Multi-thresholding Good Arm Identification with Bandit Feedback

Xuanke Jiang 3ie22005s@s.kyushu-u.ac.jp
Kyushu University/RIKEN AIP

Sherief Hashima Sherief.hashima@riken.jp
RIKEN AIP

Kohei Hatano hatano@inf.kyushu-u.ac.jp
Kyushu University/RIKEN AIP

Eiji Takimoto eiji@inf.kyushu-u.ac.jp
Kyushu University

Editors: Hung-yi Lee and Tongliang Liu

Abstract

We consider a good arm identification problem in a stochastic bandit setting with multi-objectives, where each arm $i \in [K]$ is associated with a distribution D_i defined over R^M. For each round t, the player pulls an arm i_t and receives an M-dimensional reward vector sampled according to D_{i_t}. The goal is to find, with high probability, an ϵ-good arm whose expected reward vector is larger than $\boldsymbol{\xi} - \epsilon\mathbf{1}$, where $\boldsymbol{\xi}$ is a predefined threshold vector, and the vector comparison is component-wise. We propose the Multi-Thresholding UCB (MultiTUCB) algorithm with a sample complexity bound. Our bound matches the existing one in the special case where $M = 1$ and $\epsilon = 0$. The proposed algorithm demonstrates superior performance compared to baseline approaches across synthetic and real datasets.

Keywords: Multi-objective Optimization; Good Arm Identification; Thresholding Bandit

1. Introduction

Single objective thresholding problem was proposed by Locatelli et al. (2016); Kano et al. (2019), where the player/algorithm aims to find a subset of all arms $[K]$ denoted as the good arm set $[K]_{\text{good}}$. An arm is good if it has an average reward that surpasses a predefined performance threshold ξ with a high probability. This problem mainly adheres to two mainstream directions. The first one is known as the *fixed budget setting* (Locatelli et al., 2016), in which a fixed computational or sampling budget T is given. This framework focuses on deriving a theoretical upper bound on the probability of errors associated with the selection of the arms. The second framework, termed the *fixed confidence setting* (Kano et al., 2019), aims to minimize the total sample complexity required to reliably identify the set of arms above the threshold while adhering to a predefined confidence level specified by an error tolerance parameter δ. In this case, it is required to ensure that the selected set of arms is correct with a confidence level of at least $1 - \delta$. This setting is particularly relevant in applications where achieving a desired level of statistical accuracy is critical and minimizing resource consumption remains a key consideration. Both frameworks offer valuable insights into pure exploration in decision-making under uncertainty by addressing these distinct but complementary goals.

© 2025 X. Jiang, S. Hashima, K. Hatano & E. Takimoto.

Concurrently, identifying arms under specific requirements with bandit feedback has garnered significant attention within topics on machine learning and sequential decision-making (Zhao et al., 2023; Kano et al., 2019). However, existing works only considered a single threshold, while in this work, we extend it into multiple thresholds with the following two motivating examples:

Example 1 (Machine Retention in Manufacturing) *Consider a factory that operates multiple types of machines, each responsible for producing different components. The factory manager is tasked with optimizing operational costs without necessarily identifying the single most efficient machine. Hence, the manager implements a policy where any machine is retained if it cannot match thresholds across multiple issues. For instance, all remaining machines should have an energy consumption lower than ξ_1, maintenance cost lower than ξ_2, labor cost lower than ξ_3, etc. This approach ensures that all machines contributing positively to the overall profit are kept in operation, even if some are not the most efficient relative to others.*

Example 2 (Internet of Things (IoT) System Deployment) *In large-scale IoT deployments, a network operator must choose one configuration from a pool of candidate sensor node designs or communication protocols, each exhibiting uncertain performance due to hardware variability, environmental noise, and dynamic workloads. Each candidate is evaluated based on multiple key performance metrics such as communication latency, energy consumption, and data reliability. Rather than selecting the optimal configuration, the objective is to identify any one setup that satisfies minimum operational thresholds across all metrics—for example, latency below 100 ms, packet loss under 1%, and energy usage within a predefined budget. Due to resource constraints, it is essential to minimize the number of evaluations required for testing each configuration.*

Existing research on Multi-Objective Optimization (MOO), e.g., Pereira et al. (2022), is not applicable to the above scenarios. Researchers on MOO typically aim to achieve Pareto optimality, e.g., Jiang et al. (2023), in which improving one objective is impossible without negatively affecting another. On the basis of this, finding the Pareto front under the Multi-Armed Bandit (MAB) setting for MOO has also gathered plenty of attention (Kaufmann et al., 2016; Kone et al., 2023), which aims to find a subset of arms that achieves the Pareto optimality to form the front with optimal sample complexity. The Track-and-Stop framework (Kaufmann et al., 2016; Garivier et al., 2024) has been implemented in this problem. However, these methods are not suitable for our setting since the goal is different, as shown in Figure 1.

We integrate the difficulties of handling multiple objectives and restrictions, and thus we consider the case where multiple thresholds are simultaneously considered across all K arms. Note that the concentration inequality for the single objective case, as $M = 1$, cannot be applied directly to multi-objectives. We address this challenge by establishing a new framework for good arm identification problems with multiple thresholds. At each round t, the player pulls one arm $i \in [K]$, in which $[K] = \{1, 2, \ldots, K\}$, and receives a M dimensional reward vector $\boldsymbol{z}_{t,i}$ sampled i.i.d. from a distribution $\mathcal{D}_i$. Concrete details are presented in section 2. The player's goal is to identify a good arm, whose expected reward

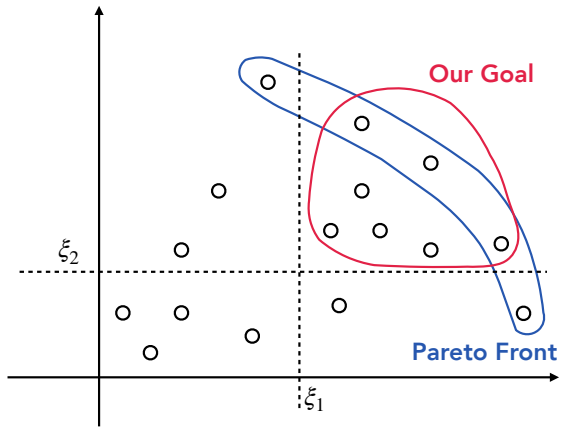

Figure 1: We compare our goal, which is to identify the arms whose expected average rewards exceed given thresholds, with the Pareto front (Jiang et al., 2023), which consists of the set of Pareto optimal arms.

vector is larger than $\boldsymbol{\xi} - \epsilon\mathbf{1}$ for $\epsilon \in [0, 1]$. If no such arm exists, the player selects $\perp$ while minimizing sample complexity with high probability.

Our contributions can be summarized as follows:

1. We introduce the Multi-Thresholding Good Arm Identification (MTGAI) framework, which addresses the challenge of identifying arms that satisfy multiple thresholds simultaneously. (i)First, we construct a reduction from the multi-thresholding problem to a single thresholding problem. However, it seems impossible to construct a Bernoulli or sub-Gaussian random variable $z_{t,i}$ with

$$\mathbb{E}[z_{t,i}] = \max\left\{\xi_1 - \mu_i^{(1)}, \xi_2 - \mu_i^{(2)}, \ldots, \xi_M - \mu_i^{(M)}\right\},$$

   and so we cannot immediately apply existing algorithms such as Locatelli et al. (2016) and Kano et al. (2019). (ii)Instead, we construct a new estimator and derive a concentration inequality for it, which implies that our modification of the single thresholding algorithms works.

2. We propose the Multi-Thresholding UCB (MultiTUCB) algorithm for the MTGAI problem and derive an upper bound on its sample complexity that is consistent with existing results when $M = 1$ and $\epsilon = 0$ (Kano et al., 2019). Our analysis significantly generalizes prior work by extending the theoretical guarantees from Bernoulli rewards to sub-Gaussian distributions and from zero accuracy ($\epsilon = 0$) to a more practical setting with non-zero accuracy ($\epsilon > 0$), which introduces substantial technical complexity in the proofs. Moreover, we design a refined arm selection rule tailored to the multi-thresholding setting, enabling more efficient exploration under multiple constraints.

3. We establish an almost matching lower bound for the MTGAI problem when $\epsilon = 0$, showing the near-optimality of our approach.

4. Experimental results on both synthetic and real datasets demonstrate that our algorithm consistently outperforms three strong baselines adapted from prior thresholding bandit methods (Kano et al., 2019; Locatelli et al., 2016; Kalyanakrishnan et al., 2012).

1.1. Other Related Work

In addition to the papers discussed in the previous part, we discuss further related work to our settings.

Best Arm Identification (BAI) The BAI is a pure exploration problem in which the agent tries to identify the best arm with the maximal mean reward within a limited sample (Audibert and Bubeck, 2010). Unlike minimizing the regret for online learning, which involves the dilemma of exploration and exploitation, BAI focuses exclusively on exploration—gathering enough information to determine the arm that maximizes the mean reward. In addition, it has been extended to varying reward distributions, or constraints such as fixed confidence (Garivier and Kaufmann, 2016; Jourdan et al., 2023), making it a versatile tool in decision-making under uncertainty.

Multi-objective Optimization As for MOO, achieving Pareto optimality is of essential importance for making in-time predictions, particularly in contexts where data explosion poses significant challenges. Various approaches have been developed for achieving Pareto optimality, including evolutionary methods (Murata et al., 1995), hypervolume scalarization (Zhang and Golovin, 2020), and multiple gradient-based methods (Sener and Koltun, 2018). This applies to various real applications, such as resource management (Liu et al., 2024) and load forecasting (Xing et al., 2024). Alternatively, researchers have also studied the problem from the perspective of regret minimization Lu et al. (2019); Cheu et al. (2018), which serves as a different goal. In parallel, researchers in economics have considered selecting one arm from the Pareto front that optimizes the Generalized Gini Index (Busa-Fekete et al., 2017).

However, all previous works do not solve our MTGAI problem, unlike our proposed MultiTUCB algorithm, which will be explained in Section 3.

2. Preliminaries

This section presents basic tools and a formal problem formulation. First, we introduce the definition of (μ, σ)-subgaussian random variable and its property.

Definition 1 ((μ, σ)-subgaussian) *A random variable z that takes values in $\mathbb{R}$ is (μ, σ)-subgaussian if $\mu = \mathbb{E}[\boldsymbol{z}]$ and for any $\lambda \in \mathbb{R}$,*

$$\mathbb{E}[\exp(\lambda(z - \mu))] \leq \exp(\lambda^2 \sigma^2 / 2).$$

Proposition 2 (See e.g. Lattimore and Szepesvári (2020)) *Let $z_1, \ldots, z_t$ be i.i.d. (μ, σ)-subgaussian random variables. Then for any $\epsilon \geq 0$,*

$$\mathbb{P}\left(\hat{\mu}_t \geq \mu + \epsilon\right) \leq \exp\left(-\frac{t\epsilon^2}{2\sigma^2}\right) \text{ and } \mathbb{P}\left(\hat{\mu}_t \leq \mu - \epsilon\right) \leq \exp\left(-\frac{t\epsilon^2}{2\sigma^2}\right),$$

where $\hat{\mu}_t = \frac{1}{t}\sum_{s=1}^{t} z_s$.

Table 1: Notation List

$T_i(t)$	The number of times arm i is chosen until round t
$\hat{a}$	The output of Algorithm 2
T_{stop}	Stopping time of Algorithm 2
$\perp$	No ϵ-good arm exists
$\tilde{g}_{t,i}$	$= \hat{g}_{t,i} - \sqrt{\frac{2\sigma^2 \ln(KMT_i(t))}{T_i(t)}}$
$\underline{g}_{t,i}$	$= \hat{g}_{t,i} - \alpha(T_i(t), \delta)$
$\overline{g}_{t,i}$	$= \hat{g}_{t,i} + \alpha(T_i(t), \delta)$
i^*	$= \arg\min_{i \in [K]} g_i$

This paper investigates the protocol of MTGAI, which is defined between one player and Nature as the following protocol with K as the number of arms and M as the number of thresholds/objectives. They are given an accuracy parameter ϵ, a confidence parameter δ and thresholds $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_M) \in [0,1]^M$. For each iteration $t = 1, 2, \ldots$, (i) the player chooses an arm $i_t \in [K]$ and (ii) Nature returns a reward vector $\boldsymbol{z}_{t,i_t} \in [0,1]^M$ sampled according to $\mathcal{D}_{i_t}$. Here, $\mathcal{D}_i$ is defined as the distribution over $\mathbb{R}^M$ associated with arm i such that for the random variable $\boldsymbol{z} = \left(z^{(1)}, \ldots, z^{(M)}\right) \sim \mathcal{D}_i$, $\boldsymbol{z}^{(m)}$ is $(\mu_i^{(m)}, \sigma)$-subgaussian for some $\boldsymbol{\mu}_i = \left(\mu_i^{(1)}, \ldots, \mu_i^{(M)}\right) \in [0,1]^M$ and $\sigma \in \mathbb{R}_+$ for each $m \in [M]$. The player decides when to stop and outputs an arm $\hat{\imath} \in [K] \cup \{\perp\}$. The stopping time is considered as sample complexity. An arm i is ϵ-good if $\boldsymbol{\mu}_i \geq \boldsymbol{\xi} - \epsilon\mathbf{1}$, where the inequality is component-wise. In particular, an arm i is good if it is 0-good. The goal of the player is to find an arm $\hat{\imath}$ with $\mathbb{P}\left(\hat{\imath} \neq \perp \wedge \boldsymbol{\mu}_{\hat{\imath}} \geq \boldsymbol{\xi} - \epsilon\mathbf{1}\right) \geq 1 - \delta$ if one good arm exists and to output $\perp$ with $\mathbb{P}\left(\hat{\imath} = \perp\right) \geq 1 - \delta$ if no ϵ-good arm exists. An algorithm that achieves this goal is said to be (δ, ϵ)-*successful*. Notice that this protocol can be easily extended to finding all good arms by repeatedly running the algorithm for at most K times. We summarize all other symbols in Table 2 for convenience.

3. Multi-thresholding UCB algorithm

In this section, we propose the MultiTUCB algorithm to efficiently solve the proposed problem. First, we introduce a gap vector and its estimator to handle the multiple objectives and prepare for the algorithm. For each $i \in [K]$, let

$$g_i = g(\boldsymbol{\mu}_i) = \max\left\{\xi_1 - \mu_i^{(1)}, \xi_2 - \mu_i^{(2)}, \ldots, \xi_M - \mu_i^{(M)}\right\}. \tag{1}$$

Note that for any ϵ-good arm i, $g_i \leq \epsilon$. The gap vector $\boldsymbol{g}$ is defined as $\boldsymbol{g} = (g_1, \ldots, g_K)$.

Definition 3 *(α-**approx estimator**) For a function $\alpha : \mathbb{N} \times (0,1) \to \mathbb{R}_+$, we define α-approx estimator of $\boldsymbol{g}$ is an oracle that satisfies the condition. For the t-th call to the oracle, it receives input $i_t \in [K]$ and returns $\hat{\boldsymbol{g}}_t = (\hat{g}_{t,1}, \ldots, \hat{g}_{t,K}) \in \mathbb{R}^K$ such that*

$$\forall \delta \in (0,1), \mathbb{P}\left(\forall i \in [K], \forall s \in [t], |\hat{g}_{s,i} - g_i| \leq \alpha(T_i(s), \delta)\right) \geq 1 - \delta, \tag{2}$$

where $T_i(s) = |\{u \in [s] | i_u = i\}|$.

We claim that $\boldsymbol{g}$ satisfies the following proposition.

Proposition 4 *For any $i \in [K]$ and any $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$, $\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_K \in [0,1]^M$*

$$|g(\boldsymbol{\mu}_i) - g(\hat{\boldsymbol{\mu}}_i)| \leq \|\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_i\|_\infty.$$

The proof is given in Appendix A.2.

Now, we show the existence of α-approx estimator for some α. The algorithm for the estimator is described in Algorithm 1.

Algorithm 1 Approx Estimator

for $i \in [K]$ **do**

$$\hat{\boldsymbol{\mu}}_{t,i} = \frac{\sum_{s=1}^{t} \mathbb{1}[i_s = i] \boldsymbol{z}_{s,i}}{T_i(t)};$$

end

Output: $\hat{\boldsymbol{g}}_t = (g(\hat{\boldsymbol{\mu}}_{t,1}), \ldots, g(\hat{\boldsymbol{\mu}}_{t,K}))$;

Hence, we calculate the exact value of the α-approx estimator under our setting.

Proposition 5 *Algorithm 1 is an α-approx estimator of $\boldsymbol{g}$ with*

$$\alpha(\tau, \delta) = \sqrt{\frac{2\sigma^2 \ln \frac{\pi^2 K M \tau^2}{3\delta}}{\tau}}.$$

The proof is given in Appendix A.1.

Then, we propose the MultiTUCB algorithm to efficiently solve the proposed problem. The detail of the algorithm is given in Algorithm 2. It uses Algorithm 1 for the α-approx estimator as shown in Proposition 5 in the stopping criteria. We denote T_{stop} as the stopping time and $\hat{a}$ as the output of Algorithm 2.

3.1. Upper Bounds with Expectation

First, we guarantee the correctness of our proposed algorithm with the following theorem.

Theorem 6 *Algorithm 2 is (δ, ϵ)-successful.*

Details are deferred to Appendix A.3.

Definition 7 *Let $t_i : \mathbb{R}_+ \to \mathbb{R}_+$ be a function defined as follows.*

For any good arm i,

$$t_i(\epsilon_0) = \max\left\{ \frac{4\sigma^2}{(\epsilon - g_i - \epsilon_0)^2} \ln\left(\frac{8\sqrt{3}\sigma^2 \pi K M/\delta}{3(\epsilon - g_i - \epsilon_0)^2} \ln \frac{4\sqrt{3}\pi\sigma^2}{3(\epsilon - g_i - \epsilon_0)^2} \right), 0 \right\}.$$

For any non-ϵ-good arm i,

$$t_i(\epsilon_0) = \max\left\{ \frac{4\sigma^2}{(g_i - \epsilon - \epsilon_0)^2} \ln\left(\frac{8\sqrt{3}\sigma^2 \pi K M/\delta}{3(g_i - \epsilon - \epsilon_0)^2} \ln \frac{4\sqrt{3}\pi\sigma^2}{3(g_i - \epsilon - \epsilon_0)^2} \right), 0 \right\}.$$

Algorithm 2 Multi-thresholding UCB algorithm (MultiTUCB)

Input: Arm set $[K]$, thresholds $\{\xi_1, \ldots, \xi_M\}$, confidence parameter δ, accuracy rate ϵ;
Init: Let $\mathcal{A} = [K]$;
Pull each arm once;
Compute $\hat{\boldsymbol{g}}_K$ by Algorithm 1;
for $t \in [K+1, \ldots,]$ **do**
  Select $i_t \in \arg\min_{i \in \mathcal{A}} \tilde{g}_{t-1,i}$ with $\tilde{g}_{t-1,i} = \hat{g}_{t-1,i} - \sqrt{\frac{2\sigma^2 \ln(KMT_i(t-1))}{T_i(t-1)}}$;
  Receive feedback $\boldsymbol{z}_{t,i_t}$;
  Update $T_i(t)$;
  Compute $\hat{\boldsymbol{g}}_t$ by Algorithm 1;
  if $\underline{g}_{t,i_t} > 0$ **then**
    Delete i_t from $\mathcal{A}$.
  end
  //**Condition 1**
  if $\overline{g}_{t,i_t} \leq \epsilon$ **then**
    Output $\hat{a} = i_t$, Stop.
  end
  //**Condition 2**
  if $\mathcal{A} = \emptyset$ **then**
    Output $\hat{a} = \perp$, Stop.
  end
end

Theorem 8 *If there exists a good arm, letting $i^* = \arg\min_{i \in [K]} g_i$, Algorithm 2 achieves*

$$\mathbb{E}[T_{stop}] \leq t_{i^*}(\epsilon_0) + \sum_{i \neq i^*} \frac{8\sigma^2 \ln\left(KM \max_{i \in [K]} \lfloor t_i(\epsilon_0) \rfloor\right)}{(g_i - g_{i^*} + \epsilon_0)^2} + \frac{2(K+1)M\sigma^2}{{\epsilon_0}^2} + \frac{K^3 M}{2{\epsilon_0}^2} e^{4{\epsilon_0}^2},$$

where ϵ_0 be any number s.t. $0 < \epsilon_0 < \epsilon - g_i$ for any good arm i and be any number s.t. $0 < \epsilon_0 < g_i - \epsilon$ with any non-ϵ-good arm i.

Proof sketch of Theorem 8: The stopping time with an output $\hat{a} \in [K]$ can be divided as follows: (i) the number of rounds at which the algorithm chooses $\hat{a}$, (ii) the number of rounds at which the algorithm chooses another ϵ-good arm, and (iii) the number of rounds at which the algorithm chooses any non-ϵ-good arms. Bounding the expectation of each term and combining them lead to the final result. Appendix A.4 details the full proof.

Theorem 9 *If there is no ϵ-good arm, Algorithm 2 achieves*

$$\mathbb{E}[T_{stop}] \leq \sum_{i \in [K]} t_i(\epsilon_0) + \frac{KM\sigma^2}{\epsilon_0^2},$$

while ϵ_0 be any number such that $0 < \epsilon_0 < g_i - \epsilon$ for any arm $i \in [K]$.

Proof sketch of Theorem 9: If no good arm exists and the algorithm continues at round t, Condition 2 is not satisfied, and thus the active set is not empty. This means some arm i_t is chosen at round t with $\hat{g}_{t,i_t} \leq \alpha(T_i(t), \delta)$. Bounding the number of such rounds over all arms leads to the result. The detailed proof of Theorem 9 is postponed to Appendix A.5.

We present a more straightforward result based on the above theorems in the following corollary.

Corollary 10 *Algorithm 2 achieves the following:*

1. *If there exists a good arm and letting $i^* = \arg\min_{i\in[K]} g_i$,*

$$\limsup_{\delta\to 0} \frac{\mathbb{E}[T_{stop}]}{\ln(1/\delta)} \leq \frac{4\sigma^2}{(\epsilon - g_{i^*} - \epsilon_0)^2},$$

   where ϵ_0 be any number such that $0 < \epsilon_0 < \epsilon - g_i$ for any good arm i.

2. *If there is no ϵ-good arm,*

$$\limsup_{\delta\to 0} \frac{\mathbb{E}[T_{stop}]}{\ln(1/\delta)} \leq \sum_{i\in[K]} \frac{4\sigma^2}{(\epsilon - g_i - \epsilon_0)^2},$$

   where ϵ_0 be any number such that $0 < \epsilon_0 < g_i - \epsilon$ with any arm $i \in [K]$.

4. Lower Bound

In the following part, we give a lower bound for the expectation of T_{stop}.

Definition 11 (Binary relative entropy) *For $x, y \in \mathbb{R}_+$, the binary relative entropy is defined as*

$$d(x, y) = x\ln(x/y) + (1-x)\ln(1-x)/(1-y))$$

with conventions $d(0,0) = d(1,1) = 0$.

Theorem 12 *There exists a bandit model s.t. the following holds for any $(\delta, 0)$-successful algorithm with stopping time T_{stop}:*

1. *If good arm i exists,*

$$\mathbb{E}[T_{stop}] \geq \frac{1}{\max_{i\in[K]_{good}} \min_{m\in[M]} d\left(\mu_i^{(m)}, \xi_m\right)} \ln\frac{1}{2\delta} - \frac{\delta}{\max_{i\in[K]_{good}} \min_{m\in[M]} d\left(\mu_i^{(m)}, \xi_m\right)},$$

   where $[K]_{good} \subseteq [K]$ is the set of all good arms.

2. *When no good arm exists,*

$$\mathbb{E}[T_{stop}] \geq \frac{1}{\max_{i\in[K]} \min_{m\in[M]} d\left(\mu_i^{(m)}, \xi_m\right)} \ln\frac{1}{2\delta} - \frac{\delta}{\max_{i\in[K]} \min_{m\in[M]} d\left(\mu_i^{(m)}, \xi_m\right)}.$$

Detailed proof is in Appendix A.6.

Discussion on the tightness of the bounds By Pinsker's inequality and Sason and Verdú (2015), we obtain

$$\frac{2g_{i^*}^2}{a_{\boldsymbol{\xi}}} \geq \max_{i\in[K]_{\text{good}}} \min_{m\in[M]} d\left(\mu_i^{(m)}, \xi_m\right) \geq 2g_{i^*}^2,$$

where $a_{\boldsymbol{\xi}} := \min_{m\in[M]} \min\{\xi_m, 1-\xi_m\}$. Then we have

$$\limsup_{\delta\to 0} \frac{\mathbb{E}[T_{\text{stop}}]}{\ln 1/\delta} \geq \frac{a_{\boldsymbol{\xi}}}{2g_{i^*}^2}.$$

When every element of $\boldsymbol{\xi}$ is not close to 0 or 1, $a_{\boldsymbol{\xi}}$ is large enough to be considered as a constant. On the other hand, when feedback follows Bernoulli distributions, which is $\frac{1}{4}$-sub-Gaussian, $\epsilon = 0$ and a good arm exists, we have the following result for the upper bound

$$\limsup_{\delta\to 0} \frac{\mathbb{E}[T_{\text{stop}}]}{\ln 1/\delta} \leq \frac{1}{g_{i^*}^2}.$$

Thus, we see that our upper bound for $(\delta, 0)-$successful algorithm almost matches the lower bound when a good arm exists.

5. Experiments

In this section, we compare our proposed MultiTUCB with three benchmarks, which are Anytime Parameter-free Thresholding algorithm for MTGAI (MultiAPT) (Locatelli et al., 2016; Jourdan et al., 2023), Lower and Upper Confidence Bounds algorithm for MTGAI (MultiLUCB) (Kalyanakrishnan et al., 2012) , Hybrid algorithm for the Dilemma of Confidence for MTGAI (MultiHDoC) (Kano et al., 2019) in terms of stopping time. All the algorithms are presented in Algorithm 3. As for MultiLUCB, we adopt a lower confidence bound for g_i, i.e., $\hat{g}_i - \alpha(T_i(t), \delta)$ as the selection rule. Note that Algorithm 3 is guaranteed to solve the MTGAI problem for any benchmark algorithms and ours when it terminates by Theorem 6. However, unlike ours, the benchmark algorithms are designed for the single thresholding problem only, and they do not have sample complexity bounds for the multi-thresholding problem. The source codes are available at Github.

5.1. Synthetic Data

We consider the following 4 different environments with $K = 10, M = 4$.

1. Gaussian distribution with means $\mu_{1:3}^{(1)} = 0.1$, $\mu_4^{(1)} = 0.35$, $\mu_5^{(1)} = 0.45$, $\mu_6^{(1)} = 0.55$, $\mu_7^{(1)} = 0.65$ and $\mu_{8:10}^{(1)} = 0.2$.

2. Gaussian distributions with means $\mu_{1:4}^{(2)} = 0.4 - 0.2^{1:4}, \mu_5^{(2)} = 0.45, \mu_6^{(2)} = 0.55$, $\mu_{7:10}^{(2)} = 0.6 + 0.1^{5-(1:4)}$.

3. Gaussian distributions with increasing means $\mu_{1:4}^{(3)} = (1:4)\cdot 0.05$, $\mu_5^{(3)} = 0.45$, $\mu_6^{(3)} = 0.55$ and $\mu_{7:10}^{(3)} = 0.65 + (0:3)\cdot 0.05$.

Algorithm 3 MultiHDoC/MultiLUCB/MultiAPT-G

Input: Arm set $[K]$, thresholds $\{\xi_1, \ldots, \xi_M\}$, confidence parameter δ, accuracy parameter ϵ;
Init: Let $\mathcal{A}_{\text{HDoC}} = [K]/\mathcal{A}_{\text{LUCB}} = [K]/\mathcal{A}_{\text{APT}} = [K]$;
Pull each arm once;
Compute $\hat{\boldsymbol{g}}_K$ by Algorithm 1;
for $t \in [K+1, \ldots,]$ **do**
  MultiHDoC: Pull arm $i_t = \arg\min_{i \in \mathcal{A}_{\text{HDoC}}} \hat{g}_{t,i} - \sqrt{\frac{\ln t}{2T_i(t)}}$.
  MultiLUCB: Pull arm $i_t = \arg\min_{i \in \mathcal{A}_{\text{LUCB}}} \hat{g}_{t,i} - \alpha(T_i(t), \delta)$.
  MultiAPT: Pull arm $i_t = \arg\min_{i \in \mathcal{A}_{\text{APT}}} \sqrt{T_i(t)}\,|\hat{g}_{t,i} - \epsilon|$.
  Receive feedback $\boldsymbol{z}_{t,i_t}$;
  Update $T_i(t)$;
  Compute $\hat{\boldsymbol{g}}_t$ by Algorithm 1;
  if $\underline{g}_{t,i_t} > 0$ **then**
    Delete i_t from $\mathcal{A}_{\text{HDoC}}/\mathcal{A}_{\text{LUCB}}/\mathcal{A}_{\text{APT}}$.
  end
  if $\overline{g}_{t,i_t} \leq \epsilon$ **then**
    Output $\hat{a} = i_t$ as a good arm.
    Stop.
  end
  if $\mathcal{A}_{HDoC}/\mathcal{A}_{LUCB}/\mathcal{A}_{APT} = \emptyset$ **then**
    Output $\hat{a} = \perp$, Stop.
  end
end

4. Gaussian distributions with grouped means $\mu^{(4)}_{1:4} = 0.4$, $\mu^{(4)}_{5:8} = 0.5$, $\mu^{(6)}_{9:10} = 0.6$.

We set thresholds $\{\xi_1, \xi_2, \xi_3, \xi_4\} = \{0.6, 0.5, 0.6, 0.5\}$ and σ is chosen as 1.2 while all distributions share the same variance for convenience. All the baselines in Algorithm 3 share the same stopping criteria as Algorithm 2. We set a time limit of $200,000$ arm-pulls, and passing the limitation is counted as an error. All results are averaged over 5000 repetitions.

Figure 2. exhibits the average stopping time against different values of δ from 0.005 to 0.05 for the compared schemes at $\epsilon = 0.005$. MultiTUCB outperforms other compared schemes since it is theoretically guaranteed. As δ increases, the difficulty of exiting the algorithm decreases, which may result in a shorter stopping time. Additionally, the performance curve of MultiLUCB is rugged, which indicates that it is more sensitive to the change of δ and needs more rounds to be stabilized. The others are not very sensitive to the change of δ, but we observe that the number of required rounds decreases as the confidence parameter increases.

Figure 3 previews the stopping rounds of the compared schemes with the increasing $\epsilon = \{0.002, 0.004, \ldots, 0.02\}$ and $\delta = 0.005$. Here, we notice that MultiHDoC performs close to MultiTUCB since it is designed for finding a good arm set with a single threshold. It

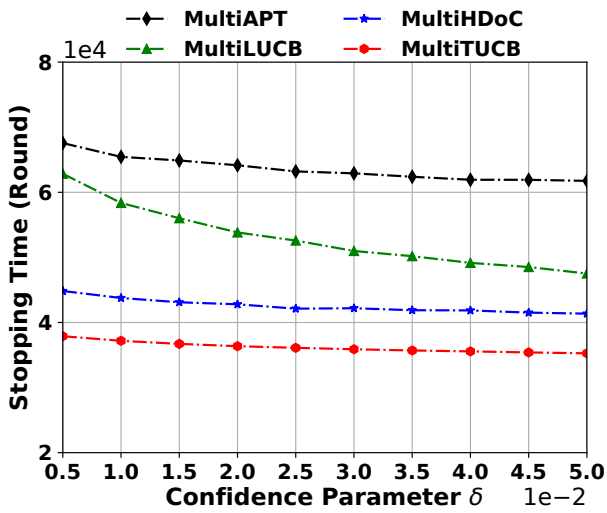

Figure 2: Stopping times w.r.t. δ with synthetic data.

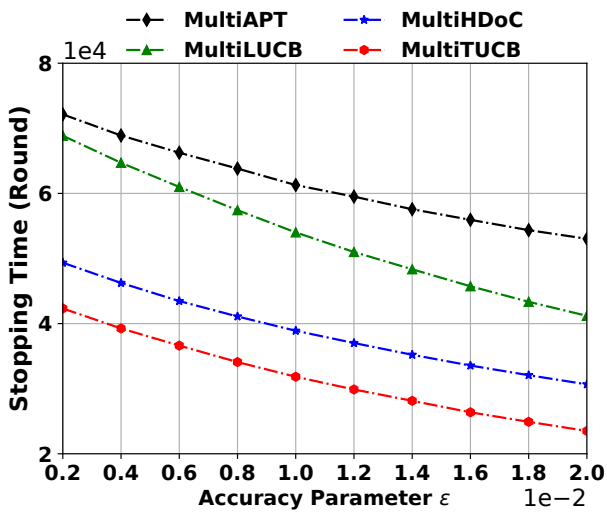

Figure 3: Stopping times w.r.t. ϵ with synthetic data.

also indicates our modifications to the baseline algorithm are valid. The stopping time of all algorithms decreases along with the relaxed stopping condition. Our proposed algorithm outperforms others not only in terms of straightforward stopping time but also in stability and less turbulence. We generalize the detailed data in Table 2 and 3 for comparison.

Table 2: Standard Deviation of Stopping Time w.r.t. δ with Synthetic Data

δ	MultiAPT	MultiHDoC	MultiLUCB	**MultiTUCB**
0.005	9408.58	14589.53	9640.70	**7465.55**
0.010	9197.93	14374.44	9464.46	**7393.21**
0.015	9359.25	14167.67	9371.94	**7354.64**
0.020	8861.37	13947.92	8785.98	**7382.50**
0.025	9154.26	13569.75	8885.88	**7313.47**
0.030	9020.14	13688.36	8389.23	**7302.48**
0.035	8967.06	13513.67	8278.27	**7233.80**
0.040	8909.38	13612.20	8022.23	**7225.78**
0.045	9161.23	13393.54	7816.71	**7237.46**
0.050	8914.17	13429.67	8034.80	**7213.29**

Table 3: Standard Deviation of Stopping Time w.r.t. ϵ with Synthetic Data

ϵ	MultiAPT	MultiHDoC	MultiLUCB	**MultiTUCB**
0.002	10213.43	15203.78	10236.82	**8116.02**
0.004	9704.71	14821.15	9840.27	**7615.36**
0.006	9035.50	14501.22	9520.23	**7284.30**
0.008	8803.26	14413.35	9018.05	**6953.02**
0.010	8610.67	14300.52	8436.28	**6672.03**
0.012	8188.97	14164.40	7995.11	**6369.55**
0.014	8227.20	14083.09	7582.54	**6099.59**
0.016	8126.56	14008.76	7144.09	**5886.65**
0.018	8042.25	13878.92	6765.63	**5679.36**
0.020	7956.99	13810.01	6422.68	**5404.42**

5.2. Dose Confirmation for Cocktail Therapy

Cocktail therapy refers to a treatment approach that combines multiple drugs or therapeutic agents to enhance effectiveness and achieve synergistic benefits. By targeting different pathways simultaneously, cocktail therapy can improve treatment outcomes, reduce drug resistance, and minimize side effects compared to single-drug treatments. Our work provides insights into confirming the dose of each medicine in cocktail therapy. Current research shows that overweight and type 2 diabetes are related (Ruze et al., 2023), and we give insight into determining the dose of a possible cocktail therapy on two medicines for a better treatment. Here, we regularize all data into the range $(0, 1]$ for convenience, and the larger means indicate the better curative effect.

1. (Dose finding of LY3437943 compared with Dulaglutide (1.5mg) on glycated haemoglobin change w.r.t. placebo (Urva et al., 2022)): Five Gaussian distributions with mean $\mu_1^{(1)} = 0.36$, $\mu_2^{(1)} = 0.59$, $\mu_3^{(1)} = 0.85$, $\mu_4^{(1)} = 0.95$, $\mu_5^{(1)} = 0.79$, corresponds to dose with unit mg as $\{1.5, 3, 3/6, 3/6/9/12\}$ (3/6 means prescribe 3mg for the first half of treatments and 6mg for the next half), and threshold $\xi_1 = 0.48, \sigma = 1.0$.

2. (Dose finding of cagrilintide compared with Liraglutide (3.0mg) on bodyweight change w.r.t. placebo (Lau et al., 2021)): Five Gaussian distributions with mean $\mu_1^{(2)} = 0.375$, $\mu_2^{(2)} = 0.475$, $\mu_3^{(2)} = 0.7625$, $\mu_4^{(2)} = 0.8375$, $\mu_5^{(2)} = 0.975$, corresponds to dose with unit mg as $\{0.3, 0.6, 1.2, 2.4, 4.5\}$ with $\sigma = 1.0$ and threshold $\xi_2 = 0.75$.

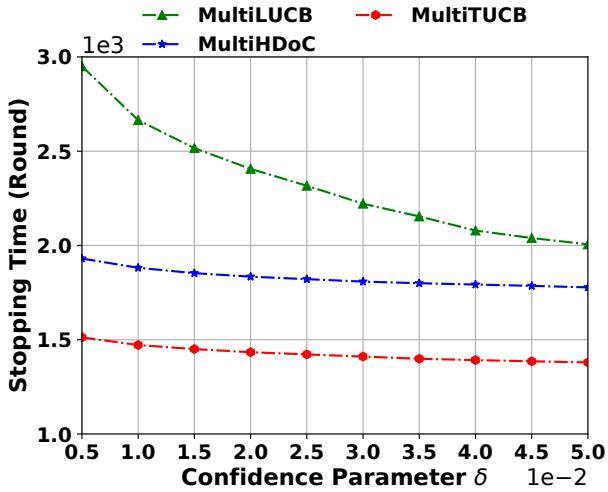

Figure 4: Stopping times w.r.t. δ with medical data.

We combine LY3437943 (Urva et al., 2022) and cagrilintide (Lau et al., 2021) as shown above and put the results in Figure 4 and Figure 5. The algorithm MultiAPT has a large stopping time compared with others in this setting, and we omit it from the figures for clarity. The behavior of MultiHDoC, MultiLUCB, and MultiTUCB is similar to the one in Section 5.1.

Other complementary results are listed in Appendix B.

6. Conclusion

This work introduced the multi-thresholding good arm identification framework in a stochastic bandit setting, where an arm is considered good if all components of its expected reward vector exceed given thresholds. We proposed the MultiTUCB algorithm, which achieves an upper bound on sample complexity scaling as $O(\ln M)$ w.r.t. $M > 1$ in the dominant term. Meanwhile, we derive corresponding lower bounds, and we prove that it is an almost match when a good arm exists and $\epsilon = 0$. Experimental results demonstrated that MultiTUCB outperforms three baseline methods on synthetic and real-world datasets.

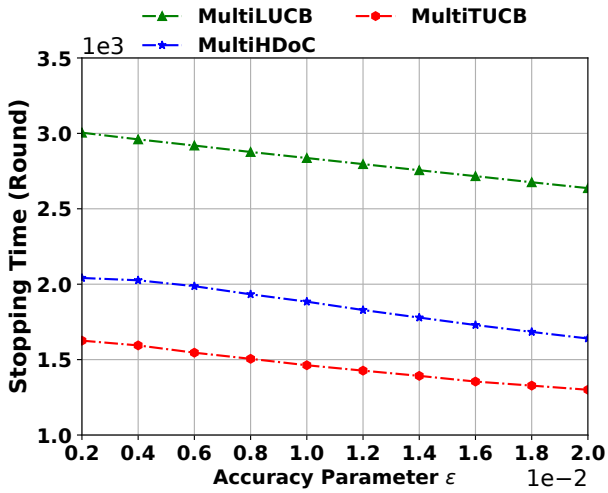

Figure 5: Stopping times w.r.t. ϵ with medical data.

As a direction for future work, it would be valuable to derive a better bound for the cases when no good arm exists or $\epsilon > 0$. Additionally, exploring more challenging settings, such as high-dimensional feedback, is another promising direction.

Acknowledgements

We thank the reviewers for their helpful comments and Atsuyoshi Nakamura for insightful discussions. This work is supported by JSPS KAKENHI Grant Numbers JP23H03348, JP25K07743, JP23K24905, and JP24H00685, respectively.

References

Jean-Yves Audibert and Sébastien Bubeck. Best arm identification in multi-armed bandits. In *COLT-23th Conference on learning theory-2010*, pages 13–p, 2010.

Róbert Busa-Fekete, Balázs Szörényi, Paul Weng, and Shie Mannor. Multi-objective bandits: Optimizing the generalized gini index. In *International Conference on Machine Learning*, pages 625–634. PMLR, 2017.

Albert Cheu, Ravi Sundaram, and Jonathan Ullman. Skyline identification in multi-arm bandits. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 1006–1010. IEEE, 2018.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR, 2016.

Aurélien Garivier, Wouter M Koolen, et al. Sequential learning of the pareto front for multi-objective bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3583–3591. PMLR, 2024.

Jiyan Jiang, Wenpeng Zhang, Shiji Zhou, Lihong Gu, Xiaodong Zeng, and Wenwu Zhu. Multi-objective online learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Marc Jourdan, Rémy Degenne, and Emilie Kaufmann. An ϵ-best-arm identification algorithm for fixed-confidence and beyond. *Advances in Neural Information Processing Systems*, 36:16578–16649, 2023.

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.

Hideaki Kano, Junya Honda, Kentaro Sakamaki, Kentaro Matsuura, Atsuyoshi Nakamura, and Masashi Sugiyama. Good arm identification via bandit feedback. *Machine Learning*, 108:721–745, 2019.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

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

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

David C W Lau, Lars Erichsen, Ann Marie Francisco, Altynai Satylganova, Carel W le Roux, Barbara McGowan, Sue D Pedersen, Kirsi H Pietiläinen, Domenica Rubino, and Rachel L Batterham. Once-weekly cagrilintide for weight management in people with overweight and obesity: a multicentre, randomised, double-blind, placebo-controlled and active-controlled, dose-finding phase 2 trial. *The Lancet*, 398(10317):2160–2172, 2021. ISSN 0140-6736. doi: https://doi.org/10.1016/S0140-6736(21)01751-7.

Qi Liu, Ruichao Mo, Xiaolong Xu, and Xu Ma. Multi-objective resource allocation in mobile edge computing using paes for internet of things. *Wireless networks*, 30(5):3533–3545, 2024.

Andrea Locatelli, Maurilio Gutzeit, and Alexandra Carpentier. An optimal algorithm for the thresholding bandit problem. In *International Conference on Machine Learning*, pages 1690–1698. PMLR, 2016.

Shiyin Lu, Guanghui Wang, Yao Hu, and Lijun Zhang. Multi-objective generalized linear bandits. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3080–3086, 2019.

Tadahiko Murata, Hisao Ishibuchi, et al. Moga: multi-objective genetic algorithms. In *IEEE international conference on evolutionary computation*, volume 1, pages 289–294. IEEE Piscataway, 1995.

João Luiz Junho Pereira, Guilherme Antônio Oliver, Matheus Brendon Francisco, Sebastiao Simoes Cunha Jr, and Guilherme Ferreira Gomes. A review of multi-objective optimization: methods and algorithms in mechanical engineering problems. *Archives of Computational Methods in Engineering*, 29(4):2285–2308, 2022.

Rexiati Ruze, Tiantong Liu, Xi Zou, Jianlu Song, Yuan Chen, Ruiyuan Xu, Xinpeng Yin, and Qiang Xu. Obesity and type 2 diabetes mellitus: connections in epidemiology, pathogenesis, and treatments. *Frontiers in endocrinology*, 14:1161521, 2023.

Igal Sason and Sergio Verdú. Upper bounds on the relative entropy and rényi divergence as a function of total variation distance for finite alphabets. In *2015 IEEE Information Theory Workshop - Fall (ITW)*, pages 214–218, 2015.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

Shweta Urva, Tamer Coskun, Mei Teng Loh, Yu Du, Melissa K Thomas, Sirel Gurbuz, Axel Haupt, Charles T Benson, Martha Hernandez-Illas, David A D'Alessio, and Zvonko Milicevic. Ly3437943, a novel triple gip, glp-1, and glucagon receptor agonist in people with type 2 diabetes: a phase 1b, multicentre, double-blind, placebo-controlled, randomised, multiple-ascending dose trial. *The Lancet*, 400(10366):1869–1881, 2022. ISSN 0140-6736. doi: https://doi.org/10.1016/S0140-6736(22)02033-5.

Abraham Wald. *Sequential analysis*. Courier Corporation, 2004.

Qianyi Xing, Xiaojia Huang, Jianzhou Wang, and Shuai Wang. A novel multivariate combined power load forecasting system based on feature selection and multi-objective intelligent optimization. *Expert Systems with Applications*, 244:122970, 2024.

Richard Zhang and Daniel Golovin. Random hypervolume scalarizations for provable multi-objective black box optimization. In *International conference on machine learning*, pages 11096–11105. PMLR, 2020.

Yao Zhao, Connor Stephens, Csaba Szepesvári, and Kwang-Sung Jun. Revisiting simple regret: Fast rates for returning a good arm. In *International Conference on Machine Learning*, pages 42110–42158. PMLR, 2023.