes finite-sample guarantees. In *Artificial Intelligence and Statistics*, pages 384–392. PMLR, 2014.

Yunlong Hou, Vincent YF Tan, and Zixin Zhong. Almost optimal variance-constrained best arm identification. *IEEE Transactions on Information Theory*, 69(4):2603–2634, 2022.

Xiaoguang Huo and Feng Fu. Risk-aware multi-armed bandit problem with application to portfolio selection. *Royal Society open science*, 4(11):171377, 2017.

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. Pac subset selection in stochastic multi-armed bandits. In *ICML*, volume 12, pages 655–662, 2012.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *International conference on machine learning*, pages 1238–1246. PMLR, 2013.

Ramtin Keramati, Christoph Dann, Alex Tamkin, and Emma Brunskill. Being optimistic to be conservative: Quickly learning a cvar policy. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 4436–4443, 2020.

Cyrille Kone, Emilie Kaufmann, and Laura Richert. Adaptive algorithms for relaxed pareto set identification. *Advances in Neural Information Processing Systems*, 36:35190–35201, 2023.

Cyrille Kone, Emilie Kaufmann, and Laura Richert. Bandit pareto set identification: the fixed budget setting. In *International Conference on Artificial Intelligence and Statistics*, pages 2548–2556. PMLR, 2024.

Cyrille Kone, Emilie Kaufmann, and Laura Richert. Bandit pareto set identification in a multi-output linear model. In *Seventeenth European Workshop on Reinforcement Learning*, 2025.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms.* Cambridge University Press, Cambridge, United Kingdom, 2020.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 297–306, 2011.

Neel S Madhukar, Prashant K Khade, Linda Huang, Kaitlyn Gayvert, Giuseppe Galletti, Martin Stogniew, Joshua E Allen, Paraskevi Giannakakou, and Olivier Elemento. A new big-data paradigm for target identification and drug discovery. *Biorxiv*, page 134973, 2017.

Josef Pannee, Johan Gobom, Leslie M Shaw, Magdalena Korecka, Erin E Chambers, Mary Lame, Rand Jenkins, William Mylott, Maria C Carrillo, Ingrid Zegers, et al. Round robin test on quantification of amyloid-β 1–42 in cerebrospinal fluid by mass spectrometry. *Alzheimer's & Dementia*, 12(1):55–59, 2016.

Lijing Qin, Shouyuan Chen, and Xiaoyan Zhu. Contextual combinatorial bandit and its application on diversified online recommendation. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 461–469. SIAM, 2014.

R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

Amir Sani, Alessandro Lazaric, and Rémi Munos. Risk-aversion in multi-armed bandits. *Advances in neural information processing systems*, 25, 2012.

Yi Shen, Jessilyn Dunn, and Michael M Zavlanos. Risk-averse multi-armed bandits with unobserved confounders: A case study in emotion regulation in mobile health. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 144–149. IEEE, 2022.

Alex Tamkin, Ramtin Keramati, Christoph Dann, and Emma Brunskill. Distributionally-aware exploration for cvar bandits. In *NeurIPS 2019 Workshop on Safety and Robustness on Decision Making*, 2019.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12:389–434, 2012.

Tamara Ulrich, Dimo Brockhoff, and Eckart Zitzler. Pattern identification in pareto-set approximations. In *Proceedings of the 10th annual conference on Genetic and evolutionary computation*, pages 737–744, 2008.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Kaifeng Yang, Michael Emmerich, André Deutz, and Thomas Bäck. Efficient computation of expected hypervolume improvement using box decomposition algorithms. *Journal of Global Optimization*, 75:3–34, 2019.

Eckart Zitzler, Dimo Brockhoff, and Lothar Thiele. The hypervolume indicator revisited: On the design of pareto-compliant indicators via weighted integration. In *Evolutionary Multi-Criterion Optimization: 4th International Conference, EMO 2007, Matsushima, Japan, March 5-8, 2007. Proceedings 4*, pages 862–876. Springer, 2007.