ROBUST OPTIMIZATION OF MULTI-OBJECTIVE MULTI-ARMED BANDITS WITH CONTAMINATED BANDIT FEEDBACK

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

MASTER OF SCIENCE

IN

ELECTRICAL AND ELECTRONICS ENGINEERING

By

Kerem Bozgan

June 2022

Robust Optimization of Multi-Objective Multi-Armed Bandits with Contaminated Bandit Feedback
By Kerem Bozgan
June 2022

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Cem Tekin(Advisor)

Çağın Ararat

Elif Vural

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

ROBUST OPTIMIZATION OF MULTI-OBJECTIVE MULTI-ARMED BANDITS WITH CONTAMINATED BANDIT FEEDBACK

Kerem Bozgan
M.S. in Electrical and Electronics Engineering
Advisor: Cem Tekin
June 2022

Multi-objective multi-armed bandits (MO-MAB) is an important extension of the standard MAB problem that has found a wide variety of applications ranging from clinical trials to online recommender systems. We consider Pareto set identification problem in the adversarial MO-MAB setting, where at each arm pull, with probability $\epsilon \in (0, \frac{1}{2})$, an adversary corrupts the reward samples by replacing the true samples with the samples from an arbitrary distribution of its choosing. Existing MO-MAB methods in the literature are incapable of handling such attacks unless there are strict restrictions on the contamination distributions. As a result, these methods perform poorly in practice where such restrictions on the adversary are not valid in general. To fill this gap in the literature, we propose two different robust, median-based optimization methods that can approximate the Pareto optimal set from contaminated samples. We prove a sample complexity bound of the form $\mathcal{O}\big(\frac{1}{\alpha^2}\log(\frac{1}{\delta})\big)$ for the proposed methods, where $\alpha > 0$ and $\delta \in (0, 1)$ are accuracy and confidence parameters, respectively, that can be set by the user according to his/her preference. This bound matches, in the worst case, the bounds from [1, Theorem 4] and [2, Theorem 3] that consider the adversary free setting. We compare the proposed methods with a mean-based method from the MO-MAB literature on real-world and synthetic experiments. Numerical results verify our theoretical expectations and show the importance of robust algorithm design in the adversarial setting.

Keywords: Multi-armed bandits, multi-objective optimization, robust optimization, adversarial attack, median based optimization, Pareto set identification.

ÖZET

ÇOKLU KOLLU ÇOKLU HEDEFLİ HAYDUTLARDA DAYANIKLI ÖĞRENME

Kerem Bozgan
Elektrik ve Elektronik Mühendisliği, Yüksek Lisans
Tez Danışmanı: Cem Tekin
Haziran 2022

Çoklu hedefli, çoklu kollu haydut optimizasyonu problemi (MO-MAB), standard çoklu kollu haydut optimizasyonu probleminin (MAB) önemli bir varyasyonu olup, klinik deneylerden, çevrimici öneri sistemlerine kadar uzanan pek çok çeşitte aplikasyonda kullanılmaktadır. Bu çalışmada, karşı saldırı varlığında, çok kollu optimizasyon problemi ele alınmıştır. Her kol çekilmesinde, kol çekilmesinden elde edilen gerçek örneklem, ϵ ile ifade edilen, 0 ile 1/2 arasında bir olasılıkla saldırıya maruz kalmakta ve örneklem kirlenmektedir. Ayrıca, saldırının, rastgele bir olasılık dağılımından seçilebildiği kabul edilmekte ve olasılık dağılımı üzerinde hiçbir sınırlama getirilmemektedir. Varolan MO-MAB calışmalarında önerilen metodlar, saldırı üzerinde çok katı sınırlamalar olmadığı sürece, dayanıksız kalmaktadır. Bu durum, bu algoritmaların, saldırının katı bir sınırlamaya tabi tutulmasının genellikle mümkün olmadığı, gerçek dünya problemlerinde, kötü bir performans göstermesine sebep olmaktadır. Literatürdeki bu boşluğu doldurmak için, dayanıklı, medyan temelli; Pareto setini, kirlenmiş örneklemlerden yola çıkarak, kullanıcı tarafından belirlenen α isabet ve δ güven parametrelerine uygun olarak tahmin edebilen, iki ayrı metod önerilmiştir. Önerilen algoritmaların, tüm kollardan aldıkları toplam örneklem sayısının, bitişe kadar, $\mathcal{O}\big(\frac{1}{\alpha^2}\log(\frac{1}{\delta})\big)$ ile verilen ifadeyle sınırlı olduğu ispat edilmiştir. Bu ifade aynı zamanda, saldırı olmadığı durumu ele alan daha önceki çalışmalarda bulunan üst sınırla eşleşmektedir [1, Theorem 4], [2, Theorem 3]. Önerilen algoritma, sentetik ve gerçek veri kullanılarak yapılan deneylerle, literatürden, ortalama bazlı bir metod ile karşılaştırılmıştır. Sonuçlarımız, teorik beklentilerimizi karşılamakta ve karşı saldırı varlığında, dayanıklı öğrenmenin gerekliliğini ispat etmektedir.

Anahtar sözcükler: çoklu kollu haydutlar, çoklu hedefli optimizasyon, dayanıklı optimizasyon, karşı saldırı, medyan bazlı optimizasyon, Pareto seti tahmini.

Acknowledgement

I would like to begin by expressing my deepest gratitude to my supervisor Assoc. Prof. Dr. Cem Tekin, for giving me the opportunity to be a member of the CYBORG research group and for his guidance and unwavering support throughout my graduate studies at Bilkent University.

I would also like to acknowledge the members of the thesis jury, ..., for the time and effort they put into reviewing my thesis and providing valuable feedback which contributed to the improvement of this work to its final form.

I would like to thank the members of my research group CYBORG: Anjum, Andi, Alireza, Alp, Sepehr, and İlker for helping me whenever I needed support and for the enlightening conversations we had that inspired me in my research. I also thank Alp for sharing his experience and giving useful tips about graduation and the thesis writing process.

I would like to thank all the members of my family who have always supported me for my entire life and who have respected and stood behind all the decisions I took towards realizing my goals. In particular, I am grateful to my mother Gülçin, who has tolerated all the ups and downs in my mood throughout the enjoyable but also the hard process of completing this work. Her unwavering discipline in life and work has inspired me during this time. I also thank her for the food she has cooked which greatly contributed to my well-being and kept me alive during this process.

Lastly, I would like to thank Dr. Hilmi Güven for his guidance and advice on the matters of academia and life in general. I am deeply grateful for his constant support during my undergraduate and graduate studies.

This work was supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) under the Grant 215E342 and through the 2210/A scholarship program.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Multi-armed bandit (MAB) problem is a decision-making problem under uncertainty in which a finite amount of resources are allocated between a limited number of choices (arms) in order to optimize gain over time (or equally minimize regret). In the stochastic MAB setting, each arm is associated with a probability distribution known as the reward distribution. The properties of these distributions might be only partially known (or not known at all) at the time of allocation (arm pull). These properties are better understood over time as more samples are obtained from the arms [3–5].

In the standard MAB setting, the goal is to maximize the reward over time by pulling arms that yield relatively high rewards. Equivalently, this can be formulated as a regret minimization problem. Regret is defined as the difference between the expected reward that one would obtain by following the optimal strategy and the actual reward obtained by the agent. For instance, if the arms are stationary, i.e., the stochastic reward distributions associated with the arms do not change over time, the optimal policy is just pulling the arm with the highest expected reward at every round. For a detailed argument on regret, the reader is referred to [6, pp. 10-11].

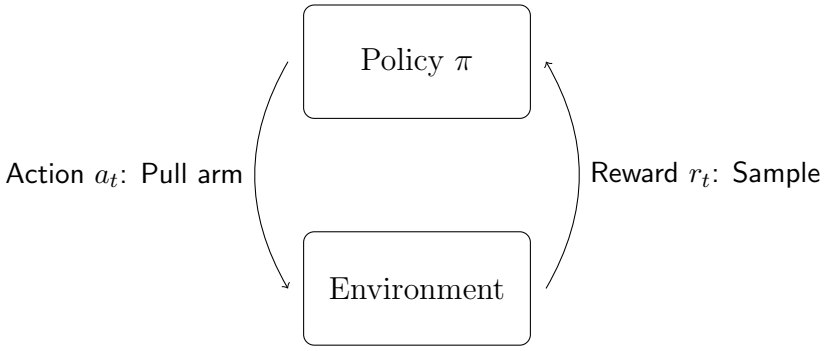


Figure 1.1: MAB problem viewed as a reinforcement learning problem.

Since the information on reward distributions is limited or non-existent initially, one relies on empirical observations to predict the performance of an arm. In general, it takes multiple pulls of an arm to reliably determine its performance. One would want to determine a policy (also known as an allocation rule) that pulls arms with high expected reward more often than the others. When constructing such a policy, one needs to consider the trade-off between exploration and exploitation. Exploitation is the act of choosing the arm that appears to be the best based on the observations made so far. Exploration on the other hand, refers to choosing a relatively less observed arm in an attempt to increase future rewards by decreasing uncertainty. In this paper, we are interested in pure exploration since our goal is to identify the Pareto optimal arms using as few samples as possible.

In its standard form, the MAB problem can be considered as a single-state reinforcement learning problem (although different bandit frameworks exist in which this correspondence might not hold, e.g., Restless Multi-Armed Bandit (RMAB) problem). At each round, the agent takes an action by pulling an arm based on past observations, and in turn, observes a reward sample (see Figure 1.1). The arms to pull are determined based on a policy which can be considered as a mapping from the past observations to the arm space.

Bandit problems are first introduced as a means for "sequential design of experiments" by Thompson [7,8] and Robbins [9]. They emphasized the importance of gathering data in an online manner, i.e., designing experiments on the fly based

on the data available so far, instead of collecting data all at once, without any analysis being made on the data until the data collection process is over. In particular, Thompson [7] suggests a strategy to guide the data collection process for medical trials where patients receive one of 2 alternative treatments. He argues that, based on the currently available data, if one of the treatments is deemed inferior and excluded in future trials, there is a possibility for all the consecutive patients to receive an inferior treatment if the rejected treatment was in fact, the superior one. On the other hand, if in the subsequent trials, the treatments are administered to the patients in a probabilistic manner based on their estimated probabilities of being superior to the other treatment, then, the expected number of patients that receive inferior treatment is smaller than that in the case of deciding on a fixed treatment and administering that in all future trials. [7, Sec. 1].

Later, in their seminal paper, Lai and Robbins [10] determined an asymptotic lower bound for the cumulative regret. Further, they proposed asymptotically optimal allocation rules (optimal policy) for common distributions such as Gaussian, Bernoulli, Poisson, and exponential, that give an expected cumulative regret that matches the lower bound.

MAB problems have been studied for almost a hundred years now. There are already hundreds of articles published in this field every year and the interest seems to keep growing [6, p. 1, p. 9]. Over the years, the MAB problem has been studied in many different contexts. Below, we will introduce some of them that are relevant to this study.

Best Arm Identification (BAI)

In this MAB framework, the goal is to find a 'good enough' approximation for the optimal arm with high probability (w.h.p.) by taking as few samples as possible. The accuracy of the approximation and the allowable failure rate of the algorithm is set by the user-defined accuracy parameters α and δ, respectively. Smaller α and δ values result in larger sample complexities. The effects of this trade-off should be taken into consideration when specifying these parameters.

BAI is a pure exploration problem where the aim is not to minimize the cumulative regret but to minimize the sample complexity, i.e., the total number of samples taken until termination. For instance, in [11, p. 2], the manufacture of cosmetic products is given as an example, where the product that is developed is subjected to a trial phase before the commercialization happens. In this case, it is desirable to explore different choices as much as possible until a satisfying product that meets the design requirements is discovered. The cost of making a choice that results in an unsatisfactory product is not important as long as the final product meets the criteria, which makes the BAI framework a better fit for this application.

Multi Objective Optimization in Bandits

Multi-objective MAB (MO-MAB) problem is another significant extension of the MAB setting where the aim is to optimize multiple, potentially conflicting objectives simultaneously. Often, trade-offs must be made between these conflicting objectives [12].

In general, there is no single optimal arm that achieves the highest scores in all objectives at the same time. Therefore, one should consider Pareto optimality in multi-objective optimization problems. Pareto optimality can be considered as a generalization of single objective optimality to multi-objective case. In MO-MAB case, an arm is considered Pareto optimal if there is no other arm that perform equally or better in all objectives. A mathematical definition of the Pareto optimality is given in Chapter 2. Alternatively, one can convert the multi-objective optimization problem to a single objective problem by weighing each objective with a scalar and then adding the weighed objectives to obtain a single scalar reward for each arm. However, this causes some of the Pareto optimal arms to appear suboptimal in the transformed single objective problem. In this work, the scalarization method is not considered, since our aim is to detect or closely approximate all the Pareto optimal arms without making any preferences for one Pareto optimal arm over the other.

Multi-objective optimization finds application in many fields such as finance, drug development, and neural network tuning where the aim is to optimize multiple performance metrics simultaneously. For instance, in a neural network tuning problem, training time and classification performance are conflicting objectives that need to be optimized simultaneously. Many studies experimented on the neural network tuning problem in the χ-armed bandit literature (bandit setting with infinitely many arms) where correlated arms are modeled with Gaussian process priors [13, 14].

In the single-objective case, the BAI framework is studied by Even-Dar et al. [2] which they refer to as the PAC (Probably Approximately Accurate) optimization framework. They propose PAC algorithms with $\mathcal{O}(\frac{1}{\alpha^2}\log\frac{1}{\delta})$ asymptotic sample complexity. They also prove a matching mini-max type lower sample complexity bound with the asymptotic complexity of $\Omega(\frac{1}{\alpha^2}\log\frac{1}{\delta})$.

In the multi-objective setting, the BAI problem can be generalized to what we call a "Pareto set Identification (PSI)" problem in which there are multiple optimal arms to be approximated accurately instead of a single optimal arm as in the single objective setting. In this work, we consider the PSI problem. We define the conditions for successfully approximating the Pareto set in Chapter 2.

Adversarial Attack

In practical applications, there are adversarial factors that can misguide the decision process by corrupting the reward samples. Many studies tried to integrate adversarial effects into analysis by proposing different types of attack models [15]. Some of the popular attack models considered in the literature can be listed as follows:

- Limited attack value: Attack intensity is subjected to upper limits. The attack occurs in every round.
- Budget constrained attack model: Attack value is limited with a certain

budget, i.e, the sum of the attack values over all rounds cannot exceed the designated budget.

- Bounded probability attack model: Attack can occur at every round with a certain probability. There are no restrictions on the intensity of the attack.

We consider 3 different adversarial models that fall under the bounded probability attack model category, namely oblivious, prescient and malicious adversaries introduced in [16]. In these attack models, at each round, with probability $\epsilon \in (0, \frac{1}{2})$, which we call as the attack probability, the sample from the reward distribution associated with the pulled arm is replaced with the sample that comes from the distribution chosen by the adversary, which we call as the contamination distribution. Otherwise, the agent observes the sample from the reward distribution directly (In the rest, we sometimes refer to the reward distribution as the true reward sample). Note that, each arm is associated with multiple reward distributions, each corresponding to a different objective. Therefore, at each arm pull, we observe a reward vector corresponding to the pulled arm. Note that this is in contrast to the single objective optimization case where a scalar reward sample is observed instead of a vector. Also note that the corruptions in the objectives are assumed to occur independently and therefore, the corruption of one objective does not necessarily imply the corruption of all the other objectives in the pulled arm. For instance, it is possible for some of the elements in the observed reward vector to be corrupted while others being true reward samples.

At each round, only the reward vector associated with the pulled arm is observed and no other information is available regarding the reward vectors of other arms. This is called the bandit feedback [5, p. 2].

Among the 3 adversaries mentioned above, the oblivious adversary is not aware of any reward outcomes and chooses the contamination distributions prior to the observation of rewards. The prescient and malicious adversary, on the other hand, know all the past and future reward outcomes and can strategically choose contamination distributions based on this knowledge to cause maximum disturbance to the learner. Different than the prescient adversary, a malicious adversary can

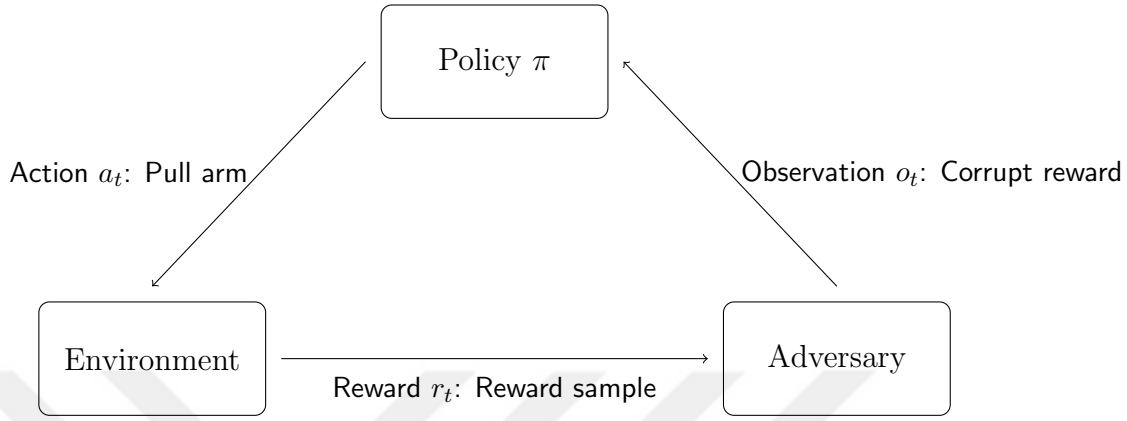


Figure 1.2: Reinforcement learning analogue of adversarial MAB setting

also correlate the attack probability with the reward distributions. For instance, in the case of an malicious adversary, the corruption probability might be larger than ϵ for some values of the true reward samples as long as the marginal distribution of the Bernoulli variable that determines whether an attack occurs or not has a bias of ϵ.

In conclusion, the malicious adversary is the strongest one and the oblivious adversary is the weakest, among these 3 adversaries. The rigorous definitions for each of them are given in Chapter 2.

Note that we put no restrictions on the contamination distributions or any kind of bound on the maximum value of the contamination. Also note that, unlike the reward distributions, the contamination distributions are not fixed and can change from round to round depending on the strategy of the adversary. The contamination distributions can also vary across arms and objectives. The general framework of the adversarial MAB problem is illustrated in Figure 1.2.

1.1 Related Work

In this section, we survey some related studies on the MAB and adversarial MAB problems.

Related work on MO-MABs

Multi-objective bandits are extensively studied in the literature for adversary free setting. Auer et al. [1] consider the Pareto set identification problem in the MO-MAB setting. They propose a confidence-bound algorithm and prove lower and upper bounds on the sample complexity. Ararat and Tekin [17] consider a generalized version of the Pareto set identification problem in the MO-MAB setting. They propose a method that allows for the specification of different preferences on the relative importance of the objectives by the practitioner. Drugan and Nowe [18] consider regret minimization problem in MO-MAB setting. They employ scalarization methods to turn the multi-objective problem into a single objective one. Yahyaa and Manderick [19] extend Thomson sampling to the MO-MAB problem and propose scalarization and partial ordering-based algorithms. Yahyaa et al. [20] adapts the Knowledge Gradient (KG) algorithm that was originally proposed for the single objective bandit problem to the MO-MAB setting.

In addition to the standard MO-MAB problem, there is also extensive study in the literature on multi-objective Gaussian Process (GP) bandits. In this setting, the mean reward vectors of the arms are assumed to be correlated with each other. This correlation is modeled with a Gaussian Process prior. The Gaussian bandits came into prominence with the seminal paper of Srinivas et al. [21] in which they proposed the GP-UCB algorithm for the standard MAB problem and proved information-theoretic regret bounds. Belakaria et al. [22] and Belakaria et al. [23] are other studies that propose methods that come with theoretical regret guarantees. Nika et al. [24] and Zuluaga et al. [25] proposed confidence bound algorithms for pure exploration problem and provided sublinear sample

complexity bounds. Also, there are studies that propose heuristic methods in GP bandit setting [26, 27].

Related Work on Adversarial MAB setting

Multi-armed bandits have found a broad range of applications such as medical treatment allocation [28]; financial portfolio design [29, 30]; adaptive routing [31]; news article recommendation [32], online advertising [33], and clinical trials [34]. Security concerns in these applications, motivated the studies on adversarial MABs that aim to develop robust methods against adversarial attacks. So far, adversarial MABs have attracted considerable attention [15]. A variety of adversary models are studied in the literature. A notable study that considers the bounded attack value model mentioned previously, is the exponential weight algorithm Exp3, proposed by Auer et al. [35] as a means for robust optimization. There are other studies that consider this type of attack model as well [36–38]. Lykouris et al. [39] proposed various robust algorithms in the budget constraint attack setting. Gupta et al. [40] provided an algorithm with improved regret bound.

The bounded attack probability model, which is another attack model mentioned before, is considered in [15, 16, 41]. Kapoor et al. [41] considered an adversarial model that is very similar to the oblivious adversary mentioned before. They proposed two variants of robust, median-based UCB type algorithm and provided asymptotic regret bounds. Guan et al. [15] considered an adversarial model that is similar to the prescient adversary but is weaker in the sense that it is assumed to know the past samples but not the future ones. Similar to [41], they considered cumulative regret minimization problem and provided optimality gap dependent asymptotic regret bounds. Altschuler et al. [16] on the other hand, considered 3 different adversarial attack models, namely oblivious, prescient and malicious that are also considered in this work. They propose a method in the BAI setting that is robust to the adversarial attacks. Note that in [41] and [15], the algorithms terminate when the number of samples exceed a certain budget.

On the other hand, the proposed algorithms in our work terminate automatically when a prediction that is guaranteed to meet the accuracy and confidence requirements can be returned.

1.2 Our Contribution

We study the MO-MAB problem in the adversarial setting. In many real-world multi-objective optimization problems, there are adversarial effects that cannot be modeled with a stochastic distribution. In these problems, robust methods are required to achieve accurate predictions. For instance, in a Neural Network tuning problem where the goal is to discover the Neural Network architectures that give near-optimal performance, there might be some exploding gradients in the training phase in one of the random initializations which result in poor prediction performance. If such corrupted iterations are not treated properly, some of the promising neural network architectures might be rejected early on by an online learning algorithm that is not robust to such effects. As another example, consider the blood glucose prediction application where we aim to accurately predict the blood glucose levels of type 1 diabetes patients throughout a day. Accurate blood glucose predictions are important for keeping patient's blood glucose levels within the acceptable range. This is achieved by finding the right amount of insulin dose injections based on the blood glucose level predictions. However, there are environmental factors such as insomnia, stress and physical activity of the patient that influence the blood glucose levels in an unpredictable way. Therefore, the accurate prediction of the blood glucose is only possible with a robust method that takes these adversarial affects into account.

Motivated by these examples, we formulate the Pareto set identification problem under unbounded adversarial attack models and propose two robust methods that comes with high probability accuracy guarantees.

Although the adversarial setting is extensively studied in the single objective MAB setting, the extension of robust optimization to MO-MABs offers some

unique challenges which are elaborated below:

- The mean-based optimization methods developed for the classical, adversary-free MO-MAB setting drastically fail when faced with an unbounded adversarial attack. This prohibits using the mean statistic under the type of adversaries considered in this work. A new statistic has to be employed which is robust to such adversarial effects.

- Because of the unboundedness of the adversarial attacks considered in this work, the empirical estimation of the median has a fixed inherent bias that cannot be decreased no matter how much sampling data is collected from the reward distributions.

- As mentioned in [16, Remark 4], "additive property of suboptimality" does not apply in the adversarial case. This calls for special design choices in the algorithm. Below, we state the additive property for single objective optimization case:
  "Suppose that the suboptimality gap of the arm j with respect to arm i is denoted by $\Delta_{j,i}$ and the suboptimality gap of i with respect to the optimal arm is denoted by Δ_i. Then the suboptimality gap of arm j with respect to the optimal arm is given by $\Delta_j = \Delta_i + \Delta_{j,i}$."
  For single objective case, $\Delta_{j,i} = \max\{0, \mu_i - \mu_j\}$ where μ_i and μ_j are the mean of the reward distributions associated with arm i and j, respectively. The definition of suboptimality gap in the multi-objective case will be provided in Chapter 2. A similar additive property is also valid for the multi-objective case, as a generalization of the additive property in the single objective case. Similar to the adversarial single objective case, additive property in multi-objective case breaks down in the adversarial setting as well. This presents a new challenge in the way of Pareto set approximation. In particular, it must be ensured that none of the arms in the Pareto set are eliminated until the final round.

We address the above challenges as follows:

- We turn to median-based optimization approach which is proved to be robust against the adversarial attacks. In particular, we propose a median-based method that achieves the maximum possible prediction accuracy that can be guaranteed under the adversarial setting considered in this work (see Remark 1).

- Since the additive property of suboptimality does not apply in the adversarial setting as mentioned above, in our algorithm design, we make sure that none of the Pareto optimal arms are eliminated until the last round by designing a strict elimination condition (see Chapter 3). Note that in the adversary-free multi-objective setting, since the additive property of suboptimality holds in the adversary-free case, it is not necessary to abide by this restriction.

Finally we summarize our contributions below:

1. We propose a robust algorithm that returns, w.h.p., a prediction that satisfies the accuracy requirements.

2. We show that in the adversarial MO-MAB setting, no algorithm can be guaranteed to return a prediction that is better than a certain limit which depends on the adversary, the environment under consideration, and the attack probability ϵ. An algorithm with an accuracy guarantee that can be made arbitrarily close to this limit is called a Pareto accurate algorithm (In addition, a Pareto accurate algorithm should satisfy a coverage condition, see Section 2.6 for more details). We prove that the proposed method is Pareto accurate.

3. We provide a sample complexity bound for the proposed algorithm that depends on the accuracy parameter α inverse squarely.

4. We provide a variation of the proposed algorithm which is guaranteed to return all the Pareto optimal points (full coverage). For this algorithm, we prove an accuracy guarantee that is looser than that given for the original

method. We also prove that the same sample complexity bound proved for the original method holds for this algorithm as well (see Chapter 4).

5. By considering Gaussian reward distributions, we show that our sample complexity bound has the same accuracy dependence (α parameter dependence) as Algorithm 1 of [1]. Although Algorithm 1 is mean-based, the equivalence of mean and median for Gaussian distributions makes comparison possible.

6. Our experiments show that the proposed algorithms satisfy our theoretical predictions in the adversarial setting. Also, the results show that the proposed algorithms make accurate predictions in the adversary-free setting as well, with a sample complexity comparable to Algorithm 1 of [1]. Algorithm 1 fails drastically for large values of attack probability which shows the need for robust algorithm design in the adversarial setting.

1.3 Comparison with the related work

In the literature, a single objective MAB problem is studied under various attack models. In some of these studies, the goal is to identify the arm with the highest first-order statistic (mean). This can only be justified when the attack model is assumed to be bounded in value, since, otherwise, the mean cannot be estimated accurately from the samples contaminated with an attack that is large in value. In many applications, however, it is more plausible to use adversarial models that are restricted in attack probability rather than the attack value. For instance, in the neural network corruption example mentioned previously, the adversarial effect is more suitable for a bounded probability attack model rather than a bounded attack value model since the gradient explosion problem usually alters the performance of the network in an unpredictable way.

The median statistic is shown to be robust for the bounded probability attack model even if the contaminations are unbounded in value [16]. In this work, we

Table 1.1: *Comparison with related work.*

Work	Setup	Goal	Adversary	Bound type
[1]	MO-MAB	Pareto Set Id.	Adv. free	Sample comp.
[16]	Single obj. MAB	Best arm Id.	Obl., Presc., Mal.	Sample comp.
[42]	Single obj. MAB	Cum. regret min.	Bounded attack	Cum. regret
[18]	MO-MAB	Cum. regret min.	Adv. free	Cum. regret
[24]	MO GP bandit	Pareto Set Id.	Adv. free	Sample comp.
[25]	MO GP bandit	Pareto Set Id.	Adv. free	Sample comp.
[41]	Single obj. MAB	Cum. regret min.	Oblivious	Cum. regret
[15]	Single obj. MAB	Cum. regret min.	Prescient like	Cum. regret
[39]	Single obj. MAB	Cum. regret min.	Prescient like	Cum. regret
This work	**MO-MAB**	**Pareto Set Id.**	**Obl., Presc., Mal.**	**Samp. comp.**

also adapt the median-based approach to achieve robust optimization in MO-MAB setting.

The MO-MAB setting is also studied extensively in the literature. In this setting, the mean (median) reward vectors of each arm is assumed to be independent of each other. In other words, the arms are assumed to be 'uncorrelated'. However, the MO-MAB framework can be employed when the arms are correlated as well, since the same theoretical results proved for the uncorrelated arms still hold in this case. It is possible, however, to obtain better sample complexity performance if one formulates the problem in a framework that utilizes the correlation information (if it is available) between the arms. Some studies model the correlation by assuming that the mean rewards are sampled from a Gaussian Process where the correlations between the arms are determined based on their proximity to each other in the design space (input space). Quantitatively, correlation is determined with a function known as a kernel function that maps the distances in the input space to scalar values. Note that in these studies, it is assumed that the kernel is exactly known and is a good fit for the real correlation between the arms.

In this work, the MO-MAB model is adopted with the goal of approximating the Pareto optimal set by taking as few samples as possible from the reward

distributions. As far as we are aware, this is the first work in the literature that considers adversarial attack in the MO-MAB setting. The comparison of this work with the other studies from the literature is summarized in Table 1.1.

The remainder of the thesis is organized as follows. In Chapter 2, we introduce the notation, give some preliminaries, define the adversarial attack models considered in this work in detail and give the accuracy and coverage conditions for Pareto accurate algorithms. In Chapter 3, a Pareto accurate algorithm is introduced and an upper sample complexity bound is derived for this algorithm. In Chapter 4, a variation of the Pareto accurate algorithm proposed in Chapter 3 is presented and the theoretical analysis of this new algorithm is given. In Chapter 5, the numerical results are presented where 2 algorithms proposed in this work are compared with the Algorithm 1 of [1]. Finally, in Chapter 6, concluding remarks are given.

Chapter 2

Problem Formulation

2.1 Notation

The Bernoulli distribution with bias $\rho \in [0, 1]$ is denoted by $\text{Ber}(\rho)$. The set of positive integers are denoted by $\mathbb{N}_+$. For the set $\{1, ..., n\}$, we use the short hand notation $[n]$ where $n \in \mathbb{N}$. Similarly, for $n \in \mathbb{Z}_+$, where $\mathbb{Z}_+$ denotes the positive integers, the notation $[a_i]_{i=1}^n$ is used to denote the set $\{a_1, \cdots a_n\}$ that consists of the elements of a finite sequence $a_1, \cdots, a_n$. The short hand notation $[a \pm b]$ is used to denote the interval $[a - b, a + b]$. If A is a set, $|A|$ denotes the cardinality of A.

Let F represent the cumulative density function (c.d.f) of a random variable X, i.e, $X \sim F$. The left and right quantile functions of X are denoted by $Q_{R,F}(p) \coloneqq \inf\{x \in \mathbb{R} : F(x) > p\}$ and $Q_{L,F}(p) \coloneqq \inf\{x \in \mathbb{R} : F(x) \geq p\}$ for $p \in [0, 1]$ respectively. The left and right quantile functions differ when F is not a strictly increasing function. Due to this, for some probability $p \in [0, 1]$, there are multiple values of x such that $F(x) = p$. A demonstration for such a case where the right and left quantiles take different values for some p is illustrated in Figure 2.1.

The set of medians is denoted by $m(F) \coloneqq [Q_{L,F}(\frac{1}{2}), Q_{R,F}(\frac{1}{2})]$. When the median is unique, for convention, we assume that $m(F)$ denotes the unique median instead of the singleton set containing it.

The empirical median of a sequence of samplings $x_1, \ldots, x_n \in \mathcal{R}$ is denoted as $\hat{m}(x_1, \ldots, x_n)$. If n is odd, this corresponds to the middle value in the sequence and if n is even, it corresponds to the average of the two middle values.

Suppose that x is an M dimensional vector. The ith element of x is denoted by x^i. Consider another M dimensional vector y. The relation $x \prec y$ denotes that the vector x is *dominated* by vector y. Mathematically, this is defined as:

$$\forall i \in [M] : x^i \leq y^i \quad \wedge \quad \exists k \in [M] : x^k < y^k \ .$$

Also, by $x \preceq y,$ we denote that x is *weakly dominated* by y. Mathematically:

$$\forall i \in [M] : x^i \leq y^i \ .$$

Note that if x is dominated by y, there exists at least one objective in which y is strictly better than x. On the other hand, in the case of weak domination, $x \preceq y$ holds when x and y vectors are equal as well. In addition, we use the notation $x \npreceq y$ to denote that x is *not weakly dominated* by y. By definition of weak domination, this is equivalent to:

$$\exists i \in [M] : x^i > y^i.$$

2.2 Adversarial Pareto Set Identification Problem

In this work, multi-objective optimization problem with M objectives and a finite number of arms indexed by $i = 1, \ldots, K$ is considered. The cdf of the reward distributions are denoted by $\{F_i^d\}_{i \in [K], d \in [M]}$, contamination distributions

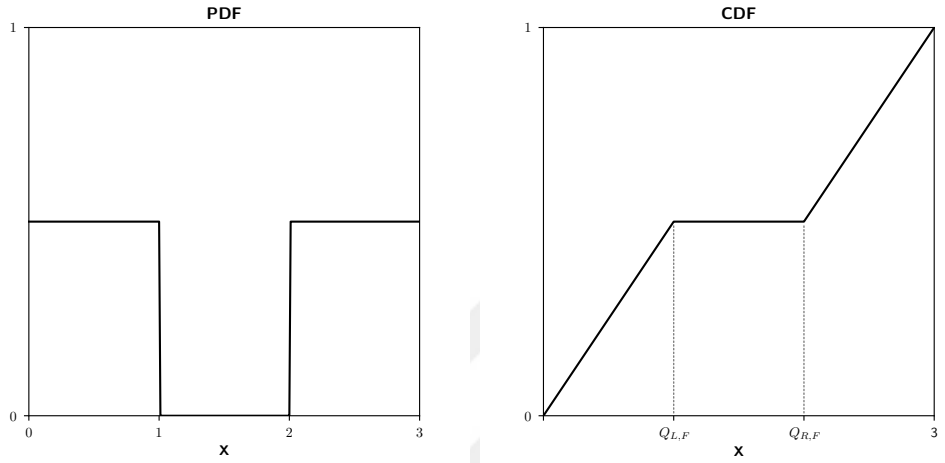


Figure 2.1: An example case where $Q_{R,F}$ and $Q_{L,F}$ take different values at $p = 0.5$. Any $x \in [Q_{L,F}(\frac{1}{2}), Q_{R,F}(\frac{1}{2})]$ is a median of this distribution.

by $\{G^d_{i,t}\}_{i\in[K],d\in[M],t\in\mathbb{N}}$, and the corresponding random variables by $Y^d_{i,t} \sim F^d_i$ and $Z^d_{i,t} \sim G^d_{i,t}$. Note that the contamination distributions depend on the time step t since the adversary is free to pick a new distribution at every time step. The contamination probability (attack probability) $\epsilon \in (0, \frac{1}{2})$ on the other hand is fixed across the arms, objectives, and time steps.

The Bernoulli random variable corresponding to arm i and objective d that determines whether a contamination occurs at a time step t on the reward distribution that corresponds to arm i and objective d, is denoted by $B^d_{i,t} \sim \text{Ber}(\epsilon)$. The contaminated reward distribution at time step t pertaining to arm i and objective d is denoted by $\tilde{F}^d_{i,t}$ and the corresponding random variable by $\tilde{Y}^d_{i,t}$. If contamination occurs, the observed reward is replaced with the sample from the contamination distribution. Otherwise, the sample from the true distribution is observed at the output. Formally:

$$\tilde{Y}^d_{i,t} = \begin{cases} Y^d_{i,t} & \text{if } B^d_{i,t} = 0 \ . \\ Z^d_{i,t} & \text{if } B^d_{i,t} = 1 \ . \end{cases} \tag{2.1}$$

Note that the random variable $\tilde{Y}_{i,t}^d$ is equivalent to $(1 - B_{i,t}^d)Y_{i,t}^d + B_{i,t}^d Z_{i,t}^d$ in distribution. Also note that if $B_{i,t}^d$, $Y_{i,t}^d$, and $Z_{i,t}^d$ are independent from each other (which is true for the oblivious and prescient adversaries), then:

$$\tilde{F}_{i,t}^d(\tilde{y}) = (1-\epsilon)F_i^d(\tilde{y}) + \epsilon G_{i,t}^d(\tilde{y}),\ \forall \tilde{y}\ .$$

Note that, from time to time, we drop the time index t, arm index i or objective index d (or all of them together) from the random variables and distributions defined above when their specification is not necessary in the analysis or a general argument that holds for all t, i and d is presented. The *median of interest* m_i^d, corresponding to cdf F_i^d, is defined as the average of right and left $\frac{1}{2}$-quantiles of F_i^d, i.e:

$$m_i^d = \frac{Q_{R,F_i^d}(\frac{1}{2}) + Q_{L,F_i^d}(\frac{1}{2})}{2}\ .$$

Note that if F_i^d has a unique median, m_i^d is equivalent to this median. Also, the *median of interest vector* of arm i which is denoted by m_i is defined as the M dimensional vector whose elements correspond to the median of interest of each objective in arm i. Note that this definition of median of interest is purely for convention, and in fact, any median can be picked as the median of interest.

Given K arms, we define the Pareto optimal set P^* based on the median of interest vectors as follows:

$$P^* = \{i \in [K] :\ \nexists j \in [K] \setminus \{i\} : m^i \prec m^j\}\ .$$

Note that in the case when there exist one or multiple reward distributions with non-unique medians, the above definition of median of interest prevents any ambiguity in the definition of P^*.

A detailed notation table that contains the symbols introduced throughout this work is provided in Appendix A.

A typical MO-MAB problem with 10 arms and 2 objectives is demonstrated in Figure 2.2 where each point corresponds to the mean reward vector of an arm. Although arms such as 2 and 3 are not optimal in any of the objectives, they are

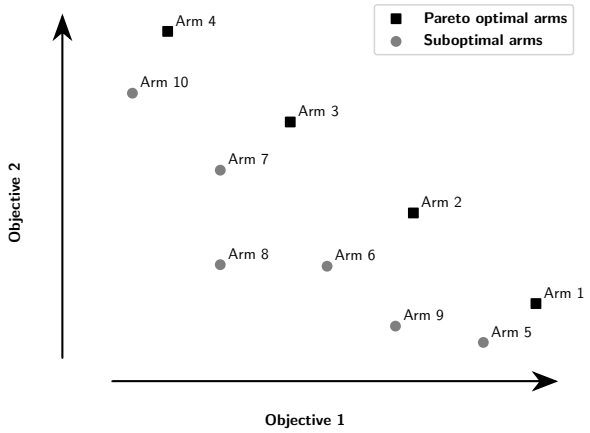


Figure 2.2: Mean reward vectors of a MO-MAB problem with $K = 10$ and $M = 2$ are represented on a 2D plot.

Pareto optimal arms since none of the other arms score higher than these arms on both objectives. This combinatorial nature of the multi-objective optimization (MOO) problem makes it more challenging than the single-objective case.

A Pareto set approximation algorithm stops after conducting a series of sequential evaluations of arms in $[K]$ with the aim of returning a predicted Pareto set P that approximates P^* in accordance with the accuracy and coverage conditions defined in Section 2.6. We note that an algorithm that makes as few samplings as possible is desirable.

2.3 Attack Models

Three attack models are considered, which are given in the order from the weakest to the strongest below in terms of adversarial power.

Oblivious. Oblivious adversary chooses all the contamination distributions apriori without the knowledge of the arm outcomes (Y) or without the knowledge

of the round t, arm i and objective d triples (t, i, d), in which the samples are corrupted, i.e., in which $B_{i,t}^d = 1$. Formally, for all $i \in [K]$ and all $d \in [M]$, $\{(Y_{i,t}^d, Z_{i,t}^d, B_{i,t}^d)\}_{t\geq 1}$ triples are independent. Furthermore, for a fixed t, $Y_{i,t}^d$, $B_{i,t}^d$, and $Z_{i,t}^d$ are independent for all i and d. In practice, a motivating example for an oblivious adversary could be occasionally occurring measurement errors in an experiment setup.

Prescient. The adversary can choose contamination distributions based on all the past and future true arm outcomes. Prescient adversary also knows all the round t, arm i and objective d triples (t, i, d) in which the samples are corrupted ($B_{i,t}^d = 1$), and it can adapt its strategy accordingly. Formally, for all $i \in [K]$ and $d \in [M]$, the pairs $\{(Y_{i,t}^d, B_{i,t}^d)\}_{t\geq 1}$ are independent. Furthermore, for a fixed t, $Y_{i,t}^d$ and $B_{i,t}^d$ are independent for all i and d and $Z_{i,t}^d$ may depend on all the realizations $\{Y_{j,s}^d, B_{j,s}^d, Z_{j,s}^d\}_{j\in[K],d\in[M],s\geq 1}$. The prescient adversary model can be a good fit for an adversary that tries to corrupt the samples of a dataset in a deliberate manner.

Malicious. Same as the prescient adversary except that the malicious adversary can also couple the probability of attack with the true reward outcomes. Formally, for all $i \in [K]$ and for all $d \in [M]$, $\{(Y_{i,t}^d, B_{i,t}^d)\}_{t\geq 1}$ pairs are independent and $Z_{i,t}$ may depend on all $\{Y_{j,s}^d, B_{j,s}^d, Z_{j,s}^d\}_{j\in[k],d\in[M],s\geq 1}$. Note that, for malicious adversary, $Y_{i,t}^d$, $B_{i,t}^d$ pairs may be dependent for a fixed t since they can be coupled by the malicious adversary. For instance, a hacker attack on a dataset that targets the data points based on their value could be a good fit for a malicious adversary model. The hacker might want to concentrate on corrupting the key data-points whose corruption would cause the largest disturbance to the integrity of the data-set. As a result, the corruption probability of these key points are higher compared to the other points.

In all three attack models, the random variables corresponding to the true arm reward distributions are allowed to be correlated with each other, i.e., for all $i, j \in [K]$ and $a, b \in [M]$, the random variables $Y_{i,n}^a$ and $Y_{j,n}^b$ can be correlated.

2.4 Median Concentration and Unavoidable Bias

Because our adversarial attack model allows for arbitrary contamination distributions, the mean statistic cannot be accurately predicted from the contaminated samples. Furthermore, the median statistic is subject to an unavoidable bias which makes the median identifiable only up to a certain interval. In the following, we determine this unavoidable bias. First, the classification of the distributions according to the behaviour of their right and left quantile functions around the median is given below.

Definition 1. *[16, Definition 5]. For any $\bar{t} \in (0, \frac{1}{2})$ and positive, non-decreasing function R defined on the range $[0, \bar{t}]$, we define $C_{R,\bar{t}}$ to be the family of distributions such that for a cdf F that belongs to $C_{R,\bar{t}}$ the following holds:*

$$R(t) \geq \max\{Q_{R,F}(\frac{1}{2} + t) - m, m - Q_{L,F}(\frac{1}{2} - t) \ , \tag{2.2}$$

$\forall t \in [0, \bar{t}]$ and for any $m \in m(F)$.

In words, R bounds the maximum quantile deviation that can occur from the median for each given t. Note that the above definition captures the relevant features of the reward distributions that can be used in the analysis. Also note that a common R and $\bar{t}$ is chosen for all reward distributions $\{F_i^d\}_{i,d}$ to facilitate the analysis.

Example 1. For subgaussian distributions with parameter σ, (2.2) holds $\forall t \in [0, \frac{1}{2})$ and $\forall m \in m(F)$, for the R given below:

$$R(t) = \sigma\sqrt{2}\left(\sqrt{\log\left(\frac{1}{1/2 - t}\right)} + \sqrt{\log(2)}\right) \ . \tag{2.3}$$

Proof. Note that given a random variable X, if $X - \mathbb{E}(X)$ is Subgaussian with

parameter σ, then the following concentration inequalities hold:

$$\mathbb{P}\big(X - \mathbb{E}[X] \geq q\big) \leq \ e^{\frac{-q^2}{2\sigma^2}} \ , \tag{2.4}$$

$$\mathbb{P}\big(X - \mathbb{E}[X] \leq -q\big) \leq \ e^{\frac{-q^2}{2\sigma^2}} \ , \tag{2.5}$$

for $q > 0$. Also, by definition of the right quantile function:

$$\mathbb{P}(X < Q_{R,F}(1/2 + t)) \leq 1/2 + t, \ \forall t \in (0, \frac{1}{2}) \ . \tag{2.6}$$

Note that, $\mathbb{P}(X \leq Q_{R,F}(1/2 + t)) \geq 1/2 + t, \ \forall t \in (0, \frac{1}{2})$ and the equality holds in this case and in the above inequality if the c.d.f. of X is continuous at $Q_{R,F}(1/2 + t)$. In the rest of the proof, let $q_t := Q_{R,F}(1/2 + t)$. In the following, we will derive a different upper bound on $(q_t - m)$ for each case given below:

1. $\mathbb{E}[X] \leq m \leq q_t$,

2. $m \leq \mathbb{E}[X] \leq q_t$,

3. $m \leq q_t \leq \mathbb{E}[X]$.

Considering that $(q_t - \mathbb{E}[X]) > 0$ for the first case, by (2.4) and (2.6):

$$e^{\frac{-(q_t - \mathbb{E}[X])^2}{2\sigma^2}} \geq \mathbb{P}\big(X - \mathbb{E}[X] \geq q_t - \mathbb{E}[X]\big) = \mathbb{P}\big(X \geq q_t\big) \geq 1 - (1/2 + t) \ .$$

Simplifying above gives:

$$q_t \leq \mathbb{E}[X] + \sqrt{2\sigma^2 \log\big(\frac{1}{1/2 - t}\big)} \ . \tag{2.7}$$

Further, considering that $m \geq \mathbb{E}[X]$, we obtain the following bound on $(q_t - m)$ for the first case:

$$q_t - m \leq q_t - \mathbb{E}(X) \leq \sqrt{2\sigma^2 \log\big(\frac{1}{1/2 - t}\big)} \ .$$

For the second case, note that (2.7) is valid since $q_t \geq \mathbb{E}[X]$. Also, by (2.5), and the definition of the median, we can obtain a bound on $(\mathbb{E}[X] - m)$ as follows:

$$1/2 \leq \mathbb{P}(X \leq m) = \mathbb{P}(X - \mathbb{E}[X] \leq m - \mathbb{E}[X]) \leq e^{\frac{-(\mathbb{E}[X]-m)^2}{2\sigma^2}} .$$

Hence:

$$1/2 \leq e^{\frac{-(E[X]-m)^2}{2\sigma^2}} .$$

By simplifying the above, we obtain:

$$\mathbb{E}[X] - m \leq \sqrt{2\sigma^2 \log 2} . \tag{2.8}$$

Combining this with (2.7) we obtain:

$$q_t - m \leq \sqrt{2\sigma^2 \log \big(\frac{1}{1/2 - t}\big)} + \sqrt{2\sigma^2 \log 2} .$$

For the third case, considering that (2.8) is valid and $q_t \leq \mathbb{E}(X)$, we obtain:

$$q_t - m \leq \sqrt{2\sigma^2 \log 2} .$$

Combining the results, we obtain the following bound on $(q_t - m)$ that holds for all 3 cases:

$$q_t - m \leq \sqrt{2\sigma^2 \log \big(\frac{1}{1/2 - t}\big)} + \sqrt{2\sigma^2 \log 2} .$$

Through similar steps, the following bound can be obtained on $(m - q_t)$ as well:

$$m - q_t \leq \sqrt{2\sigma^2 \log \big(\frac{1}{1/2 - t}\big)} + \sqrt{2\sigma^2 \log 2} .$$

□

Below we restate two results from [16] that allows for the determination of the unavoidable bias.

Lemma 1. ***Concentration of empirical median around the true median for prescient and oblivious adversaries.*** *[16, Lemma 7]. Let* $\bar{t} \in (0, \frac{1}{2})$, $\epsilon \in (0, \frac{2\bar{t}}{1+2\bar{t}})$, $\delta \in (0,1)$ *and* $F \in C_{\bar{t},R}$. *Let* $Y_i \sim F$ *and* $B_i \sim Ber(\epsilon)$ *all be independently drawn for* $i \in [n]$. *Let* $\{Z_i\}_{i\in[n]}$ *be arbitrary random variables possibly depending on* $\{Y_i, B_i\}_{i\in[n]}$, *and* $\tilde{Y}_i = (1-B_i)Y_i + B_i Z_i$. *Then for* $n \geq 2(\bar{t} - \frac{\epsilon}{2(1-\epsilon)})^{-2}\log(\frac{2}{\delta})$*:*

$$\mathbb{P}\Big(\sup_m |\hat{m}(\tilde{Y}_1, ..., \tilde{Y}_n) - m| \leq R\Big(\frac{\epsilon}{2(1-\epsilon)} + \sqrt{\frac{2\log(2/\delta)}{n}}\Big)\Big) \geq 1-\delta \ . \tag{2.9}$$

Lemma 2. ***Concentration of empirical median around the true median for malicious adversary.*** *[16, Lemma 8]. Let* $\bar{t} \in (0, \frac{1}{2})$, $\epsilon \in (0, \bar{t})$, $\delta \in (0,1)$ *and* $F \in C_{\bar{t},R}$. *Let* (Y_i, B_i) *pairs be independently drawn for* $i \in [n]$ *with marginals* $Y_i \sim F$ *and* $B_i \sim Ber(\epsilon)$. *Let* $\{Z_i\}_{i\in[n]}$ *be arbitrary random variables possibly depending on* $\{Y_i, B_i\}_{i\in[n]}$, *and* $\tilde{Y}_i = (1-B_i)Y_i + B_i Z_i$. *Then for* $n \geq 2(\bar{t} - \epsilon)^{-2}\log(\frac{3}{\delta})$*:*

$$\mathbb{P}\Big(\sup_m |\hat{m}(\tilde{Y}_1, ..., \tilde{Y}_n) - m| \leq R\Big(\epsilon + \sqrt{\frac{2\log(3/\delta)}{n}}\Big)\Big) \geq 1-\delta \ . \tag{2.10}$$

By taking the limit $n \to \infty$ in (2.9) and (2.10), we determine the unavoidable bias term D for oblivious and prescient adversaries:

$$D = R(\frac{\epsilon}{2(1-\epsilon)}) \ , \tag{2.11}$$

and for malicious adversary:

$$D = R(\epsilon) \ . \tag{2.12}$$

Note that the above lemmas hold for the median of interest since the median of interest is also a median of the given distribution. In the rest of the paper, we will simply refer to the median of interest as the median and the median of interest vector as the median vector.

2.5 Suboptimality in the Multi-Objective Setting

Below, we restate the suboptimality gap definition of Auer et al. [1], which is a measure of the amount that an arm is dominated by the Pareto set.

Definition 2. *The suboptimality gap of an arm i with respect to arm j is defined as follows:*

$$\Delta_{i,j} := \max\left\{0, \min_d (m_j^d - m_i^d)\right\} \ .$$

Based on the above, the suboptimality gap of arm i is defined as follows:

$$\Delta_i := \max_{j \in P^*} \Delta_{i,j} \ .$$

Given a positive real number α, an arm i is called α-suboptimal if its suboptimality gap is smaller than α, i.e, $\Delta_i \leq \alpha$. Note that in the rest of the paper, the arms that are not Pareto optimal is referred as suboptimal which should not be confused with the α-suboptimality defined above. For instance, all the Pareto optimal arms are α-suboptimal for any positive real number α since $\Delta_j = 0$ for a Pareto optimal arm j.

In general, it is not possible to detect Pareto optimal arms with more than $2D$ accuracy due to the unavoidable bias, as proven in the following remark.

Remark 1. *Suppose that an adversary can alter a reward sample as much as D. Then, there are bandit environments in which it is impossible to distinguish Pareto optimal arms from the suboptimal arms with suboptimality gaps smaller than $2D$.*

Proof. Consider the simple environment with only 2 arms. Suppose that arm 1 is Pareto optimal and arm 2 is suboptimal with $\Delta_2 = \Delta_{2,1} < 2D$. Then, by definition of suboptimality gap:

$$\exists d \in [M] : m_1^d - m_2^d < 2D \ .$$

Adversary can manipulate arms so that, as $n \to \infty$, $\hat{m}_1^d = m_1^d - D$ and $\hat{m}_2{}^d = m_2^d + D$, in which case, $\hat{m}_2^d > \hat{m}_1^d$ and hence $\hat{m}_2 \not\prec \hat{m}_1$. □

2.6 Pareto Accuracy

In the following, we define the class of algorithms that is of interest in the adversarial MO-MAB setting.

Definition 3. ***Pareto Accurate Algorithm:*** *Suppose that the reward distributions belong to $C_{R,\bar{t}}$ for some $\bar{t} \in (0, \frac{1}{2})$. Then, given the accuracy parameter $\alpha \geq 0$, the confidence probability $0 < \delta < 1$ and the adversarial attack probability $\epsilon \in (0, \frac{2\bar{t}}{1+2\bar{t}})$ ($\epsilon \in (0, \bar{t})$ if malicious adversary), we call an algorithm Pareto accurate in the adversarial MO-MAB setting if the set of arms P that the algorithm returns at termination satisfy the following conditions:*

1. *Accuracy: All the returned arms are at least $(2D + \alpha)$ suboptimal:*

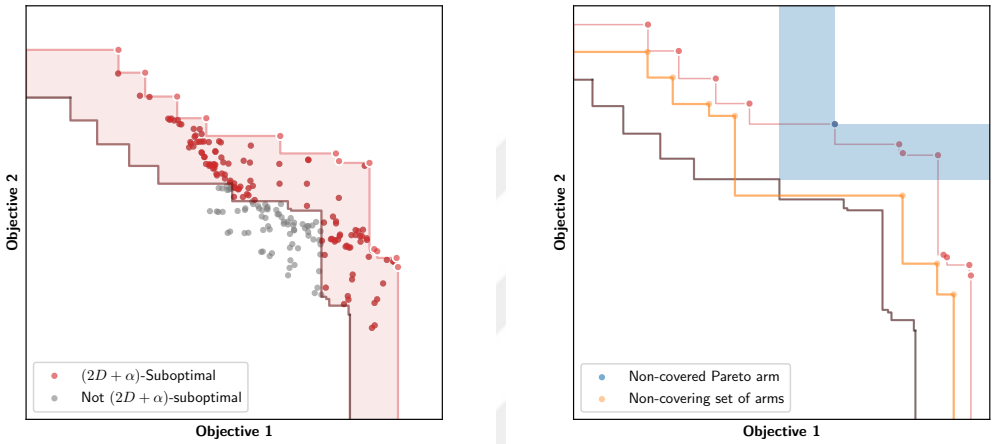
$$\forall i \in P,\ \Delta_i \leq 2D + \alpha \ .$$

2. *Coverage: If a Pareto optimal arm is not returned in the predicted set, then, there exists at least one arm in the predicted set that $(2D + \alpha)$-covers the eliminated Pareto optimal arm, i.e.:*

$$\forall j \in P^*,\ \exists i \in P : m_j - m_i \preceq 2D + \alpha \ .$$

In Figure 2.3, we demonstrate accuracy and coverage arguments on a 2-objective example. In 2.3a, the red area represents the region that $(2D + \alpha)$-suboptimal arm can be in, i.e., the arms that are below the red area are not $(2D + \alpha)$-suboptimal. Therefore, w.h.p., a Pareto accurate algorithm returns arms only from the red region. In 2.3b, the blue area shows the region that a $(2D + \alpha)$-covering arm of the blue Pareto point can be in. A Pareto accurate algorithm returns at least one arm from the blue area, w.h.p., in order to cover the

blue Pareto point. Orange points in the plot sets an example for a non-covering set of arms for the Pareto set, since the Pareto arm represented by the blue point is not covered.



(a) (b)

Figure 2.3: Demonstration of accuracy (2.3a) and coverage (2.3b) requirements in a 2-objective bandit environment. In the plots, each point corresponds to the median reward vector of an arm. In (2.3a), $(2D+\alpha)$-suboptimal arms are shown with red color and the arms that are not $(2D+\alpha)$-suboptimal is shown with grey. The light red-colored area represents the region where a $(2D+\alpha)$-suboptimal arm can be situated. In (2.3b), blue region shows the coverage region for the blue Pareto point (Note that the area that is to the upper right of the blue point is not included in the coverage region since there cannot be any arm in that region if the blue point is Pareto optimal. Otherwise, blue point would have been suboptimal). Orange points form a set that covers all the Pareto optimal arms except for the Pareto point shown in blue. In addition, all the orange points are $(2D+\alpha)$-suboptimal. This means that, if an arm had been included from the blue region as well, the orange set would have satisfied the conditions given in Definition 3.

The Pareto accuracy can be considered as the generalization of the PAC concept from the single objective MAB setting to the adversarial MO-MAB setting. However, in the adversarial MO-MAB setting, since the Pareto optimal arms cannot be distinguished from other $(2D)$-suboptimal arms as shown in the Remark 1,

it is not possible to approximate the Pareto set arbitrarily well. This restriction reflects itself in the the accuracy and the coverage conditions given in Definition 3 as the addition of $2D$ term.

Chapter 3

Algorithm

In this section, a Pareto accurate algorithm called *Robust Pareto Set Identification* (R-PSI) is presented whose flowchart and pseudocode is given in Algorithm 1 and Figure 3.1 respectively. The Pareto accuracy of this algorithm will be proved in Section 3.1 as well as an sample complexity upper bound. The procedure of R-PSI is described as follows: At the beginning, all the arms are assigned to an *undecided set* denoted by S. Then algorithm starts by sampling each arm n_0 times, given by the following expression:

$$n_0 = \left\lceil 2\beta_{\bar{t},\epsilon} \log\left(\frac{\pi^2 MK}{6\tilde{\delta}}\right)\right\rceil \ . \tag{3.1}$$

where

$$\beta_{\bar{t},\epsilon} = (\bar{t} - \frac{\epsilon}{2(1-\epsilon)})^{-2} \ , \tag{3.2a}$$

$$\tilde{\delta} = \delta/2 \ , \tag{3.2b}$$

for the prescient and oblivious adversaries. For the malicious adversary, they are given as follows:

$$\beta_{\bar{t},\epsilon} = (\bar{t} - \epsilon)^{-2} \ , \tag{3.3a}$$

$$\tilde{\delta} = \delta/3 \ . \tag{3.3b}$$

Then, the algorithm computes the empirical median. Also, the statistical bias is computed according to (3.8) before entering the main loop. At time steps $t \geq 1$, algorithm selects the arm with the largest statistical bias (largest U). This arm is denoted by i^*. The selected arm is successively sampled $n_{\tau_{i^*}}$ times. For $\tau \geq 2$, n_τ is defined as follows:

$$n_\tau = \left\lceil 4\tau\beta_{\bar{t},\epsilon} \log\left(\frac{\tau}{\tau-1}\right) + 2\beta_{\bar{t},\epsilon} \log\left(\frac{(\tau-1)^2 MK\pi^2}{6\tilde{\delta}}\right) \right\rceil \ , \tag{3.4}$$

where τ denotes the number of sampling rounds that an arm has gone through (Generally, we use the arm indices as a subscript in τ. For instance, τ_i denotes the sampling round of arm i). This should not be confused with the total number of samplings made from this arm since each sampling round consists of multiple samplings of the selected arm.

As will be shown in Lemma 3, n_τ is chosen in such a way that the sample complexity requirements of Lemma 1 and Lemma 2 are met. The median deviation bound of Lemma 1 and Lemma 2 corresponding to arm i^* is given by:

$$R(h_\epsilon + \frac{1}{\sqrt{\beta_{\bar{t},\epsilon}\tau_{i^*}}}) \ , \tag{3.5}$$

where

$$h_\epsilon = \frac{\epsilon}{2(1-\epsilon)} \ , \tag{3.6}$$

for the oblivious and malicious adversaries and:

$$h_\epsilon = \epsilon \ , \tag{3.7}$$

for the malicious adversary. We note that, (3.5) is obtained by calculating a lower bound on the total number of samplings made from arm i^*, which is equal to $\left(n_0 + \sum_{t=2}^{\tau_{i^*}} n_t\right)$, and putting this in the place of n in Lemma 1 and Lemma 2. This derivation will be more clear in the proof of Lemma 3.

Given a sampling round τ, we define the statistical bias term U_τ as:

$$U_\tau = R(h_\epsilon + 1/\sqrt{\beta_{\bar{t},\epsilon}\tau}) - R(h_\epsilon) \ . \tag{3.8}$$

which corresponds to the remaining part of the empirical median deviation given in Lemma 1 and Lemma 2 after subtracting the unavoidable bias. Note that statistical bias can be attributed to the randomness of the reward samples whereas the unavoidable bias is due to the adversarial attack. Before moving on with the description of the algorithm procedure, we give the following remark about the sampling rounds of the arms in S.

Remark 2. *Since at each sampling step, the arm with the largest statistical bias is selected by the algorithm for sampling, the difference between the sampling round numbers of any two arms in S cannot be larger than 1, i.e.:*

$$\forall i, j \in S, \ |\tau_i - \tau_j| \leq 1.$$

After the sampling step, algorithm enters the *elimination step* where the arms that are guaranteed to be worse than $(2D)$-suboptimal are eliminated. As shown in Lemma 4, none of the Pareto optimal arms can be eliminated at this step which is crucial for our theoretical analysis to hold since the "additive property of suboptimality" is not satisfied in the adversarial setting as shown in [16, Remark 4].

After the elimination step, algorithm enters the *identification step* where the arms that are guaranteed to satisfy the accuracy condition given in Definition 3 are collected in O_1. Among these arms, the ones that can potentially be used to eliminate an arm in S at a future round are dropped back to S. The rest of the arms in O_1 are collected in O_2. Dropping some arms in O_1 back to S is

necessary to prevent arms that are not $(2D+\alpha)$-suboptimal to potentially end up in P. Note that, identification step contains an if statement that checks whether $U_k > \alpha/4$. When $U_k \leq \alpha/4$, the algorithm terminates by moving all the arms in O_1 to P. This is to guarantee the termination of the algorithm as there might be some arms left in S with suboptimality gaps between $(4D + \alpha)$ and $(2D + \alpha)$ that might cause the algorithm to stuck in an infinite loop in the absence of this step.

3.1 Theoretical Analysis of R-PSI

In this section, we give our accuracy and sample complexity analysis. We start by showing that the *good event* in which the sample median stays within the limits given in Lemma 1 and Lemma 2 occurs with high probability. The rest of our analysis is based on this *good event*.

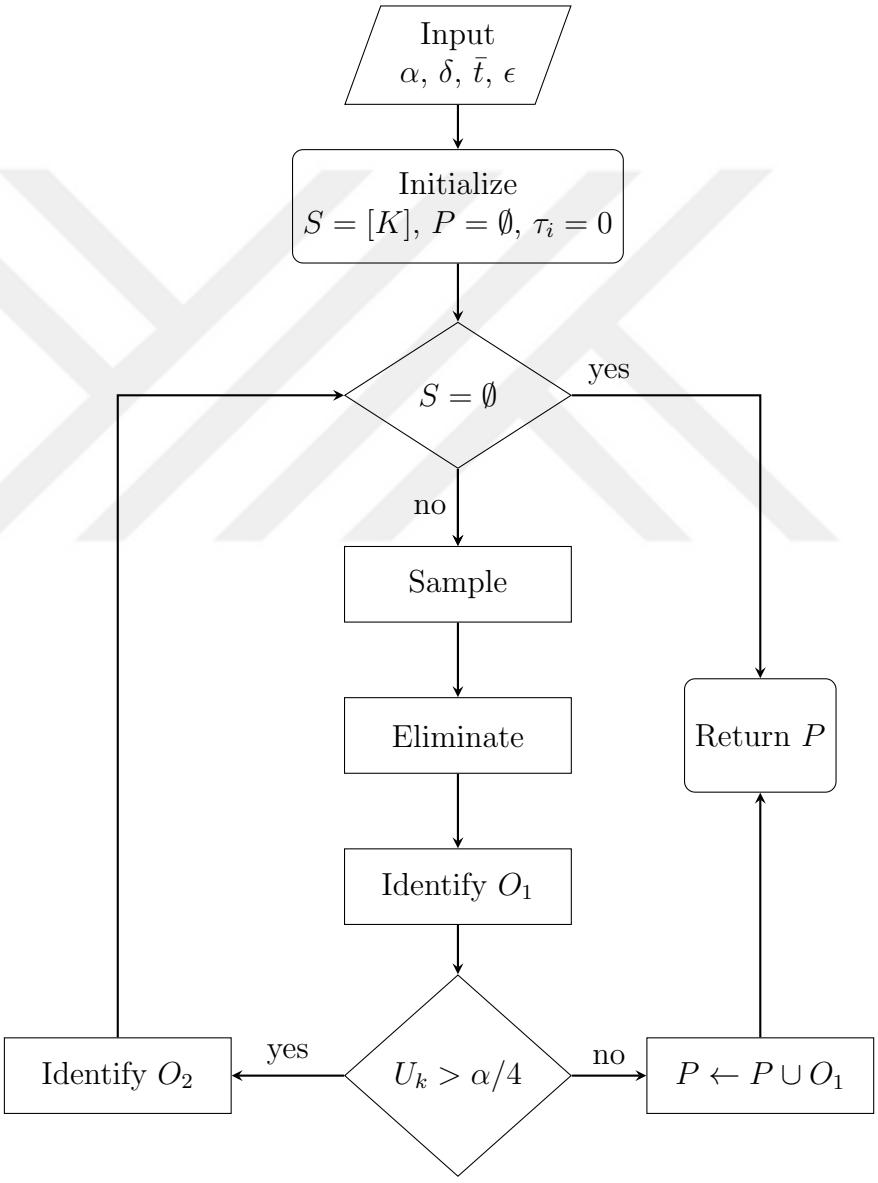


Figure 3.1: Simplified flowchart of R-PSI

Algorithm 1 R-PSI

Input: α, δ, ϵ, $\bar{t}$
Initialize: $S = [K]$, $P = \emptyset$, $\tau_i = 0 \ \ \forall i \in [K]$, $t = 0$.
$\tau_i \leftarrow \tau_i + 1 \ \forall i \in [K]$
Sample each arm n_0 times as in (3.1)
Update U_{i,τ_i} according to (3.8) $\ \forall i \in S$
Update $\hat{m}_i \ \ \forall i \in S$

while $S \neq \emptyset$ **do**
  if $t > 0$ **then**
    Sampling:
    Choose arm $i^* = \arg\max_{k \in [K]} U_{k,\tau_k}$
    $\tau_{i^*} \leftarrow \tau_{i^*} + 1$
    Sample i^* successively $n_{i^*,\tau_{i^*}}$ times as in (3.4)
    Update $U_{i^*,\tau_{i^*}}$ according to (3.8)
    Update $\hat{m}_i^*$
  end if
  Elimination:
  $S \leftarrow S \setminus \{i \in S | \, \exists j \in S \setminus \{i\} : \hat{m}_i + D + U_i \prec \hat{m}_j - D - U_j\}$
  Identification:
  $O_1 \leftarrow \{i \in S | \, \nexists j \in S \setminus \{i\} : \hat{m}_i - U_i + \alpha \prec \hat{m}_j + U_j\}$
  if $\exists k \in S : U_k > \alpha/4$ **then**
    $O_2 \leftarrow \{i \in O_1 | \, \nexists j \in S \setminus \{i\} : \hat{m}_j - U_j + \alpha \prec \hat{m}_i + U_i\}$
    $S \leftarrow S \setminus O_2$
    $P \leftarrow P \cup O_2$
  else
    $P \leftarrow P \cup O_1$
    return P
  end if
  $t \leftarrow t + 1$
end while
return P

Lemma 3. *Define E as the event in which for all $i \in [K]$, $d \in [M]$ and $\tau_i \geq 1$, the following is satisfied:*

$$m_i^d + D + U_{\tau_i} \geq \hat{m}_i{}^d \geq m_i^d - D - U_{\tau_i} \ ,$$

or equivalently, $|\hat{m}_i{}^d - m_i^d| \leq D + U_{\tau_i}$. Then:

$$\mathbb{P}(E) \geq 1 - \delta \ .$$

Proof. When the sampling round of arm i is τ_i, the total number of samplings N_{τ_i} made from arm i is given by:

$$\begin{aligned}
N_{\tau_i} &= n_0 + \sum_{\tau=2}^{\tau_i} n_\tau = \lceil 2\beta_{\bar{t},\epsilon} \log(\frac{\pi^2 MK}{6\tilde{\delta}}) \rceil + \sum_{\tau=2}^{\tau_i} \lceil 4\tau\beta_{\bar{\tau},\epsilon} \log(\frac{\tau}{\tau-1}) + 2\beta_{\bar{t},\epsilon} \log \frac{(\tau-1)^2 MK\pi^2}{6\tilde{\delta}} \rceil \\
&\geq 2\beta_{\bar{t},\epsilon} \log(\frac{\pi^2 MK}{6\tilde{\delta}}) + \sum_{\tau=2}^{\tau_i} \left(4\tau\beta_{\bar{t},\epsilon} \log(\frac{\tau}{\tau-1}) + 2\beta_{\bar{t},\epsilon} \log \frac{(\tau-1)^2 MK\pi^2}{6\tilde{\delta}} \right) \\
&= 2\beta_{\bar{t},\epsilon} \log(\frac{\pi^2 MK}{6\tilde{\delta}}) + \\
&\qquad \sum_{\tau=2}^{\tau_i} \left(2\tau\beta_{\bar{t},\epsilon} \log(\frac{\tau^2 MK\pi^2}{6\tilde{\delta}}) - 2\tau\beta_{\bar{t},\epsilon} \log(\frac{(\tau-1)^2 MK\pi^2}{6\tilde{\delta}}) + 2\beta_{\bar{t},\epsilon} \log \frac{(\tau-1)^2 MK\pi^2}{6\tilde{\delta}} \right) \\
&= 2\beta_{\bar{t},\epsilon} \log(\frac{MK\pi^2}{6\tilde{\delta}}) + \sum_{\tau=2}^{\tau_i} \left(2\tau\beta_{\bar{t},\epsilon} \log(\frac{\tau^2 MK\pi^2}{6\tilde{\delta}}) - 2(\tau-1)\beta_{\bar{t},\epsilon} \log(\frac{(\tau-1)^2 MK\pi^2}{6\tilde{\delta}}) \right) \\
&= 2\tau_i\beta_{\bar{t},\epsilon} \log(\frac{\tau_i^2 MK\pi^2}{6\tilde{\delta}}) = 2\tau_i\beta_{\bar{t},\epsilon} \log(\frac{\delta}{\tilde{\delta}\delta_{\tau_i}}) = 2\tau_i\beta_{\bar{t},\epsilon} \log(\frac{1}{\tilde{\delta_{\tau_i}}}) \ ,
\end{aligned}$$

where $\delta_{\tau_i} = \frac{6\delta}{\tau_i^2 MK\pi^2}$, $\tilde{\delta}_{\tau_i} = (\delta_{\tau_i}\tilde{\delta})/\delta$ and $\tilde{\delta}$ is as given in (3.2b) and (3.3b). Hence, by Lemma 1 and 2, and noting that $\tilde{\delta}_{\tau_i} = \delta_{\tau_i}/2$ for oblivious and prescient adversaries and $\tilde{\delta}_{\tau_i} = \delta_{\tau_i}/3$ for malicious adversary, and $N_{\tau_i} \geq 2\beta_{\bar{t},\epsilon} \log(\frac{1}{\tilde{\delta}_{\tau_i}})$, $\forall \tau_i \geq 1$,

the following holds for $\forall i \in [K]$ and $\forall d \in [M]$:

$$\mathbb{P}\Big(\sup_{m_i^d} |\hat{m}_i{}^d - m_i^d| > R\big(h_\epsilon + \sqrt{\frac{2\log(1/\tilde{\delta}_{\tau_i})}{N_{\tau_i}}}\big)\Big) \leq \delta_{\tau_i} \ .$$

Since $N_{\tau_i} \geq 2\tau_i \beta_{\bar{t},\epsilon} \log(\frac{1}{\tilde{\delta}_{\tau_i}})$ and R is a non-decreasing function:

$$R\big(h_\epsilon + \sqrt{\frac{1}{\beta_{\bar{t},\epsilon}\tau_i}}\big) \geq R\big(h_\epsilon + \sqrt{\frac{2\log(1/\tilde{\delta}_{\tau_i})}{N_{\tau_i}}}\big) \ .$$

Above inequality implies:

$$\mathbb{P}\Big(\sup_{m_i^d} |\hat{m}_i{}^d - m_i^d| > R\big(h_\epsilon + \sqrt{\frac{1}{\beta_{\bar{t},\epsilon}\tau_i}}\big)\Big) \leq \mathbb{P}\Big(\sup_{m_i^d} |\hat{m}_i{}^d - m_i^d| > R\big(h_\epsilon + \sqrt{\frac{2\log(1/\tilde{\delta}_{\tau_i})}{N_{\tau_i}}}\big)\Big) \leq \delta_{\tau_i} \ .$$

Inserting U_{τ_i} and D in the left hand side of the inequality above:

$$\mathbb{P}\Big(\sup_{m_i^d} |\hat{m}_i{}^d - m_i^d| > U_{\tau_i} + D\Big) \leq \delta_{\tau_i} \ .$$

The result follows by taking union bound:

$$\mathbb{P}(E) \geq 1 - \sum_{i\in[K]} \sum_{j\in[M]} \sum_{\tau_i \geq 1} \delta_{\tau_i} = 1 - \sum_{i\in[K]} \sum_{j\in[M]} \sum_{\tau_i \geq 1} \frac{6\delta}{\tau_i^2 MK\pi^2}$$

$$= 1 - MK\frac{6}{MK\pi^2}\delta \sum_{\tau_i \geq 1} \frac{1}{\tau_i^2} = 1 - \delta \ .$$

□

Next, we state the first main result which gives an accuracy guarantee for R-PSI and establishes a high probability upper bound on its sample complexity below.

Theorem 1. *Assume that the reward distributions belong to $C_{R,\bar{t}}$ given in Definition 1 and the event E defined in Lemma 3 holds. Then, given any $\alpha \in \mathbb{R}^+$, and $\epsilon < \frac{2\bar{t}}{1+2\bar{t}}$ for oblivious and prescient adversaries ($\epsilon < \bar{t}$ if malicious adversary), R-PSI is Pareto accurate, with sample complexity N bounded by:*

$$N \leq \sum_{i\in[K]\setminus P^*:\, \Delta_i > 4D+\alpha} 2\tau_{(\bar{\Delta}_i)}(\beta_{\bar{t},\epsilon}\log(\frac{\tau^2_{(\bar{\Delta}_i)}MK\pi^2}{6\tilde{\delta}}) + 1/2) +$$

$$\sum_{i\in P^*:\, \hat{\Delta}_i > 4D+\alpha} 2\tau_{(\tilde{\Delta}_i)}(\beta_{\bar{t},\epsilon}\log(\frac{\tau^2_{(\tilde{\Delta}_i)}MK\pi^2}{6\tilde{\delta}}) + 1/2) +$$

$$\sum_{\{i\in P^*:\, \hat{\Delta}_i \leq 4D+\alpha\}\cup\{i\in[K]\setminus P^*:\, \Delta_i \leq 4D+\alpha\}} 2\tau_{(\alpha)}(\beta_{\bar{t},\epsilon}\log(\frac{\tau^2_{(\alpha)}MK\pi^2}{6\tilde{\delta}}) + 1/2)$$

$$\leq K\tau_{(\alpha)}(2\beta_{\bar{t},\epsilon}\log(\frac{\tau^2_{(\alpha)}MK\pi^2}{6\tilde{\delta}}) + 1)\ . \tag{3.9}$$

where $\bar{\Delta}_i = \Delta_i - 4D$, $\tilde{\Delta}_i = \hat{\Delta}_i - 4D$, and $\hat{\Delta}_i$ is given in Definition 4. Also, for $a \in \mathbb{R}^+$:

$$\tau_{(a)} := \inf\left\{\tau :\ U_\tau \leq a/4\right\}\ .$$

The above theorem gives the most general expression for the sample complexity without making any assumptions on the reward distributions. If a suitable R can be determined for the given reward distributions, then it is possible to derive an explicit gap-dependent bound on the sample complexity. Below, such a result is provided for subgaussian reward distributions.

Corollary 1. *Suppose that the reward distributions are Subgaussian with parameter σ. Then, the asymptotic sample complexity is given by:*

$$\mathcal{O}\Big(\frac{1}{(1/2-\epsilon)^2}\,\frac{K}{\alpha^2}\log\big(\frac{MK}{\alpha\tilde{\delta}}\big)\Big)\ .$$

Proof. Let $\tau_- := \tau_{(\alpha)} - 1$. By definition of $\tau_{(\alpha)}$, we have:

$$U_{\tau_-} = R(h_\epsilon + 1/\sqrt{\beta_{\bar{t},\epsilon}(\tau_-)}) - R(h_\epsilon) > \alpha/4\ .$$

Using $R(t)$ from Example 1, we have:

$$\sigma\sqrt{2}\Bigg(\sqrt{\log\Bigg(\frac{1}{1/2-h_\epsilon-\sqrt{\frac{1}{\beta_{\bar{t},\epsilon}\tau_-}}}\Bigg)} - \sqrt{\log\big(\frac{1}{1/2-h_\epsilon}\big)}\Bigg) > \alpha/4\ .$$

Arranging the terms gives:

$$\tau_- < \frac{1}{(1/2-h_\epsilon)^2\beta_{\bar{t},\epsilon}}\left(\frac{1}{1-e^{-c_1\alpha^2-c_2\alpha}}\right)^2\ .$$

where $c_1 = \frac{1}{32\sigma^2} > 0$ and $c_2 = \frac{1}{2\sigma\sqrt{2}}\sqrt{\log\big(\frac{1}{1/2-h_\epsilon}\big)} > 0$. Hence:

$$\tau_{(\alpha)} < 1 + \frac{1}{(1/2-h_\epsilon)^2\beta_{\bar{t},\epsilon}}\left(\frac{1}{1-e^{-c_1\alpha^2-c_2\alpha}}\right)^2\ .$$

Note that $\frac{1}{1-e^{-x}} = 1+\frac{1}{e^x-1} \le 1+\frac{1}{x}$ which follows from the relation $e^x - 1 > x,\ \forall x$. Applying this to above, we obtain:

$$\tau_{(\alpha)} < 1 + \frac{1}{(1/2-h_\epsilon)^2\beta_{\bar{t},\epsilon}}\left(1+\frac{1}{c_1\alpha^2+c_2\alpha}\right)^2\ .$$

Note that as $\alpha \to 0$, the dominant term in the denominator of $\frac{1}{c_1\alpha^2+c_2\alpha}$ becomes $c_2\alpha$. Hence, by (3.9):

$$N = \mathcal{O}\Big(\frac{1}{(1/2-h_\epsilon)^2}\frac{K}{\alpha^2}\log\big(\frac{MK}{\tilde{\delta}\alpha}\big)\Big)\ .$$

□

The sample complexity bound derived for Subgaussian distributions asymptotically matches (in the worst case) the bound derived for Algorithm 1 in [1] and the bounds in [16] in terms of the suboptimality gap dependence. However, note that a direct comparison between the results may not be meaningful since Auer et al. [1] consider a mean based ranking of arms while our method is median based. Also note that unlike these studies, we have an ϵ dependent factor $\frac{1}{(1/2-h_\epsilon)}$ in our bound due to the adversarial attack. As expected, this term increases as ϵ increases, and goes to infinity when $\epsilon \to 1/2$.

3.1.1 Proof of Theorem 1

In the first part of the proof, we will show that R-PSI is Pareto accurate and obtain α-only dependent sample complexity given in (3.9). In the second part, we will improve on this bound by obtaining gap-dependent bounds for arms with median reward vectors that satisfy certain conditions.

3.1.1.1 Proof of Pareto accuracy and α-only dependent sample complexity

We start by proving that the Pareto optimal arms are not eliminated in the 'Elimination' step.

Lemma 4. *R-PSI does not eliminate Pareto optimal arms in the 'Elimination' step.*

Proof. At elimination step, an arm i is eliminated if

$$\exists j \in S : \hat{m}_i + D + U_i \prec \hat{m}_j - D - U_j \ .$$

Applying Lemma 3 to above:

$$(m_i - D - U_i) + D + U_i \preceq \hat{m}_i + D + U_i \prec \hat{m}_j - D - U_j \preceq (m_j + D + U_j) - D - U_j \ .$$

Hence, $m_i \prec m_j$. By definition of the Pareto optimality, this is not possible if i is a Pareto optimal arm. Hence, Pareto optimal arms cannot be discarded at the elimination step of R-PSI. □

Next, we prove that R-PSI has a suboptimality guarantee of $(2D+\alpha)$, i.e, any arm returned in P is $(2D+\alpha)$ suboptimal. But first, in Lemma 5, we state a technical result needed in the proof.

Lemma 5. *Suppose an arm i is moved to O_2 at round t_1. Then, the following is satisfied for all $t \geq t_1$:*

$$\forall j \in S,\ \exists d_j \in [M] : m_j^{d_j} + D + \alpha > m_i^{d_j} - D\ .$$

Proof. Since i is moved to O_2 at round t_1, the condition for O_2 requires that:

$$\nexists j \in S : \hat{m}_j - U_j + \alpha \prec \hat{m}_i + U_i\ .$$

This implies:

$$\forall j \in S,\ \exists d_j \in [M] : \hat{m_j}^{d_j} - U_j + \alpha \geq \hat{m}_i^{d_j} + U_i\ .$$

Applying Lemma 3 gives:

$$\begin{aligned}\forall j \in S,\ \exists d_j \in [M] : (m_j^{d_j} + D + U_j) - U_j + \alpha \geq \\ \hat{m_j}^{d_j} - U_j + \alpha \geq \\ \hat{m_i}^{d_j} + U_i \geq \\ (m_i^{d_j} - D - U_i) + U_i\ .\end{aligned}$$

Hence, if arm i is moved to O_2, then arm i satisfies the following:

$$\forall j \in S,\ \exists d_j \in [M] : m_j^{d_j} + D + \alpha \geq m_i^{d_j} - D\ .$$

The result follows by considering that S does not admit any new arms over time. □

Now, we are ready to prove the suboptimality guarantee for the predicted arms.

Lemma 6. *P can only contain arms that are $(2D + \alpha)$-suboptimal.*

Proof. Consider $i \in S$ that is moved to P in round t_2. Note that an arm in S needs to be first moved to O_1 to end up in P. Denote the Pareto optimal arms in S by $S^{(p)}$. Suppose that an arm $i \in S$ is moved to O_1. Then, by O_1 condition, and considering that $S^{(P)} \subseteq S$:

$$\nexists j \in S^{(p)} \setminus \{i\} : \hat{m}_i - U_i + \alpha \prec \hat{m}_j + U_j \ .$$

Then, by definition of dominance, the above implies that:

$$\forall j \in S^{(p)} \setminus \{i\},\ \exists d_j \in [M] : \hat{m_i}^{d_j} - U_i + \alpha \geq \hat{m_j}^{d_j} + U_j \ .$$

By Lemma 3:

$$\begin{aligned} \forall j \in S^{(p)} \setminus \{i\},\ \exists d_j \in [M] : (m_i^{d_j} + D + U_i) - U_i + \alpha \geq \\ \hat{m_i}^{d_j} - U_i + \alpha \geq \\ \hat{m_j}^{d_j} + U_j \geq \\ (m_j^{d_j} - D - U_j) + U_j \ . \end{aligned}$$

Hence:

$$\forall j \in S^{(p)} \setminus \{i\},\ \exists d_j \in [M] : m_i^{d_j} + D + \alpha \geq m_j^{d_j} - D \ . \tag{3.10}$$

Next, consider the set that consists of the optimal arms in P, which we denote by $P^{(p)}$. By Lemma 5:

$$\forall j \in P^{(p)},\ \exists d_j \in [M] : m_i^{d_j} + D + \alpha \geq m_j^{d_j} - D \ . \tag{3.11}$$

Note that at any round, $S^{(p)} \cup P^{(p)} = P^*$ since Lemma 4 implies that a Pareto optimal arm is either in S or P up until the identification step at the last round.

Hence, by (3.10) and (3.11):

$$\forall j \in P^*,\ \exists d_j \in [M] : m_i^{d_j} + D + \alpha \geq m_j^{d_j} - D\ .$$

By Definition 2, this implies that $\Delta_i \leq 2D + \alpha$. Hence, we conclude that if arm i is moved to O_1, it needs to be $(2D + \alpha)$-suboptimal and the result follows. $\square$

Next, we show that for any optimal arm that is not returned in P, there exists an arm returned in P which is not worse than the optimal arm more than the amount of $2D$ in any of the objectives.

Lemma 7. *If a Pareto optimal arm p^* is not returned in P, then there exists an arm j returned in P such that $m_{p^*} - m_j \prec 2D$.*

Proof. First, note that, if any of the optimal arms is not returned in P, this implies that, at the final round before termination, the following is satisfied: $\forall i \in S,\ U_i \leq \alpha/4$, since this is the only possible way to enter the if statement in the identification step where it is possible for a Pareto optimal arm to not end up in P in the end. Otherwise, all the Pareto optimal arms are returned in P because of Lemma 4. Also, considering that p^* is not included in O_1 at the final round, we have an arm $l_1 \in S$ such that:

$$\hat{m}_{p^*} - U_{\tau_{p^*}} + \alpha \prec \hat{m}_{l_1} + U_{\tau_{l_1}}\ . \tag{3.12}$$

Also, considering that $U_i, U_j \leq \alpha/4$, the above inequality implies:

$$\hat{m}_{l_1} - U_{\tau_{l_1}} + \alpha \not\prec \hat{m}_{p^*} + U_{\tau_{p^*}}\ . \tag{3.13}$$

In the rest of the proof, we use the notation $p^* \prec l_1$ and $l_1 \not\prec p^*$ to denote the relations in (3.12) and (3.13) respectively. Now, let $[l_i]_{i=1}^n$ be a subset of S that satisfies the following conditions:

1. $p^* \prec l_1 \cdots \prec l_n$.
2. There are no repeating arms in the set.

Below, we will prove by induction that the following holds:

$$\forall z \in \{p^*\} \cup [l_i]_{i=1}^{k-1},\ l_k \not\prec z\ , \tag{3.14}$$

for any $k \in [n]$. We have already proven in (3.13) that when $k = 1$, (3.14) holds. Now, assuming that (3.14) holds for $k-1$, we show that it also holds for k. First, observe that, since $l_{k-1} \prec l_k$ and since $\alpha > 2U_{\tau_{l_{k-1}}}$ and $\alpha > 2U_{\tau_{l_k}}$:

$$\begin{aligned} \hat{m}_{l_{k-1}} + U_{\tau_{l_{k-1}}} &\prec \hat{m}_{l_{k-1}} + U_{\tau_{l_{k-1}}} + (\alpha - 2U_{\tau_{l_{k-1}}}) = \\ &\hat{m}_{l_{k-1}} - U_{\tau_{l_{k-1}}} + \alpha \prec \hat{m}_{l_k} + U_{\tau_{l_k}} \\ &\prec \hat{m}_{l_k} + U_{\tau_{l_k}} + (\alpha - 2U_{\tau_{l_k}}) = \hat{m}_{l_k} - U_{\tau_{l_k}} + \alpha\ . \end{aligned} \tag{3.15}$$

Therefore, $\hat{m}_{l_{k-1}} + U_{\tau_{l_{k-1}}} \prec \hat{m}_{l_k} - U_{\tau_{l_k}} + \alpha,$ which implies that:

$$l_k \not\prec l_{k-1}. \tag{3.16}$$

Also, by assumption, $l_{k-1} \not\prec z,\ \forall z \in \{p^*\} \cup [l_i]_{i=1}^{k-2}$, or equivalently:

$$\hat{m}_{l_{k-1}} - U_{\tau_{l_{k-1}}} + \alpha \not\prec \hat{m}_z + U_{\tau_z},\ \ \forall z \in \{p^*\} \cup [l_i]_{i=1}^{k-2}\ . \tag{3.17}$$

Also it is shown in (3.15) that:

$$\hat{m}_{l_{k-1}} - U_{\tau_{l_{k-1}}} + \alpha \prec \hat{m}_{l_k} - U_{\tau_{l_k}} + \alpha\ . \tag{3.18}$$

From (3.17) and (3.18), we conclude that:

$$\hat{m}_{l_k} - U_{\tau_{l_k}} + \alpha \not\prec \hat{m}_z + U_{\tau_z},\ \ \forall z \in \{p^*\} \cup [l_i]_{i=1}^{k-2}\ . \tag{3.19}$$

Hence by (3.16) and (3.19):

$$l_k \not\prec z,\ \ \forall z \in \{p^*\} \cup [l_i]_{i=1}^{k-2} \cup \{l_{k-1}\}\ , \tag{3.20}$$

which proves the induction. Now, suppose that we can find a new arm l_{n+1} in S such that $l_n \prec l_{n+1}$. Since we can add l_{n+1} to $[l_i]_{i=1}^{n}$ without violating the 2 conditions we set previously for $[l_i]_{i=1}^{n}$, the induction proven above holds for

the arms in the new set $[l_i]_{i=1}^{n+1}$ as well. Therefore, we conclude that l_{n+1} is not prevented by any other arm in $\{p^*\} \cup [l_i]_{i=1}^{n}$ to enter O_1. Now, suppose that we sequentially keep adding arms in a way that the 2 conditions set previously are not violated. Considering that we have a finite amount of arms in S, we will eventually reach a point where it will not be possible to add any more arms. Let L denote a set that is constructed through such a procedure and let j denote the last element of L. Then, j satisfies the following:

$$p^* \prec l_1 \cdots \prec j \ . \tag{3.21}$$

Also, considering that we cannot keep adding any more arms to this set:

$$\forall k \in S \setminus (L \cup \{p^*\}) : j \not\prec k \ . \tag{3.22}$$

In addition, considering that L satisfies the induction we have proven above and j is the last element of L:

$$\forall k \in L \cup \{p^*\} : j \not\prec k \ . \tag{3.23}$$

Combining (3.22) and (3.23), we obtain:

$$\forall k \in S : j \not\prec k \ .$$

Hence, arm j is returned in P. Next, we prove by another induction that $p^* \prec l_k$ for any $l_k \in L$. This is already given for the first element, i.e., $p^* \prec l_1$. Now assume that this is true for arm l_{k-1} , i.e., $p^* \prec l_{k-1}$, or equivalently, $\hat{m}_{p^*} - U_{\tau_{p^*}} + \alpha \prec \hat{m}_{l_{k-1}} + U_{\tau_{l_{k-1}}}$. As shown in (3.15), we also have that $\hat{m}_{l_{k-1}} + U_{\tau_{l_{k-1}}} \prec \hat{m}_{l_k} + U_{\tau_{l_k}}$. Combining these, we have:

$$\hat{m}_{p^*} - U_{\tau_{p^*}} + \alpha \prec \hat{m}_{l_k} + U_{\tau_{l_k}},$$

or $p^* \prec l_k$ and the induction is proven. Since arm j is in L, $p^* \prec j$, or equivalently:

$$\hat{m}_{p^*} - U_{\tau_{p^*}} + \alpha \prec \hat{m}_j + U_{\tau_j} \ .$$

Applying Lemma 3 above, we get:

$$\begin{aligned} m_{p^*} - D - 2U_{\tau_{p^*}} + \alpha &\prec \hat{m}_{p^*} - U_{\tau_{p^*}} + \alpha \\ &\prec \hat{m}_j + U_{\tau_j} \prec m_j + D + 2U_{\tau_j} \ . \end{aligned}$$

Hence:

$$m_{p^*} - D - 2U_{\tau_{p^*}} + \alpha \prec m_j + D + 2U_{\tau_j} \ .$$

Since $\alpha \geq 2U_{\tau_{p^*}} + 2U_{\tau_j}$:

$$m_{p^*} - D \prec m_j + D \ .$$

□

By inspecting the pseudocode of R-PSI, it can be observed that if the algorithm enters the "else" statement inside the identification step at some round, then, after performing the operations inside the "else" statement, it terminates. This leads to the following conclusion.

Remark 3. *R-PSI terminates at the latest when* $\forall i \in S,\ U_i \leq \alpha/4$.

Now, we are ready to prove the first part of Theorem 1.

First part of the proof of Theorem 1:

By Remark 2 and 3, at termination:

$$\tau_i \leq \inf\{\tau : U_\tau \leq \alpha/4\},\ \forall i \in [K] \ .$$

The total number of samples taken from an arm with a sampling round number

τ can be bounded as:

$$N_\tau = n_0 + \sum_{\tilde{\tau}=2}^{\tau} n_{\tilde{\tau}}$$

$$= \lceil 2\beta_{\bar{t},\epsilon} \log(\frac{\pi^2 MK}{6\tilde{\delta}}) \rceil + \sum_{\tilde{\tau}=2}^{\tau} \lceil 4\tilde{\tau}\beta_{\bar{t},\epsilon} \log(\frac{\tilde{\tau}}{\tilde{\tau}-1}) + 2\beta_{\bar{t},\epsilon} \log \frac{(\tilde{\tau}-1)^2 MK\pi^2}{6\tilde{\delta}} \rceil$$

$$\leq 1 + 2\beta_{\bar{t},\epsilon} \log(\frac{\pi^2 MK}{6\tilde{\delta}}) + \sum_{\tilde{\tau}=2}^{\tau} \left(1 + 4\tilde{\tau}\beta_{\bar{t},\epsilon} \log(\frac{\tilde{\tau}}{\tilde{\tau}-1}) + 2\beta_{\bar{t},\epsilon} \log \frac{(\tilde{\tau}-1)^2 MK\pi^2}{6\tilde{\delta}} \right)$$

$$= 1 + 2\beta_{\bar{t},\epsilon} \log(\frac{\pi^2 MK}{6\tilde{\delta}}) + \sum_{\tilde{\tau}=2}^{\tau} \left(1 + 2\tilde{\tau}\beta_{\bar{t},\epsilon} \log(\frac{\tilde{\tau}^2 MK\pi^2}{6\tilde{\delta}}) - 2(\tilde{\tau}-1)\beta_{\bar{t},\epsilon} \log(\frac{(\tilde{\tau}-1)^2 MK\pi^2}{6\tilde{\delta}}) \right)$$

$$= 1 + \sum_{\tilde{\tau}=2}^{\tau} (1) + 2\beta_{\bar{t},\epsilon} \log(\frac{\pi^2 MK}{6\tilde{\delta}}) + \sum_{\tilde{\tau}=2}^{\tau} \left(2\tilde{\tau}\beta_{\bar{t},\epsilon} \log(\frac{\tilde{\tau}^2 MK\pi^2}{6\tilde{\delta}}) - 2(\tilde{\tau}-1)\beta_{\bar{t},\epsilon} \log(\frac{(\tilde{\tau}-1)^2 MK\pi^2}{6\tilde{\delta}}) \right)$$

$$= 2\tau(\beta_{\bar{t},\epsilon} \log(\frac{\tau^2 MK\pi^2}{6\tilde{\delta}}) + 1/2) \ . \tag{3.24}$$

Hence, the sample complexity N of R-PSI can be bounded as:

$$N \leq K\tau_{(\alpha)}(2\beta_{\bar{t},\epsilon} \log(\frac{\tau_{(\alpha)}^2 MK\pi^2}{6\tilde{\delta}}) + 1) \ .$$

The Pareto accuracy of R-PSI follows from Lemma 6 and 7.

3.1.1.2 Improving the sample complexity bounds for special arms

In this part, we show that it is possible to obtain tighter sample complexity bounds for some of the Pareto optimal and suboptimal arms whose median reward vectors satisfy certain conditions that depend on the unavoidable bias. The new sample complexity bounds depend on the geometric properties of the median reward vectors as well as the accuracy parameter α. We start by giving a relaxed elimination condition for suboptimal arms. But first, we prove a technical result in the following lemma needed in the proof.

Lemma 8. *Consider an arm i that is not $(2D+\alpha)$ suboptimal and suppose that the Pareto optimal arm $j = \arg\max_{k\in P^*} \Delta_{i,k}$ is moved to P at t_0. Then i is eliminated at a round $t \leq t_0$.*

Proof. If j is moved to P at t_0, then by Lemma 5:

$$\forall k \in S,\ \exists d_k \in [M] : m_k^{d_k} + (2D+\alpha) \geq m_j^{d_k},$$

which implies that $\Delta_k \leq 2D+\alpha$. Therefore, it is not possible for arm i to be in S at t_0 since $\Delta_i > 2D+\alpha$. Hence, i was eliminated at a round $t \leq t_0$. □

Now, we are ready to prove the relaxed elimination condition.

Lemma 9. *An arm i that is not $(4D+\alpha)$-suboptimal is guaranteed to be eliminated at the latest when $\forall k \in S,\ U_k \leq \bar{\Delta}_i/4$ where $\bar{\Delta}_i = \Delta_i - 4D$.*

Proof. By Lemma 6, if $i \notin S$, then i is already eliminated since i cannot be in P. Now, suppose that $i \in S$. Consider the optimal arm $j = \arg\max_{k\in P^*} \Delta_{i,k}$. Note that $\Delta_{i,j} = \Delta_i > 4D+\alpha$ and $\forall d \in [M],\ m_j^d \geq m_i^d + \Delta_i$. Now, suppose that $j \in S$ so that $U_j \leq \bar{\Delta}_i/4$. Then, by Lemma 3:

$$\begin{aligned}\hat{m_j}^d - D - U_j \geq m_j^d - 2D - 2U_j &\geq m_i^d + \Delta_i - 2D - 2U_j \\ &\geq \hat{m_i}^d + \Delta_i - 3D - 2U_j - U_i, \quad \forall d \in [M]. \end{aligned} \tag{3.25}$$

Considering that $\Delta_i > 4D+\alpha$ and $U_i, U_j \leq \bar{\Delta}/4$, the R.H.S of above can be bounded as:

$$\hat{m_i}^d + \Delta_i - 3D - 2U_j - U_i > \hat{m_i}^d + D + U_i\ . \tag{3.26}$$

Therefore, by (3.25) and (3.26):

$$\hat{m_j}^d - D - U_j > \hat{m_i}^d + D + U_i\ .$$

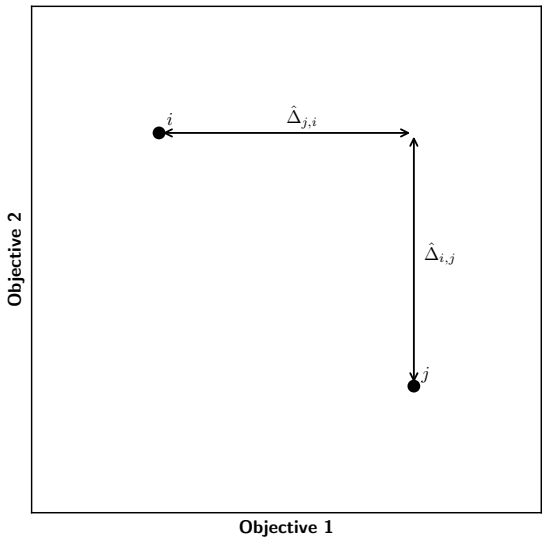


Figure 3.2: Illustration of the domination gap concept

Hence, we conclude that when $j \in S$, i is eliminated. Now suppose that $j \notin S$ and also suppose that, at the earliest round when $\forall k \in S \quad , U_k \leq \bar{\Delta}_i/4$, the algorithm does not enter the if statement in the identification step. Then, since j cannot be eliminated by Lemma 4, $\quad j \in P$ and by Lemma 8, i is eliminated. Now suppose that, as soon as $\forall k \in S \quad , U_k \leq \bar{\Delta}_i/4$ is satisfied, the algorithm enters the if statement in the identification step. Then, the algorithm terminates by Remark 3 and i is not included in P by Lemma 6. □

Next, we prove an improved gap-dependent identification condition for some of the Pareto optimal arms. We start by introducing a special gap definition for Pareto optimal arms which we call *decision gap*.

Definition 4. ***Decision and domination gaps.*** *For a Pareto optimal arm i, we define:*

$$\hat{\Delta}_{i,k} := \max\{0, \max_d \, (m_i^d - m_k^d)\} \ .$$

as the ***domination gap*** *of arm i with respect to arm k. In other words, $\hat{\Delta}_{i,k}$ defines the smallest amount that should be added to arm k in order for arm k to weakly dominate arm i. A demonstration of the domination gap is given in*

Figure 3.2. Based on the domination gap, we define the following intermediate gap definitions:

$$\hat{\Delta}_i^1 := \min_{k \in P^* \setminus \{i\}} \min\{\hat{\Delta}_{i,k}, \hat{\Delta}_{k,i}\} \ ,$$

$$\hat{\Delta}_i^2 := \min_{k \in [K] \setminus (P^* \cup A_i)} \min\{\hat{\Delta}_{i,k}, \hat{\Delta}_{k,i}\} \ ,$$

$$\hat{\Delta}_i^3 := \min_{k \in A_i} \Delta_k \ ,$$

where $A_i = \{k \in [K] \setminus \{i\} |\ m_k \prec m_i\}$ *. Based on these, we define:*

$$\hat{\Delta}_i = \min\{\hat{\Delta}_i^1, \hat{\Delta}_i^2, \hat{\Delta}_i^3\} \ ,$$

as the ***decision gap*** *of* i*.*

Note that a Pareto optimal arm i that is in the $2D$ proximity of another arm in the arm set (in terms of the Chebychev distance), have a domination gap smaller than $2D$. (Including an arm j in A_i since if j is in $2D$ proximity of i then $\Delta_j \leq 2D$, hence $\hat{\Delta}^i \leq 2D$). Finally, we note that the definition of decision gap will be clear in the next lemma. Before moving on, we give the following remark that simplifies the expression of the decision gap definition.

Remark 4. $\hat{\Delta}_i$ *can also be written in the following form:*

$$\hat{\Delta}_i = \min\{\hat{\Delta}_i^1, \hat{\Delta}_i^2, \hat{\Delta}_i^3\} = \min\{\hat{\Delta}_i^1, \min_{k \in [K] \setminus (P^* \cup A_i)} \hat{\Delta}_{k,i}, \ \hat{\Delta}_i^3\} \ .$$

Proof. First, note that for all suboptimal arms, there exists a Pareto optimal arm that dominates it. Now, consider a suboptimal arm $k \in [K] \setminus (P^* \cup A_i)$ and $p^k = \arg\max_{p \in P^*} \Delta_{k,p}$ so that $\Delta_k = \Delta_{k,p^k}$. Since $\forall d \in [M],\ m_k^d \leq m_{p^k}^d$, we have $m_i^d - m_{p^k}^d \leq m_i^d - m_k^d$. Therefore:

$$\hat{\Delta}_{i,p^k} \leq \hat{\Delta}_{i,k} \ . \tag{3.27}$$

Hence:

$$\hat{\Delta}_i^1 \leq \min_{p^k \in P^* \setminus \{i\}} \hat{\Delta}_{i,p^k} \leq \min_{k \in [K] \setminus (P^* \cup A_i)} \hat{\Delta}_{i,k} \ ,$$

and the simplification follows. □

Lemma 10. *Suppose that $\hat{\Delta}_i > 4D + \alpha$ for a Pareto optimal arm i. If $U_k \leq \tilde{\Delta}_i/4,\ \forall k \in S$ at some round t, where $\tilde{\Delta}_i = \hat{\Delta}_i - 4D$, then, arm i is moved to P at t.*

Proof. We start by showing that i is moved to O_1 at t. First, observe that, given $U_k \leq \tilde{\Delta}_i/4,\ \forall k \in S$, the following holds $\forall j \in S \cap \big[(P^* \setminus \{i\}) \cup \big([K] \setminus (P^* \cup A_i)\big)\big] = S \cap \big[[K] \setminus (A_i \cup \{i\})\big]$:

$$\begin{aligned}(m_i^{d(j)} - m_j^{d(j)} - 2D) - 2(U_i + U_j) + \alpha = \hat{\Delta}_{i,j} - 2D + \alpha - 2(U_i + U_j) \\ \geq \hat{\Delta}_i - 2D + \alpha - 2(U_i + U_j) = \tilde{\Delta}_i + 2D + \alpha - 2(U_i + U_j) \geq 2D + \alpha > 0\ .\end{aligned} \tag{3.28}$$

where $d(j) = \arg\max_{d \in [M]}(m_i^d - m_j^d)$, i.e., $\hat{\Delta}_{i,j} = m_i^{d(j)} - m_j^{d(j)}$. Applying Lemma 3 to L.H.S, we obtain:

$$\hat{m}_i^{d(j)} - \hat{m}_j^{d(j)} - (U_i + U_j) + \alpha > 0,\ \ \forall j \in S \cap \big[[K] \setminus (A_i \cup \{i\})\big]\ .$$

Rearranging the terms give:

$$\hat{m}_i^{d(j)} - U_i + \alpha > \hat{m}_j^{d(j)} + U_j,\ \ \forall j \in S \cap \big[[K] \setminus (A_i \cup \{i\})\big]\ .$$

Above implies:

$$\hat{m}_i^{d(j)} - U_i + \alpha \not> \hat{m}_j^{d(j)} + U_j,\ \ \forall j \in S \cap \big[[K] \setminus (A_i \cup \{i\})\big]\ . \tag{3.29}$$

Note that, $\forall j \in A_i,\ \ \Delta_j \geq \hat{\Delta}_i > 4D + \alpha$, which implies that $\forall k \in S,\ U_k \leq \tilde{\Delta}_i/4 = (\hat{\Delta}_i - 4D)/4 \leq (\Delta_i - 4D)/4 = \bar{\Delta}_i$ at round t. Therefore, by Lemma 9, all $j \in A_i$ is eliminated at the elimination step of round t, (if not eliminated before) which in turn guarantees that:

$$S \subseteq [K] \setminus (A_i \cup \{i\})\ . \tag{3.30}$$

Therefore, $S \cap \big[[K] \setminus (A_i \cup \{i\})\big] = S \setminus \{i\}$, after the elimination step, and before the beginning of the identification step at round t .Hence, we conclude by (3.29) and by the fact that a Pareto optimal arm cannot be eliminated until the last

round by Lemma 4, i is moved to O_1 at t.

If $U_k \leq \alpha/4,\ \forall k \in S,$ then the algorithm terminates by moving i, and the other arms in O_1, to P. Therefore, we are left to show that when $U_k > \alpha/4$, i is moved to O_2 at the same round. First, observe that the following holds $\forall j \in \big[S \cap \big([K] \setminus (A_i \cup \{i\})\big)\big]$:

$$\begin{aligned}&(m_j^{d(j)} - m_i^{d(j)}) - 2D - 2(U_j + U_i) + \alpha = \hat{\Delta}_{j,i} - 2D + \alpha - 2(U_j + U_i)\\ &\geq \hat{\Delta}_i - 2D + \alpha - 2(U_j + U_i) = \bar{\Delta}_i + 2D + \alpha - 2(U_i + U_j) \geq 2D + \alpha > 0\ .\end{aligned}$$

where $d(j) = \arg\max_{d\in[M]}(m_j^d - m_i^d)$, i.e., $\hat{\Delta}_{j,i} = m_j^{d(j)} - m_i^{d(j)}$. Note that we have applied the same steps in (3.28) to reach the above result. Applying Lemma 3 to L.H.S, we obtain:

$$\hat{m}_j^{d(j)} - \hat{m}_i^{d(j)} - (U_i + U_j) + \alpha > 0,\ \ \forall j \in \big[S \cap \big([K] \setminus (A_i \cup \{i\})\big)\big]\ .$$

Arranging the terms above, we obtain:

$$\hat{m}_j^{d(j)} - U_j + \alpha > \hat{m}_i^{d(j)} + U_i,\ \ \forall j \in \big[S \cap \big([K] \setminus (A_i \cup \{i\})\big)\big]\ ,$$

which implies:

$$\hat{m}_j - U_j + \alpha \not< \hat{m}_i + U_i,\ \ \forall j \in \big[S \cap \big([K] \setminus (A_i \cup \{i\})\big)\big]\ . \tag{3.31}$$

We have already established in (3.30) that $S \subseteq [K] \setminus (A_i \cup \{i\})$ before the identification step at round t and therefore $\big[S \cap \big([K] \setminus (A_i \cup \{i\})\big)\big] = S \setminus \{i\}$. Hence, we conclude from (3.31) that i is moved to O_2. □

The condition $\hat{\Delta}_i > (4D + \alpha)$ in the above Lemma 10 suggests that the Pareto optimal arm i is "well separated" from the other arms in the objective space. Specifically, this condition implies that Chebyshev distance between an optimal arm i which has a decision gap $\hat{\Delta}_i > 4D + \alpha$, and all the other arms in the arm set is larger than $(4D + \alpha)$.

Remaining part of the proof of Theorem 1:

Since $\bar{\Delta}_i > \alpha$, by Lemma 9, Remark 2, and Remark 3, we can get tighter bounds on the sample complexity of arms that are not $(4D+\alpha)$-suboptimal as follows:

$$\tau_i \leq \inf\{\tau : U_\tau \leq \bar{\Delta}_i/4\} \ .$$

Similarly, for Pareto optimal arms with decision gap larger than $(4D+\alpha)$, we have by Lemma 10, Remark 2 and Remark 3:

$$\tau_i \leq \inf\{\tau : U_\tau \leq \tilde{\Delta}_i/4\} \ .$$

Then, by equation (3.24):

$$\begin{aligned} N \leq & \sum_{i\in[K]\setminus P^*:\ \Delta_i > 4D+\alpha} 2\tau_{(\bar{\Delta}_i)}(\beta_{\bar{t},\epsilon}\log(\frac{\tau^2_{(\bar{\Delta}_i)}MK\pi^2}{6\tilde{\delta}}) + 1/2) + \\ & \sum_{i\in P^*:\ \hat{\Delta}_i > 4D+\alpha} 2\tau_{(\tilde{\Delta}_i)}(\beta_{\bar{t},\epsilon}\log(\frac{\tau^2_{(\tilde{\Delta}_i)}MK\pi^2}{6\tilde{\delta}}) + 1/2) + \\ & \sum_{\{i\in P^*:\ \hat{\Delta}_i \leq 4D+\alpha\ \}\cup\ \{i\in[K]\setminus P^*:\ \Delta_i \leq 4D+\alpha\}} 2\tau_{(\alpha)}(\beta_{\bar{t},\epsilon}\log(\frac{\tau^2_{(\alpha)}MK\pi^2}{6\tilde{\delta}}) + 1/2) \ . \end{aligned} \tag{3.32}$$

□

Chapter 4

R-PSI Variation

In this Chapter, we give a variation of the R-PSI algorithm, which we call R-PSI VAR, that exactly recovers all the Pareto optimal arms alongside with some $(4D+\alpha)$ suboptimal arms, hence, sacrificing the $(2D+\alpha)$ suboptimaltiy guarantee for a $(4D+\alpha)$ one. The algorithm pseudocode is given in Algorithm 2. Also we give the flowchart of the algorithm in Figure 4.1.

The difference between R-PSI and R-PSI VAR algorithms is that the identification step in R-PSI VAR is less restrictive. With this change, $(4D+\alpha)$-suboptimal arms can end up in P, instead of the previous tighter $(2D+\alpha)$ selection. Now that the infinite loop problem caused by the stucking of the $(4D+\alpha)$-suboptimal arms in the case of R-PSI is no longer present, we are able to remove the if statement in the identification step of the original R-PSI algorithm that looks for whether $U_i \leq \alpha/4,\ \forall i \in S$.

First, we start by proving that the sample complexity bound of R-PSI VAR is the same with the one found for R-PSI.

4.1 Theoretical Analysis

We state our main result for R-PSI VAR which gives accuracy guarantees on R-PSI VAR predictions and establishes a high probability upper bound on the sample complexity.

Theorem 2. *Assume that the reward distributions belong to $C_{R,\bar{t}}$ and the event E defined in Lemma 3 holds. Then, given any $\alpha > 0$, and $\epsilon \in (0, \frac{2\bar{t}}{1+2\bar{t}})$ for oblivious and prescient adversaries ($\epsilon \in (0, \bar{t})$ if malicious adversary), R-PSI VAR returns arms that are guaranteed to be $(4D+\alpha)$ suboptimal, and the predicted set includes the Pareto optimal set. The sample complexity N is bounded by:*

$$
\begin{aligned}
N \leq & \sum_{i \in [K] \setminus P^*:\, \Delta_i > 4D+\alpha} 2\tau_{(\bar{\Delta}_i)}(\beta_{\bar{t},\epsilon} \log(\frac{\tau^2_{(\bar{\Delta}_i)} MK\pi^2}{6\tilde{\delta}}) + 1/2) + \\
& \sum_{i \in P^*:\, \hat{\Delta}_i > 4D+\alpha} 2\tau_{(\bar{\Delta}_i)}(\beta_{\bar{t},\epsilon} \log(\frac{\tau^2_{(\bar{\Delta}_i)} MK\pi^2}{6\tilde{\delta}}) + 1/2) + \\
& \sum_{\{i \in P^*:\, \hat{\Delta}_i \leq 4D+\alpha\} \cup \{i \in [K] \setminus P^*:\, \Delta_i \leq 4D+\alpha\}} 2\tau_{(\alpha)}(\beta_{\bar{t},\epsilon} \log(\frac{\tau^2_{(\alpha)} MK\pi^2}{6\tilde{\delta}}) + 1/2) \\
& \leq K\tau_{(\alpha)}(2\beta_{\bar{t},\epsilon} \log(\frac{\tau^2_{(\alpha)} MK\pi^2}{6\tilde{\delta}}) + 1) \ .
\end{aligned}
$$

The above theorem gives the most general expression for the sample complexity of R-PSI VAR. Also, as stated below, the Corollary 1 holds for the R-PSI VAR as well since it has the same sample complexity bound as R-PSI.

Remark 5. *The Corollary 1 holds for R-PSI VAR.*

In the following, we divide the proof of Theorem 2 into two parts as we did in the R-PSI proof.

4.1.1 First Part of the Proof of Theorem 2

We first note that Lemma 3 of R-PSI holds for R-PSI VAR as well since the number of samples taken in a sampling round remains the same. Lemma 4 and 5 also holds since the Elimination and O_2 conditions remains the same as in the R-PSI. The rest of the analysis we give below however, differs from the analysis of R-PSI due to the change of the O_1 condition in the new algorithm. As we did in the R-PSI analysis, we assume that "the good event" of Lemma 3 holds in all the analysis of R-PSI VAR that follows.

First, we show that we can only give a $(4D + \alpha)$ suboptimality guarantee for the predicted arms compared to the $(2D + \alpha)$ guarantee of the R-PSI.

Lemma 11. *(Lemma 6 Equivalent for R-PSI VAR.) P can only contain arms that are $(4D + \alpha)$-suboptimal.*

Proof. An arm in S needs to be first moved to O_1 to end up in P. Denote the Pareto optimal arms in S by $S^{(p)}$. If an arm $i \in S$ is moved to O_1, then:

$$\nexists j \in S^{(p)} : \hat{m}_i + D - U_i + \alpha \not\prec \hat{m}_j - D + U_j \ ,$$

which implies:

$$\forall j \in S^{(p)},\ \exists d_j \in [M] : \hat{m_i}^{d_j} + D - U_i + \alpha \geq \hat{m_j}^{d_j} - D + U_j \ .$$

By Lemma 3:

$$\begin{aligned} \forall j \in S^{(p)},\ \exists d_j \in [M] : (m_i^{d_j} + D + U_i) + D - U_i + \alpha \geq \\ \hat{m_i}^{d_j} + D - U_i + \alpha \geq \\ \hat{m_j}^{d_j} - D + U_j \geq \\ (m_j^{d_j} - D - U_j) - D + U_j \ . \end{aligned}$$

Hence:

$$\forall j \in S^{(p)},\ \exists d_j \in [M] : m_i^{d_j} + 2D + \alpha \geq m_j^{d_j} - 2D \ . \tag{4.1}$$

Next, consider the set that consists of the optimal arms in P, which we denote by $P^{(p)}$. By Lemma 5:

$$\forall j \in P^{(p)},\ \exists d_j \in [M] : m_i^{d_j} + D + \alpha \geq m_j^{d_j} - D\ . \tag{4.2}$$

Note that at any round, $S^{(p)} \cup P^{(p)} = P^*$ since Lemma 4 implies that a Pareto optimal arm is either in S or P. Hence, combining (4.1) and (4.2), we obtain:

$$\forall j \in P^*,\ \exists d_j \in [M] : m_i^{d_j} + 2D + \alpha \geq m_j^{d_j} - 2D\ .$$

By Definition 2, this means that $\Delta_i \leq 4D + \alpha$. Hence, we conclude that if arm i is moved to P, it is $(4D + \alpha)$-suboptimal. □

Next, we show that all the Pareto optimal arms are returned in P .

Lemma 12. ***Zero-margin coverage.*** *R-PSI VAR does not eliminate Pareto optimal arms.*

Proof. At elimination step, an arm i is eliminated only if:

$$\exists j \in S : \hat{m}_i + D + U_i \prec \hat{m}_j - D - U_j\ .$$

By Lemma 3 this implies:

$$(m_i - D - U_i) + D + U_i \prec (m_j + D + U_j) - D - U_j\ .$$

By simplifying the above, we obtain:

$$m_i \prec m_j\ .$$

Therefore, i cannot be Pareto optimal and the result follows. □

Next, we prove a termination condition for R-PSI VAR.

Lemma 13. *R-PSI VAR is guaranteed to terminate when* $\forall i \in S, U_i \leq \alpha/4$.

Proof. Suppose that $|S| \geq 2$. Consider an arm $i \in S$ that is not eliminated at round t. Then, the elimination condition implies:

$$\forall j \in S \setminus \{i\},\ \exists d_j : \hat{m_i}^{d_j} + D + U_i \geq \hat{m_j}^{d_j} - D - U_j \ .$$

Using the fact that $U_i, U_j \leq \alpha/4$, we obtain from the above:

$$\forall j \in S \setminus \{i\},\ \exists d_j : \hat{m_i}^{d_j} + D - U_i + \alpha \geq \hat{m_j}^{d_j} - D + U_j \ .$$

Hence, i is moved to O_1. Now, we show that i is moved to O_2. First, we note that the following holds for arms that are left in S after the elimination step:

$$\forall j \in S \setminus \{i\},\ \hat{m}_j + D + U_j \not\prec \hat{m}_i - D - U_i \ .$$

By definition of domination, this implies that for any $j \in S \setminus \{i\}$, either:

$$\exists d_j : \hat{m_j}^{d_j} + D + U_j > \hat{m_i}^{d_j} - D - U_i \ ,$$

or

$$\forall d : \hat{m_j}^{d} + D + U_j = \hat{m_i}^{d} - D - U_i \ .$$

Again considering that $U_i, U_j \leq \alpha/4$, we obtain from the first case above:

$$\exists d_j : \hat{m_j}^{d_j} + D - U_j + \alpha > \hat{m_i}^{d_j} - D + U_i \ ,$$

and for the second case:

$$\forall d : \hat{m_j}^{d} + D - U_j + \alpha = \hat{m_i}^{d} - D + U_i \ .$$

This implies:

$$\forall j \in S \setminus \{i\},\ \hat{m}_j + D - U_j + \alpha \not\prec \hat{m}_i - D + U_i \ ,$$

and therefore, i is moved to O_2. Hence, we conclude that an arm is either eliminated or moved to O_2 when $U_i \leq \alpha/4,\ \forall i \in S$. □

First part of the proof of Theorem 2:

By Remark 2 and Lemma 13:

$$\tau_i \leq \inf\{\tau : U_\tau \leq \alpha/4\},\ \forall i \in [K]\ .$$

Hence, by (3.24) the sample complexity N of R-PSI can be bounded as follows:

$$N \leq K\big(2\tau_{(\alpha)}(\beta_{\bar{t},\epsilon}\log(\frac{\tau_{(\alpha)}^2 MK\pi^2}{6\tilde{\delta}}) + 1/2\big)$$

$$= K\tau_{(\alpha)}(2\beta_{\bar{t},\epsilon}\log(\frac{\tau_{(\alpha)}^2 MK\pi^2}{6\tilde{\delta}}) + 1)\ .$$

The accuracy and coverage guarantees follow from Lemma 11 and 12 respectively.

Now, we continue with the second part of the proof.

4.1.2 Second Part of the Proof of Theorem 2

Lemma 8 holds for R-PSI VAR.

Proof. The proof is the same. □

Lemma 9 holds for R-PSI VAR.

Proof. The proof is the same. The only difference is that now we use Lemma 11 instead of Lemma 6 in the proof and we do not have to consider the case $U_i \leq \alpha/4,\ \forall i \in S$ separately. □

Lemma 10 holds for R-PSI VAR.

Proof. Note that (3.29) and (3.30) together imply that arm i is moved to O_1 in the case of R-PSI VAR as well (since O_1 condition is less strict in R-PSI VAR case). Also, we do not have to consider the case $U_i \leq \alpha/4,\ \forall i \in S$ separately. The rest of the proof is the same. □

Second part of the proof of Theorem 2:

Since Lemma 9 and 10 holds, and since we can bound the sampling round of other arms that do not satisfy the conditions of Lemma 9 and 10 by τ_α, we get the same sample complexity bound expression as in (3.32). □

Note that although the sample complexity bound of R-PSI VAR derived above is the same as the R-PSI bound, in practice, R-PSI VAR generally tends to have significantly smaller sample complexity than R-PSI (See Chapter 5). We think that this is due to the less strict O_1 set condition in the identification step of R-PSI VAR.

4.2 Concluding Remarks

R-PSI VAR might be preferable over R-PSI if we want to guarantee to return all the Pareto optimal arms in our prediction set. We propose the following example scenario where R-PSI VAR might be more useful than R-PSI. Suppose that we have a readily-available corrupted dataset obtained from some previous experiments and suppose that we have the ability to perform new experiments in an adversary-free environment. This scenario might have occurred, for instance, due to a faulty experimental setup that interferes with the reward outcomes. After noticing the flaw, it might be possible to conduct new experiments with the corrected setup. In another case, a new measurement technique might have been invented that removes some adversarial noise that was present in the previous experiments. In both of these scenarios, instead of discarding the outdated dataset entirely, and starting the experiments all over again, one can extract some promising arms from the corrupted dataset and use only these arms in the new

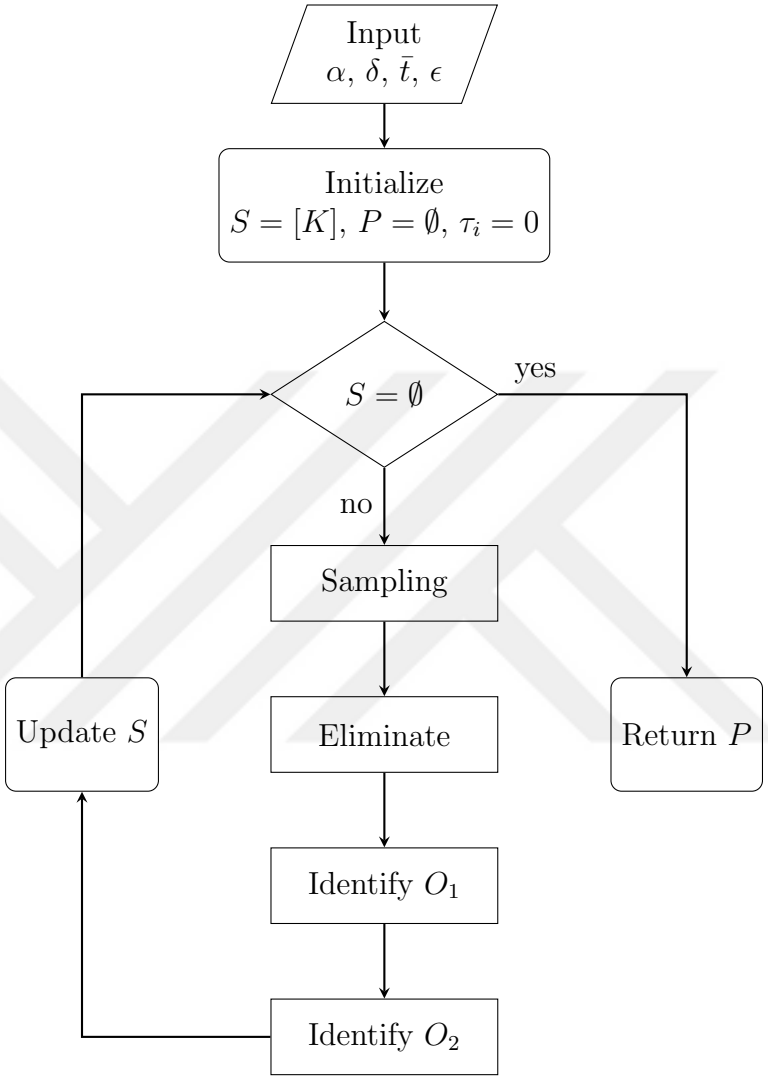


Figure 4.1: Simplified flowchart of R-PSI VAR

adversary-free experiments to decrease the experiment cost induced by arm samplings. In this scenario, the R-PSI VAR might be preferable over R-PSI in the extraction of the promising arms (especially if the adversarial effect was strong) since it is guaranteed to include the Pareto optimal set in its prediction. Then, the Pareto optimal arms can be detected in the adversary free experiments, which results in fully accurate detection of the Pareto set overall. On the other hand, if one uses R-PSI to extract arms from the corrupted dataset, the overall accuracy guarantee of the arms obtained from the new experiments is restricted to $(2D + \alpha)$.

Algorithm 2 R-PSI VAR

Input: α, δ, ϵ, $\bar{t}$
Initialize: $S = [K]$, $P = \emptyset$, $\tau_i = 0 \ \ \forall i \in [K]$, $t = 0$.
$\tau_i \leftarrow \tau_i + 1 \ \forall i \in [K]$
Sample each arm n_0 times as in (3.1)
Update U_{i,τ_i} according to (3.8) $\ \forall i \in S$
Update $\hat{m}_i \ \ \forall i \in S$

while $S \neq \emptyset$ **do**

if $t > 0$ **then**

Sampling:
Choose arm $i^* = \arg\max_{k \in [K]} U_{k,\tau_k}$
$\tau_{i^*} \leftarrow \tau_{i^*} + 1$
Sample i^* successively $n_{i^*,\tau_{i^*}}$ times as in (3.4)
Update $U_{i^*,\tau_{i^*}}$ according to (3.8)
Update $\hat{m}_i^*$

end if

Elimination:
$S \leftarrow S \setminus \{i \in S | \, \exists j \in S \setminus \{i\} : \hat{m}_i + D + U_i \prec \hat{m}_j - D - U_j\}$

Identification:
$O_1 \leftarrow \{i \in S | \, \nexists j \in S \setminus \{i\} : \hat{m}_i + D - U_i + \alpha \prec \hat{m}_j - D + U_j\}$

$O_2 \leftarrow \{i \in O_1 | \, \nexists j \in S \setminus \{i\} : \hat{m}_j - U_j + \alpha \prec \hat{m}_i + U_i\}$

$S \leftarrow S \setminus O_2$

$P \leftarrow P \cup O_2$

return P

$t \leftarrow t + 1$

end while
return P

Chapter 5

Numerical Results

In this chapter, we compare R-PSI and R-PSI VAR with the Algorithm 1 from [1] which considers the Pareto set approximation problem for adversary free MO-MAB setting. Algorithm 1 of [1] is a Pareto accurate algorithm in the adversary free setting (in the case when $D = 0$) and its ranking of the arms is based on the mean of the distributions instead of the median as in this work. For Algorithm 1 to be comparable, we consider Gaussian reward distributions which have the same median and mean values. We conduct experiments on synthetic Gaussian reward distributions, SW-LLVM dataset, and the data obtained from the UVA/PADOVA Type 1 Diabetes Simulator [43]. Note that in [1], the success condition set for Algorithm 1 differs from the accuracy and coverage requirements we use to define Pareto accurate algorithms in the adversarial MO-MAB setting. In particular, their success condition requires for the predicted arms to be at least α-accurate and all the Pareto optimal arms to be returned in the predicted set. In terms of our accuracy and coverage arguments, this success condition is equivalent to an α-accuracy and 0 margin coverage requirement. For a fair comparison, while evaluating Algorithm 1, we evaluate the success of the predicted set returned by Algorithm 1 on $(2D + \alpha)$ accuracy and $(2D + \alpha)$ coverage requirement which is equivalent to the success requirements set for R-PSI and which is looser than their own success conditions.

Table 5.1: *Synthetic Gaussian experiment results.* RSR: Ratio of successful runs, AS: Average Samples, RO: Ratio of optimal arms returned by the algorithm, VSC: Arms that violate success condition item 2.

	ϵ	RSR	AS	RO	VSC
R-PSI	0	1	922.5	1	0
	0.05	1	1,519	1	0
	0.1	1	1,675	1	0
	0.15	1	1,789.7	1	0
	0.2	1	2,240.6	1	0
	0.25	1	2,706.3	1	0
	0.3	1	3,343.1	1	0
	0.35	1	4,957	1	0
	0.4	1	9,951.5	1	0
R-PSI VAR	0	1	933.8	1	0
	0.05	1	734.1	1	0
	0.1	1	831.1	1	0
	0.15	1	985.3	1	0
	0.2	1	1,207.1	1	0
	0.25	1	1,538.4	1	0
	0.3	1	2,077.5	1	0
	0.35	1	3,327.8	1	0
	0.4	1	7,056.3	1	0
Auer	0	1	29,165.3	1	0
	0.05	1	17,878	1	0
	0.1	0.8	9,734.3	0.92	0
	0.15	0.9	7,857	0.97	0.2
	0.2	0.9	18,362.2	0.95	0.1
	0.25	0.5	15,728.4	0.82	0.6
	0.3	0.4	21,943.9	0.85	0.5
	0.35	0.5	35,297.5	0.93	0.5
	0.4	0.2	29,698.1	0.85	1.2

5.1 Experiments on Synthetic Gaussian Distributions

First, we conduct experiments on synthetically created Gaussian reward distributions. For each ϵ, we run algorithms for 10 independent iterations. The reward

samples are generated from Gaussian reward distributions with 0.1 standard deviation. We create 10 arms with 2 objectives. The reward medians (or equivalently the means) are generated randomly from the interval $[0, 10]$ using a uniform distribution. We create a synthetic adversarial attack that contaminates with a fixed contamination value of 1 when the arms are suboptimal and -1 when the arms are Pareto optimal. This adversarial attack fits the prescient type of attack model and therefore we use the parameters that correspond to the prescient model. We choose $\bar{t} = 0.49$ and maximum ϵ value of 0.4 so that $\epsilon \leq \frac{2\bar{t}}{1+2\bar{t}}$ is satisfied. We choose $\alpha = 0.1$ and $\delta = 0.1$. We report the results in Table 5.1. From left to right, columns display the attack probability, the ratio of successful runs (the runs in which coverage and accuracy requirements are satisfied), the total number of samples taken by the algorithm averaged over 10 runs, the ratio of optimal arms returned in P to the total number of optimal arms, and the average number of arms returned in P that are not $(2D + \alpha)$-accurate. In the case of $\epsilon = 0$, the results are consistent with the theoretical expectations. When ϵ exceeds 0.25, it can be seen that Algorithm 1 success ratio drops drastically while R-PSI and R-PSI VAR consistently return successful predictions for all ϵ values.

5.2 Experiments on SW-LLVM Dataset

Next, we use the SW-LLVM dataset from [44] which consists of 1023 compiler settings characterized by 11 binary features. The objectives are the performance and memory footprint of a software program when compiled with these settings. Similar to [1], to obtain a stochastic-like data for our algorithm, we use the combinations of 4 of the binary features to form 16 arms. We ignore other 7 binary features. At the end, we obtain 64 samples for each arm. We simulate an arm pull by randomly selecting one of the 64 samples belonging to that arm and returning the corresponding objectives as the reward sample. We standardize the data to obtain a similar range in both objectives. We use $R(t)$ of Example 1 with a σ value equal to the standard deviation of the standardized SW-LLVM data. We take the mean of 64 data points assigned to an arm and use this as the mean (median) of the corresponding reward distribution. We pick σ in (2.3) as 0.2,

Table 5.2: *SW-LLVM experiment results.*

	ϵ	**RSR**	**AS**	**RO**	**VSC**
R-PSI	0	1	1,530.1	1	0
	0.05	1	54,888.8	1	0
	0.1	1	57,742.3	1	0
	0.15	1	62,109.4	1	0
	0.2	1	67,331.3	1	0
	0.25	1	83,658.8	1	0
	0.3	1	98,361.6	1	0
	0.35	1	134,267	1	0
	0.4	1	208,561	1	0
R-PSI VAR	0	1	1,617	1	0
	0.05	1	10,043	1	0
	0.1	1	7,861.6	1	0
	0.15	1	7,235.2	1	0
	0.2	1	12,489.8	1	0
	0.25	0	20,421.1	1	1
	0.3	0	10,986.7	1	1
	0.35	0	11,023.2	1	1
	0.4	1	45,481.5	1	0
Auer	0	1	7,176.5	1	0
	0.05	0.8	211,242	0	0.1
	0.1	0.9	180,127	0	0.1
	0.15	0.7	117,769	0	0.1
	0.2	1	88,274.8	0	0
	0.25	0.4	56,406.9	0	0.2
	0.3	0.8	43,183.8	0	0
	0.35	0.9	36,553.6	0	0.2
	0.4	0.7	28,129.2	0	0.4

which fits the data well. We use the same type of adversarial model as in the synthetic Gaussian setting with a contamination value of ± 10. We choose $\delta = 0.1$ and $\alpha = 0.1$. We use $\bar{t} = 0.49$ and the maximum ϵ value of 0.4 as in the previous setting. For each ϵ value we test, we run algorithms for 10 independent iterations. We report the SW-LLVM results in Table 5.2. For some of the ϵ-values, R-PSI VAR returns 1 arm that is not $(2D+\alpha)$ suboptimal and fails to return a successful prediction. This is expected since we can only guarantee $(4D + \alpha)$-suboptimality for R-PSI VAR predictions. Similar to the synthetic Gaussian setting, Algorithm 1 of [1] fails for most of the ϵ values.

Table 5.3: *Blood Glucose Simulation results.*

	ϵ	RSR	AS	RO	VSC
R-PSI	0.0	1.0	904.4	1.0	0.0
	0.05	1.0	1039.2	1.0	0.0
	0.1	1.0	1280.0	1.0	0.0
	0.15	1.0	1514.2	1.0	0.0
	0.2	1.0	2031.9	1.0	0.0
	0.25	1.0	3039.7	1.0	0.0
	0.30	1.0	7151.9	1.0	0.0
	0.35	1.0	29171.7	1.0	0.0
	0.4	1.0	632488.3	1.0	0.0
R-PSI VAR	0.0	1.0	913.2	1.0	0.0
	0.05	1.0	1049.0	1.0	0.0
	0.1	1.0	1247.4	1.0	0.0
	0.15	1.0	1831.0	1.0	0.0
	0.2	1.0	2293.1	1.0	0.0
	0.25	1.0	3860.6	1.0	0.0
	0.30	1.0	3536.3	1.0	0.0
	0.35	0.8	8512.2	1.0	0.2
	0.4	1.0	9204.5	1.0	0.0
Auer	0.0	1.0	13.5	1.0	0.0
	0.05	0.9	14.4	1.0	0.1
	0.1	0.6	14.6	0.95	0.3
	0.15	0.5	14.3	0.95	0.4
	0.2	0.6	15.7	0.95	0.3
	0.25	0.2	14.3	0.8	0.7
	0.30	0.1	12.8	0.7	1.0
	0.35	0.3	15.3	0.8	0.8
	0.4	0.2	14.3	0.75	0.4

5.3 Experiments on UVA/PADOVA Type 1 Diabetes Simulator

Next, we run experiments on The UVA/PADOVA Type 1 Diabetes Simulator from [45] which simulates the blood glucose (BG) levels of type 1 diabetes patients over time. We choose an adult patient and use the parameters corresponding to this patient in the simulator. We aim to determine the insulin doses that keep

the patient's BG levels close to the optimal level as much as possible. Specifically, we try to optimize the BG levels of the patient measured after 90 and 180 minutes following an insulin injection. The objectives are calculated based on the difference between BG measurements and the ideal BG level, which is deemed 140 dL/mg for the chosen patient. To be precise, the objectives are computed in a way that an insulin dose which results in a BG measurement that is close to the ideal BG level is assigned a large objective value. We choose 11 different insulin doses and assign each of them to an arm to test the effect of each dose on the BG level of the patient. Given an insulin dose level and the time that passes after the injection, the simulator computes a deterministic value for the BG level based on various patient-dependent parameters [43]. For a more realistic simulation model, we add a Gaussian noise that has zero mean and 1 standard deviations to the deterministic result computed by the simulator to simulate the errors in the measurement device. We also model the effect of other factors such as stress, insomnia, and heavy physical activity which are events that occur occasionally, with an oblivious adversary that chooses a fixed uniform contamination distribution. The choice for the parameters α, δ and $\bar{t}$ is the same as in the previous settings. For each ϵ value, we run algorithms for 10 iterations. We report the average results in Table 5.3. Further experiments on UVA/PADOVA simulator are given in Appendix B.

Overall, we conclude that Algorithm 1 of [1] fails drastically when faced with an adversary. On the other hand, the proposed R-PSI and R-PSI VAR consistently return accurate predictions for all ϵ value in agreement with the theoretical predictions. Also, R-PSI VAR overall makes smaller number of samplings as expected since the identification step in R-PSI VAR is less strict.

Chapter 6

Conclusion and Future Work

We considered the Pareto set identification problem under contaminated bandit feedback. We proposed a robust algorithm that can make accurate predictions when subjected to strong adversaries. We proved an upper sample complexity bound that depends on the accuracy parameter α inverse squarely. This upper bound matches the dependence of the bounds proved in previous studies that consider adversary-free setting. We further provided a tighter upper bound that depends on α and some geometric properties of the median reward vectors. Our numerical results prove that our theoretical analysis holds as expected and shows the inability of the multi-objective methods developed for the adversary-free setting to handle strong adversarial attacks.

Interesting future research directions would be to extend our analysis to large-scale MO-MAB problems with correlated arms and consider the robust optimization for multi-armed bandits with hard safety constraints.

Bibliography

[1] P. Auer, C.-K. Chiang, R. Ortner, and M. Drugan, "Pareto front identification from stochastic bandit feedback," in *Proc. 19th Int. Conf. Artif. Intell. Stat.*, vol. 51, pp. 939–947, 2016.

[2] E. Even-Dar, S. Mannor, and Y. Mansour, "PAC bounds for multi-armed bandit and Markov decision processes," in *Computational Learning Theory* (J. Kivinen and R. H. Sloan, eds.), (Berlin, Heidelberg), pp. 255–270, Springer Berlin Heidelberg, 2002.

[3] J. Gittins, K. Glazebrook, and R. Weber, *Multi-armed bandit allocation indices*, ch. 1, pp. 1–17. John Wiley & Sons, 2011.

[4] D. A. Berry and B. Fristedt, *Bandit problems: Sequential Allocation of Experiments*, ch. 1, pp. 1–8. Monographs on Statistics and Applied Probability, Springer, 1985.

[5] A. Slivkins, "Introduction to multi-armed bandits," 2019.

[6] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge University Press, 2020.

[7] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, pp. 285–294, 12 1933.

[8] W. R. Thompson, "On the theory of apportionment," *Am. J. Math.*, vol. 57, no. 2, pp. 450–456, 1935.

[9] H. Robbins, "Some aspects of the sequential design of experiments," *Bull. Am. Math. Soc.*, vol. 58, no. 5, pp. 527–535, 1952.

[10] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, pp. 4–22, Mar. 1985.

[11] S. Bubeck, R. Munos, and G. Stoltz, "Pure exploration for multi-armed bandit problems," 2010.

[12] K. Van Moffaert, K. Van Vaerenbergh, P. Vrancx, and A. Nowe, "Multi-objective χ-armed bandits," in *2014 Proc. Int. Jt. Conf. Neural Netw. (IJCNN)*, pp. 2331–2338, 2014.

[13] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine Learning algorithms," in *Adv. Neural Inf. Process. Syst.* (F. Pereira, C. Burges, L. Bottou, and K. Weinberger, eds.), vol. 25, pp. 2951–2959, Curran Associates, Inc., 2012.

[14] J. Snoek, H. Larochelle, and R. Adams, "Opportunity cost in Bayesian optimization," in *NIPS Workshop on Bayesian Optimization, Sequential Experimental Design, and Bandits*, 2011.

[15] Z. Guan, K. Ji, D. J. Bucci Jr., T. Y. Hu, J. Palombo, M. Liston, and Y. Liang, "Robust stochastic bandit algorithms under probabilistic unbounded adversarial attack," *AAAI Conf. Artif. Intell.*, vol. 34, pp. 4036–4043, Apr. 2020.

[16] J. Altschuler, V.-E. Brunel, and A. Malek, "Best arm identification for contaminated bandits," *J. Mach. Learn. Res.*, vol. 20, no. 91, pp. 1–39, 2019.

[17] c. Ararat and C. Tekin, "Vector optimization with stochastic bandit feedback," 2021.

[18] M. M. Drugan and A. Nowe, "Designing multi-objective multi-armed bandits algorithms: A study," in *Proc. Int. Jt. Conf. Neural Netw.*, pp. 1–8, 2013.

[19] S. Q. Yahyaa and B. Manderick, "Thompson sampling for multi-objective multi-armed bandits problem," in *Proc. Eur. Symp. Artif. Neural Netw. Comput. Intell. Mach. Learn.*, pp. 47–52, 2015.

[20] S. Yahyaa, M. Drugan, and B. Manderick, "Knowledge gradient for multi-objective multi-armed bandit algorithms," in *ICAART 2014 - Proc. 6th Int. Conf. Agents Artif. Intell. (ICAART)*, vol. 1, pp. 74–83, 2014.

[21] N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger, "Information-theoretic regret bounds for Gaussian process optimization in the bandit setting," *IEEE Trans. Inf. Theory*, vol. 58, p. 3250–3265, May 2012.

[22] S. Belakaria, A. Deshwal, and J. R. Doppa, "Max-value entropy search for multi-objective Bayesian optimization," in *Adv. Neural Inf. Process. Syst.* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.

[23] S. Belakaria, A. Deshwal, N. K. Jayakodi, and J. R. Doppa, "Uncertainty-aware search framework for multi-objective Bayesian optimization," *Proc. AAAI Conf. Artif. Intell.*, vol. 34, pp. 10044–10052, Apr. 2020.

[24] A. Nika, K. Bozgan, S. Elahi, Ç. Ararat, and C. Tekin, "Pareto active learning with Gaussian processes and adaptive discretization," 2020.

[25] M. Zuluaga, A. Krause, and M. Püschel, "ϵ-pal: An active learning approach to the multi-objective optimization problem," *J. Mach. Learn. Res.*, vol. 17, no. 104, pp. 1–32, 2016.

[26] J. M. Hernández-Lobato, M. W. Hoffman, and Z. Ghahramani, "Predictive entropy search for efficient global optimization of black-box functions," in *Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 27, 2014.

[27] A. Shah and Z. Ghahramani, "Pareto frontier learning with expensive correlated objectives," in *Proc. 33rd Int. Conf. Mach. Learn.* (M. F. Balcan and K. Q. Weinberger, eds.), vol. 48 of *PMLR*, (New York), pp. 1919–1927, PMLR, 20–22 June 2016.

[28] V. Kuleshov and D. Precup, "Algorithms for multi-armed bandit problems," 2014.

[29] E. Brochu, M. W. Hoffman, and N. de Freitas, "Portfolio allocation for Bayesian optimization," 2010.

[30] W. Shen, J. Wang, Y.-G. Jiang, and H. Zha, "Portfolio choices with orthogonal bandit learning," in *24th IJCAI Int. Jt. Conf. Artif. Intell.*, pp. 974–980, 2015.

[31] B. Awerbuch and R. D. Kleinberg, "Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches," in *Proc. 36th Annu. ACM Symp. Theory Comput.*, p. 45–53, 2004.

[32] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proc. 19th World Wide Web Conf.*, p. 661–670, 2010.

[33] S. Pandey, D. Agarwal, D. Chakrabarti, and V. Josifovski, "Bandits for taxonomies: A model-based approach," in *Proc. 2007 SIAM Int. Conf. Data Min.*, pp. 216–227, 2007.

[34] S. S. Villar, J. Bowden, and J. Wason, "Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges," *Stat. Sci.*, vol. 30, no. 2, pp. 199–215, 2015.

[35] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM J. Comput.*, vol. 32, no. 1, pp. 48–77, 2002.

[36] G. Stoltz, *Incomplete information and internal regret in prediction of individual sequences.* Theses, Université Paris Sud - Paris XI, May 2005.

[37] J.-Y. Audibert and S. Bubeck, "Minimax policies for adversarial and stochastic bandits," in *Proc. 22nd Annu. ACM Symp. Theory Comput. (COLT)*, vol. 7, pp. 217–226, Jan. 2009.

[38] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," 2012.

[39] T. Lykouris, V. Mirrokni, and R. Paes Leme, "Stochastic bandits robust to adversarial corruptions," in *Proc. 50th Annu. ACM Symp. Theory Comput.*, STOC 2018, (New York), p. 114–122, Association for Computing Machinery, 2018.

[40] A. Gupta, T. Koren, and K. Talwar, "Better algorithms for stochastic bandits with adversarial corruptions," in *Proc. 32nd Conf. on Learn. Theory (COLT)* (A. Beygelzimer and D. Hsu, eds.), vol. 99 of *Proc. Mach. Learn. Res.*, pp. 1562–1578, PMLR, 25–28 June 2019.

[41] S. Kapoor, K. K. Patel, and P. Kar, "Corruption-tolerant bandit learning," *Mach. Learn.*, vol. 108, no. 4, pp. 687–715, 2019.

[42] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM J. Comput.*, vol. 32, no. 1, pp. 48–77, 2002.

[43] C. D. Man, F. Micheletto, D. Lv, M. Breton, B. Kovatchev, and C. Cobelli, "The UVA/PADOVA type 1 diabetes simulator: new features," *J. Diabetes Sci. Technol.*, vol. 8, no. 1, pp. 26–34, 2014.

[44] M. Zuluaga, G. Sergent, A. Krause, and M. Püschel, "Active learning for multi-objective optimization," in *Proc. 30th Int. Conf. Mach. Learn.* (S. Dasgupta and D. McAllester, eds.), vol. 28 of *Proc. Mach. Learn. Res.*, (Atlanta, GA), pp. 462–470, PMLR, 17–19 June 2013.

[45] X. Jinyu, "Simglucose v0.2.1." `https://github.com/jxx123/simglucose`, 2018.

Appendix A

Table of Notation

Table A.1: *Table of Notation*

Symbol	Description
K	Number of arms
M	Number of objectives
F_i^d	Uncontaminated reward distribution for arm i and objective d.
Y_i^d	Random variable with distribution F_i^d
G_i^d	Contamination distribution corresponding to objective d and arm i.
Z_i^d	Random variable with distribution G_i^d
$\tilde{F}_i^d$	Contaminated reward distribution.
$\tilde{Y}_i^d$	Random variable with distribution $\tilde{F}_i^d$
B_i^d	Random variable with distribution $\text{Ber}(\epsilon)$. Determines whether contamination occurs or not.
R	A non-decreasing function that bound the quantile deviation
$\bar{t}$	Maximum range of R
α	Accuracy requirement for prediction set P
ϵ	Adversarial attack probability
δ	Confidence probability
U_{τ_i}	Statistical uncertainty of arm i at sampling round τ_i.
D	Unavoidable bias
m_i^d	Median of interest corresponding to F_i^d
m_i	Median of interest vector associated with arm i
$\hat{m}_i^d$	Empirical median corresponding to $\tilde{F}_i^d$
P^*	Pareto optimal arm set
P	Predicted set
τ_i	Number of sampling rounds performed on arm i.
n_{τ_i}	Number of samples performed for arm i at the sampling round τ_i
n_0	Initial number of samplings
N_{τ_i}	Number of total samples made for arm i up until and including the sampling round τ_i
$\Delta_{i,j}$	Suboptimality gap of a suboptimal arm i with respect to a Pareto optimal arm j (Definition 1)
Δ_i	Suboptimality gap of arm i
$\hat{\Delta}_{i,j}$	Domination gap of optimal arm i with respect to arm j
$\hat{\Delta}_i$	Decision gap of an optimal arm i
$\beta_{\bar{t},\,\epsilon},\ h_\epsilon,\ \tilde{\delta}$	Adversary dependent constants

Appendix B

Further Experiments on UVA/PADOVA Simulator

Experiments are run for 5 different ϵ and α values. Average results over 10 iterations are given in the following tables.

Table B.1: *Ratio of successful runs (RSR) for R-PSI*

ϵ \ α	0.1	0.2	0.3	0.5	1.0
0.05	1	1	1	1	1
0.1	1	1	1	1	1
0.2	1	1	1	1	1
0.3	1	1	1	1	1
0.4	1	1	1	1	1

Table B.2: *RSR for R-PSI VAR*

ϵ \ α	0.1	0.2	0.3	0.5	1.0
0.05	1	1	1	1	1
0.1	1	1	1	1	1
0.2	1	1	1	1	1
0.3	1	1	0.9	1	1
0.4	1	1	1	1	1

Table B.3: *RSR for Auer*

ϵ \ α	0.1	0.2	0.3	0.5	1.0
0.05	0.8	0.9	0.6	0.9	0.9
0.1	0.7	0.7	0.7	0.7	0.5
0.2	0.4	0.5	0.6	0.5	0.5
0.3	0.3	0.2	0.6	0.8	0.3
0.4	0.5	0.3	0.2	0.4	0.2

Table B.4: *Average Samples (AS) for R-PSI VAR*

ϵ \ α	0.1	0.2	0.3	0.5	1.0
0.05	1058.8	1027.3	1027.3	1037.1	1028.7
0.1	1246.2	1211.2	1199.5	1256.2	1257.6
0.2	1999.3	2112.3	1898.2	2579.1	2329.1
0.3	7303.5	8448.5	7713.6	5990.4	4378.7
0.4	880355	198326	57059.3	23856.8	9996.1

Table B.5: *AS for RPSI-VAR*

ϵ \ α	0.1	0.2	0.3	0.5	1.0
0.05	1119.7	1038.5	1039.2	1089.3	1119
0.1	1269.3	1296.3	1233.1	1255.5	1233.1
0.2	1989.1	1864.2	2154.2	2218.9	2121.5
0.3	4595.4	7120.7	5020.5	4988.1	3773.6
0.4	9778	9604.2	8282.3	9988	7477.5

Table B.6: *AS for Auer*

ϵ \ α	0.1	0.2	0.3	0.5	1.0
0.05	14.6	14.8	16.1	14.2	15.2
0.1	15.1	14.9	15.3	15.2	12.8
0.2	17.3	13.8	13.1	15.1	14.4
0.3	14.9	14.7	16.1	14.3	15.4
0.4	16.3	14.6	14	13.4	13.6

Table B.7: *Ratio of optimal arms returned (RO) for R-PSI*

ϵ \ α	0.1	0.2	0.3	0.5	1.0
0.05	1	1	1	1	1
0.1	1	1	1	1	1
0.2	1	1	1	1	1
0.3	1	1	1	1	1
0.4	1	1	1	1	1

Table B.8: *RO for R-PSI VAR*

ϵ \ α	0.1	0.2	0.3	0.5	1.0
0.05	1	1	1	1	1
0.1	1	1	1	1	1
0.2	1	1	1	1	1
0.3	1	1	1	1	1
0.4	1	1	1	1	1

Table B.9: *RO for Auer*

ϵ \ α	0.1	0.2	0.3	0.5	1.0
0.05	0.95	1	1	1	0.95
0.1	1	1	0.95	0.85	0.85
0.2	0.8	0.95	0.85	0.85	0.8
0.3	0.8	0.85	0.95	0.9	0.85
0.4	0.95	0.75	0.75	0.85	0.7

Table B.10: *Arms that violate success condition 2 (VSC) for R-PSI*

ϵ \ α	0.1	0.2	0.3	0.5	1.0
0.05	0	0	0	0	0
0.1	0	0	0	0	0
0.2	0	0	0	0	0
0.3	0	0	0	0	0
0.4	0	0	0	0	0

Table B.11: *VSC for R-PSI VAR*

ϵ \ α	0.1	0.2	0.3	0.5	1.0
0.05	0	0	0	0	0
0.1	0	0	0	0	0
0.2	0	0	0	0	0
0.3	0	0	0.1	0	0
0.4	0	0	0	0	0

Table B.12: *VSC for Auer*

ϵ \ α	0.1	0.2	0.3	0.5	1.0
0.05	0.2	0.1	0.4	0.1	0
0.1	0.3	0.3	0.2	0.2	0.5
0.2	0.3	0.6	0.2	0.3	0.2
0.3	0.5	0.7	0.5	0.1	0.6
0.4	0.6	0.4	0.6	0.5	0.7