Bandits Algorithms for Multi-Objective Best Arm Identification

CYRILLE KONE

Master's Thesis

submitted in fulfilment of the requirements for

Master's degree

Department of mathematics
Ecole Normale Supérieure *of* Paris-Saclay
September 2022

Contents

1 Introduction

This report summarizes my research internship at *Inria ScooL*. Under the supervision of Emilie Kaufmann, I have been working on bandits algorithms for best arm identification. This topic has applications in a wide range of areas, including crowdsourcing, A/B testing for designing products, online ad selection and clinical trials for drug development and dose identification ([Karnin et al., 2013, Aziz et al., 2019]). In bandits, the best arm is traditionally defined as the arm with largest expected reward. But in the context of medical trials or A/B testing for example, the efficacy of a medical treatment or a product is often accessed by measuring several indicators, which would lead to an underlying multi-dimensional reward vector (see e.g [Katz-Samuels and Scott, 2018] for an application). The main goal of the internship is to propose new algorithms for multi-objective pure exploration, that may be phrased as identifying a Pareto front between the different objectives, with high probability and using as few samples from the arms as possible. We will in particular consider the fixed-budget setting [Audibert and Bubeck, 2010] in which the maximal number of samples is fixed and we seek to minimize the probability of error.

The report introduces some tools of interest in the sequel before diving into two chapters where some algorithms are provided and a last chapter with some applications. The last part summarizes my personal contributions before concluding.

1.1 Bandit model

A stochastic K-armed bandit model ν is parameterized K distributions $\nu_1, \dots, \nu_K$ called *arms* or *actions* with means (respectively) $\mu_1, \dots, \mu_K$. These distributions are unknown to the forecaster. At each round $t = 1 \dots T$, (T possibly infinite) the forecaster selects one arm A_t in the set $\mathbb{A} := \{1, \dots, K\}$ and observes a reward Z_t drawn from ν_{A_t} independently from the past (actions and observations). In the *pure exploration* or *best arm identification* setting, the goal of the forecaster is to identify the *best arm*, that is the arm $a^\star$ satisfying $\mu_{a^\star} = \max_{a \in \mathbb{A}} \mu_a$, which we suppose unique for simplicity. At the end of the T rounds, the forecaster selects an arm, denoted b_T, which she believes to be the best arm. We denote by $\mathbb{P}_\nu$ (resp. $\mathbb{E}_\nu$) the probability law (resp. expectation) of the stochastic process Z_t and by $\mathcal{F}^\nu$, the filtration $(\mathcal{F}_t)_{t=1,\dots,T}$ where $\mathcal{F}_t := \sigma(A_1, Z_1, \dots, A_t, Z_t)$. The problem of *best arm identification* has been studied in two distinct settings in the literature.

In the *fixed confidence* setting (see e.g., [Garivier and Kaufmann, 2016, Jamieson et al., 2013, Kaufmann and Kalyanakrishnan, 2013]), T is infinite and the forecaster tries to minimize the number of rounds needed to achieve a fixed-confidence $\delta \in (0, 1)$ about the identification of the best arm while minimizing the expected number of samples or rounds needed, that is ensure

$$\mathbb{P}_\nu(\tau_\delta < \infty, b_{\tau_\delta} \neq a^\star) \leq \delta \qquad \text{while minimizing } \mathbb{E}_\nu(\tau_\delta),$$

where τ_δ is a $\mathcal{F}^\nu$-stopping time. $\mathbb{E}_\nu(\tau_\delta)$ is also called the *sample complexity*. In the *fixed-budget* setting (see e.g [Audibert and Bubeck, 2010, Karnin et al., 2013, Gabillon et al., 2012]) which is studied here, the number of rounds T is fixed (finite) and known by the forecaster. The objective is to maximize the probability of returning the best arm, that is maximize

$$\mathbb{P}_\nu(a^\star = b_T). \tag{1.1}$$

Letting $a^\star = 1$ (without loss of generality), the hardness of the task will be characterized by

$$H = \sum_{k \in \mathbb{A}} \Delta_k^{-2} \qquad \text{and} \qquad H_2 = \max_{k \in \mathbb{A}} k \Delta_k^{-2},$$

introduced in the seminal paper [Audibert and Bubeck, 2010], where for any arm $a \neq a^\star$, $\Delta_a := \mu_{a^\star} - \mu_a$ and the arms are ordered as

$$\Delta_{a^\star} = \Delta_1 := \Delta_2 \leq \Delta_3 \leq \dots \leq \Delta_K.$$

The authors proved that theses quantities satisfies

$$H_2 \leq H \leq \log(2K)H_2 \tag{1.2}$$

These quantities, as we will see in the sequel, give an order of magnitude of the number of rounds needed to identify the best arm with a reasonable probability.

The following section recalls some notions in vector optimization that will allow to precisely (re-)define problem (1.1)when the distributions are multi-dimensional.

1.2 Vector optimization

In this section we recall some notions in vector optimization and we prove two lemmas that will be used all along the report. Letting $D \geq 1$, the bold symbol will denote $\mathbb{R}^D$ vectors. For any vector $\mathbf{X} \in \mathbb{R}^D$, for any $d \in [D]$, X^d denotes the $d-$th coordinate of $\mathbf{X}$.

Definition 1.1 (Weak dominance, Pareto dominance and strong dominance, [Auer et al., 2016, Drugan and Nowe, 2013])**.** *Let $D \geq 1$. Let $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^D$.*

a) $\boldsymbol{\mu}_2$ weakly dominates $\boldsymbol{\mu}_1$ (or $\boldsymbol{\mu}_1 \preceq \boldsymbol{\mu}_2$) if

$$\forall\, d \in [D], \mu_1^d \leq \mu_2^d,$$

b) $\boldsymbol{\mu}_2$ (Pareto) dominates $\boldsymbol{\mu}_1$ (or $\boldsymbol{\mu}_1 \prec \boldsymbol{\mu}_2$) if $\boldsymbol{\mu}_1 \preceq \boldsymbol{\mu}_2$ and

$$\exists\, d \in [D] : \mu_1^d < \mu_2^d,$$

c) $\boldsymbol{\mu}_2$ strongly dominates $\boldsymbol{\mu}_2$ (or $\boldsymbol{\mu}_1 < \boldsymbol{\mu}_2$) if

$$\forall\, d \in [D], \mu_1^d < \mu_2^d.$$

When a) (respectively b) or c)) does not hold, we may write $\boldsymbol{\mu}_1 \npreceq \boldsymbol{\mu}_2$ (respectively $\boldsymbol{\mu}_1 \nprec \boldsymbol{\mu}_2$ or $\boldsymbol{\mu}_1 \not< \boldsymbol{\mu}_2$).

Remark 1.1. *The weak dominance defines a partial order on $\mathbb{R}^D$. The Pareto dominance is transitive and non-reflexive (follows form the definition).*

The following definition recalls the notion of *Pareto optimal set* which is also called the *Pareto front* (see [Drugan and Nowe, 2013]).

Definition 1.2 (Pareto optimal set, [Drugan and Nowe, 2013]). *Let $D \geq 1$. Let $\mathcal{S} \subset \mathbb{R}^D$ be non-empty. The Pareto optimal set of $\mathcal{S}$ denoted as $P^\star(\mathcal{S})$ is the set of non dominated vectors of $\mathcal{S}$ that is the set*

$$P^\star(\mathcal{S}) := \{\boldsymbol{\mu} \in \mathcal{S} : \forall \mathbf{x} \in \mathcal{S}, \mathbf{x} \not\prec \boldsymbol{\mu}\}.$$

The vectors of the *Pareto optimal set* are called the (Pareto) optimal vectors and any vector which does not belong to the *Pareto optimal set* is called a *(Pareto) sub-optimal vector*. The following lemma shows that for finite set of vectors, the *Pareto optimal set* always exist.

Lemma 1.1 (Existence of the Pareto optimal set). *Let $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$ be vectors in $\mathbb{R}^D$. Then, the Pareto optimal set of $\mathcal{S} := \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$ is non-empty.*

Proof. Assume that any vector $\boldsymbol{\mu} \in \mathcal{S}$ is Pareto dominated (that is the Pareto optimal set is empty). Since the Pareto dominance is transitive and non-reflexive (follows from the definition), without loss of generality, there exists a permutation σ such that $\boldsymbol{\mu}_{\sigma(1)} \preceq \boldsymbol{\mu}_{\sigma(2)} \preceq \cdots \preceq \boldsymbol{\mu}_{\sigma(K)}$. Therefore since $\mathbf{x} \prec \mathbf{x}' \Longrightarrow \mathbf{x}' \not\prec \mathbf{x}$ (follows from the definition), $\boldsymbol{\mu}_{\sigma(K)}$ could not be dominated, which contradict the initial assumption. □

The following lemma which may be of independent interest is extensively used in the sequel.

Lemma 1.2 (Domination of sub-optimal points). *Let $\mathcal{S} := \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$. Let $\boldsymbol{\mu} \in \mathcal{S}$ be a sub-optimal vector. There exists a Pareto-optimal vector $\boldsymbol{\mu}' \in \mathcal{S}$ such that $\boldsymbol{\mu} \preceq \boldsymbol{\mu}'$.*

Proof. Suppose that there are $k < n$ dominated vectors in $\mathcal{S}$. Without loss of generality, we may assume they are $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$. Let $\boldsymbol{\mu}_{i_1}, i_1 \leq k$. Suppose that no Pareto-optimal vector dominates $\boldsymbol{\mu}_{i_1}$. Since $\boldsymbol{\mu}_{i_1}$ is not optimal, by the latter assumption, there exists $i_2 \leq k$ such that $\boldsymbol{\mu}_{i_1} \prec \boldsymbol{\mu}_{i_2}$. If $\boldsymbol{\mu}_{i_2}$ is dominated by a Pareto-optimal vector, this vector also dominates $\boldsymbol{\mu}_{i_1}$ (by transitivity) then we have a contradiction. If not, there exits $i_3 \leq k$ such that $\boldsymbol{\mu}_{i_1} \preceq \boldsymbol{\mu}_{i_2} \preceq \boldsymbol{\mu}_{i_3}$. Again we can use the same arguments as before for i_3. In any case we should stop in at most k steps, otherwise we would have $\boldsymbol{\mu}_{i_1} \prec \boldsymbol{\mu}_{i_2} \prec \cdots \prec \boldsymbol{\mu}_{i_k}$ and $\boldsymbol{\mu}_{i_k}$ should be dominated by a Pareto-optimal vector, otherwise it would be itself Pareto-optimal, which is not the case. □

Letting $\nu(K, D)$ be a bandit with means $(\boldsymbol{\mu}_i)_{i \in [K]}$, the Pareto optimal arms $a^\star(\nu)$ is the set of arms whose mean are in the *Pareto front* of

$$\mathcal{S} := \{\boldsymbol{\mu}_i : i \in [K]\},$$

that is $i \in a^\star(\nu)$ if and only if $\boldsymbol{\mu}_i \in P^\star(\mathcal{S})$. The initial problem (1.1) in the fixed-budget setting rephrases as: given a budget $T \in \mathbb{N}$,

$$\text{maximize} \qquad \mathbb{P}_\nu(a^\star(\nu) = \widehat{a}^\star(\nu, T)), \tag{1.3}$$

where $\widehat{a}^\star(\nu, T)$ denotes the set recommended by the forecaster at the end of the T rounds. In particular, the sequential interaction described in the introduction remains unchanged.

In the sequel, for any arms $i, j \in [K]$, we write $i\mathcal{R}j$ if $\boldsymbol{\mu}_i \mathcal{R} \boldsymbol{\mu}_j$ for $\mathcal{R}$ being one of $\preceq, \not\preceq, \prec, \not\prec, <, \not<$.

1.3 Concentration inequalities

In the main body of the report, we will be deriving results for subgaussian bandits. A bandit $\nu(K, D)$ is subgaussian if its arms are subgaussian that is

Definition 1.3 (Subgaussian random variable, [Lattimore and Szepesvári, 2020])**.** *letting* $\mathbf{X}$ *be a* $\mathbb{R}^D$*-valued random variable and* $\sigma \geq 0$*.* $\mathbf{X}$ *is* σ*-subgaussian if*

$$\forall\, \mathbf{u} \in \mathbb{R}^D,\ \mathbb{E}(\exp\left(\mathbf{u}^\top(\mathbf{X} - \mathbb{E}(\mathbf{X}))\right)) \leq \exp\left(\frac{\sigma^2}{2}\|\mathbf{u}\|_2^2\right).$$

This naturally includes Gaussian distributions (by direct computation) and bounded random variables through the Hoeffding lemma, which states that

Lemma 1.3 (Hoeffding lemma [Lattimore and Szepesvári, 2020])**.** *letting* $a < b \in \mathbb{R}$ *and* X *a centered real-valued random variable almost surely bounded in* $[a, b]$*,*

$$\forall\, u \in \mathbb{R}\ ,\ \mathbb{E}(\exp(uX) \leq \exp\left(\frac{(b-a)^2}{8}u^2\right).$$

Thus, for any random variable X *almost surely (a.s) bounded in* $[a, b]$*,* $X - \mathbb{E}(X)$ *is a centered variable almost surely bounded in* $[a - \mathbb{E}(X), b - \mathbb{E}(X)]$ *such that Lemma 1.3 applies to* $X - \mathbb{E}(X)$*,so* X *is* $\frac{b-a}{2}-$*subgaussian.*

Using Lemma 1.3 and Cauchy-Schwartz inequality yields that any bounded multidimensional random variable is also subgaussian.

Lemma 1.4 (Subgaussianity of multivariate)**.** *Let* $a < b \in \mathbb{R}$ *and* $D \geq 1$*. Any* $\mathbb{R}^D-$*valued random variable* $\mathbf{X}$ *almost surely bounded in* $[a, b]^D$ *is subgaussian.*

Proof. The proof follows by Cauchy-Schwartz inequality and Lemma 1.3. For any $\mathbf{u} \in \mathbb{R}^D$,

$$\begin{aligned}
\mathbb{E}(\exp(\mathbf{u}^\top(\mathbf{X} - \mathbb{E}(\mathbf{X})))) &= \mathbb{E}\left(\prod_{d=1}^{D} \exp(u^d(X^d - \mathbb{E}(X^d)))\right), \\
&\leq \mathbb{E}(\exp(2u^1(X^1 - \mathbb{E}(X^1))))^{1/2} \times \mathbb{E}\left(\prod_{d=2}^{D} \exp(2u^d(X^d - \mathbb{E}(X^d)))\right)^{1/2}, \\
&\leq \exp\left(\frac{(b-a)^2}{4}(u^1)^2\right) \times \mathbb{E}\left(\prod_{d=2}^{D} \exp(2u^d(X^d - \mathbb{E}(X^d)))\right)^{1/2},
\end{aligned}$$

□

where the last two inequalities follow from Cauchy-Schwartz and Lemma 1.3. By iterating the same procedure on the right-hand side expectation, one could eventually write

$$\mathbb{E}(\exp(\mathbf{u}^\top(\mathbf{X} - \mathbb{E}(\mathbf{X})))) \leq \exp(\sigma^2 \frac{\|\mathbf{u}\|}{2}),$$

where σ^2 depends on $(b-a)^2$.

Remark that the decomposition is order dependent. But when the components of the vector are independent, the following result will allow to compute σ^2 exactly.

Lemma 1.5 (Subgaussianity is linearly preserved)**.** *Let $\mathbf{X}_1, \dots \mathbf{X}_K$ be respectively $\sigma_1, \dots, \sigma_K-$subgaussian $\mathbb{R}^D-$valued, independent random variables. For any sequence $\alpha_1, \dots, \alpha_K \in \mathbb{R}$, $\alpha_1\mathbf{X}_1 + \dots + \alpha_K\mathbf{X}_K$ is $\sqrt{\sigma_1^2 + \dots \sigma_K^2}-$subgaussian. Furthermore, for any $\boldsymbol{\theta} \in \mathbb{R}^D$, $\boldsymbol{\theta}^\top\mathbf{X}_1$ is $\sigma_1\|\boldsymbol{\theta}\|_2-$subgaussian.*

Proof. The proof of the first point can be found in [Lattimore and Szepesvári, 2020]. The second point simply follows as

$$\begin{aligned}
\mathbb{E}(\exp(t\boldsymbol{\theta}^\top(\mathbf{X}_1 - \mathbb{E}(\mathbf{X}_1)))) &= \mathbb{E}(\exp((t\boldsymbol{\theta})^\top(\mathbf{X}_1 - \mathbb{E}(\mathbf{X}_1)))), \\
&\leq \exp(\|t\boldsymbol{\theta}\|_2^2 \frac{\sigma_1^2}{2}), \quad (\ \mathbf{X}_1 \text{ is } \sigma_1-\text{subgaussian}) \\
&= \exp(\frac{t^2}{2}\|\boldsymbol{\theta}\|_2^2\sigma_1^2).
\end{aligned}$$

□

Remark 1.2. *Using the latter lemma, when a $\mathbb{R}^D-$ valued random variable $\mathbf{X}$ is almost surely bounded in $[a,b]^D$ and its components are independent, for any $\mathbf{u} \in \mathbb{R}^D$, $\mathbf{u}^\top(\mathbf{X} -$*

$\mathbb{E}(\mathbf{X})$) *is a linear combination of subgaussian variables, hence* $\mathbf{X}$ *is subgaussian with* $\sigma^2 = \frac{D(b-a)^2}{4}$.

The following lemma upper-bounds the probability that a subgaussian variable deviates from its expectation.

Lemma 1.6 (Subgaussian Hoeffding's inequality, [Lattimore and Szepesvári, 2020])**.** *Let* X *be a* $\sigma-$*subgaussian random variable. The subgaussian inequality states that, for any* $\Delta \geq 0$,

$$\mathbb{P}(X - \mathbb{E}(X) \geq \Delta) \leq \exp\left(-\frac{\Delta^2}{2\sigma^2}\right).$$

Remark 1.3. *If* X *is* $\sigma-$*subgaussian, then* $-X$ *is* $\sigma-$*subgaussian, which yields the well-known symmetric form of Lemma 1.6,*

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq \Delta) \leq 2\exp\left(-\frac{\Delta^2}{2\sigma^2}\right).$$

This inequality is central as it allows to recover Hoeffing's inequality. Indeed if $X_1, \ldots, X_n$ are *i.i.d* variables bounded in $[0,1]$, so $\frac{1}{2}-$subgaussian, $X := n^{-1}\sum_{i=1} X_i$ is $\sqrt{\frac{1}{4n}}-$subgaussian (by Lemma 1.5) so, Lemma 1.3 yields for any $\Delta \geq 0$,

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq \Delta) \leq 2\exp(-2n\Delta^2).$$

The subgaussian inequality will be extensively used in the sequel to upper-bound the probability of some key events. Together with this fundamental inequality, we recall another extensively used result. The following concentration inequality whose proof can be found in many applied probability books uses Ville's inequality to upper-bound the probability that a martingale crosses a line.

Lemma 1.7 (Subgaussian maximal inequality, [Locatelli et al., 2016])**.** *Let* $X_1, \ldots, X_n \sim \mathbf{X}$ *be i.i.d real-valued* $\sigma-$*subgaussian random variables. Letting for any* $k \leq n, S_k := \sum_{t=1}^{k} X_t$*, the subgaussian maximal inequality states that, for any* $\Delta \geq 0$

a) $\mathbb{P}(\exists\, k \leq n : S_k - k\mathbb{E}(X) \geq \Delta) \leq \exp\left(-\frac{\Delta^2}{2n\sigma^2}\right)$,

b) $\mathbb{P}(\exists\, k \leq n : S_k - k\mathbb{E}(X) \leq -\Delta) \leq \exp\left(-\frac{\Delta^2}{2n\sigma^2}\right)$,

c) $\mathbb{P}(\exists\, k \leq n : |S_k - k\mathbb{E}(X)| \geq \Delta) \leq 2\exp\left(-\frac{\Delta^2}{2n\sigma^2}\right)$.

Using these tools, our goal in the next chapters will be to provide some algorithms and their theoretical guarantees for the problem (1.3).

2

Best Arm Identification

In this chapter, given a subgaussian bandit $\nu(K, D)$ with means $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$, and a function $f : \mathbb{R}^D \longrightarrow \mathbb{R}$ such that there exists $a^\star \in [K]$ satisfying

$$f(\boldsymbol{\mu}_{a^\star}) > \max_{k \in [K] \setminus a^\star} f(\boldsymbol{\mu}_k), \tag{2.1}$$

we are interested in identifying this arm $a^\star$ by using the sequential interaction described in the introduction for the *fixed-budget* setting . While this problem may seem independent from problem (1.3), we will see in this chapter that when the function f is well-chosen, $a^\star \in a^\star(\nu)$, the Pareto optimal set of arms.

2.1 Problem setup

Let $K, D \geq 1$. Let $\nu := (\nu_1, \ldots, \nu_K)$ a subgaussian bandit problem where the ν_i's are independent distributions. We denote the expectations of those distributions by $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K \in \mathbb{R}^D$, that is

$$\mathbb{E}_{\mathbf{X} \sim \nu_i}(\mathbf{X}) = \boldsymbol{\mu}_i.$$

At round t, the forecaster selects an arm A_t from $\mathbb{A} := \{1, \ldots, K\}$ and observes a reward $\mathbf{X}_t$ drawn independently from ν_{A_t}. For any round t, for any arm a, let $T_a(t)$ denotes the

number of times arm a has been selected up to time t, that is

$$T_a(t) := \sum_{s=1}^{t} \mathbb{1}_{\{A_s=a\}},$$

and, letting $s \geq 1$, $\mathbf{X}_{a,s}$ denotes the $s-$th reward collected from arm a. Let $\widehat{\boldsymbol{\mu}}_{a,s}$ denotes the empirical mean of arm a after s observations,

$$\widehat{\boldsymbol{\mu}}_{a,s} := \frac{1}{s}\sum_{u=1}^{s} \mathbf{X}_{a,u},$$

and, let $\widehat{\boldsymbol{\mu}}_a(t)$ denotes the empirical mean reward of arm a, at time t,

$$\widehat{\boldsymbol{\mu}}_a(t) := \widehat{\boldsymbol{\mu}}_{a,T_a(t)} = \frac{1}{T_a(t)}\sum_{s=1}^{t} \mathbf{X}_s \mathbb{1}_{\{A_s=a\}} = \frac{1}{T_a(t)}\sum_{u=1}^{T_a(t)} \mathbf{X}_{a,u}.$$

We recall that for a vector $\mathbf{X} \in \mathbb{R}^D$ (boldfaced), for any $d \in [D]$, X^d denotes the d-th component of $\mathbf{X}$. Letting $\mathcal{S} := \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$, $a^\star(\nu)$, the Pareto optimal set of arms of the bandit ν is the set of arms whose expectation are in the *Pareto front* of $\mathcal{S}$. For any function $f : \mathbb{R}^D \longleftarrow \mathbb{R}$, satisfying (2.1), we will denote by $a^\star(\nu, f)$ the unique arm

$$\operatorname*{argmax}_{k\in\mathbb{A}} f(\boldsymbol{\mu}_k),$$

and when it is clear from the context, this arm will be simply denoted $a^\star$. Letting $a^\star(\nu, f) = 1$ (without loss of generality), the hardness of the task will be characterized by

$$H^f = \sum_{k\in\mathbb{A}} (\Delta_k^f)^{-2} \qquad \text{and} \qquad H_2^f = \max_{k\in\mathbb{A}} k(\Delta_k^f)^{-2}, \tag{2.2}$$

analogous to the quantities introduced in the seminal paper [Audibert and Bubeck, 2010] (and recalled in (1.2)), where for any arm $a \neq a^\star$, $\Delta_a^f := f(\boldsymbol{\mu}_{a^\star}) - f(\boldsymbol{\mu}_a)$ and the arms are ordered as

$$\Delta_{a^\star}^f = \Delta_1^f := \Delta_2^f \leq \Delta_3^f \leq \cdots \leq \Delta_K^f.$$

Similarly to (1.2), it holds that

$$H_2^f \leq H^f \leq \log(2K) H_2^f. \tag{2.3}$$

2.2 Successive Rejects(f)

This algorithm slightly modifies *Successive Rejects*, which is a fixed-budget algorithm introduced in [Audibert and Bubeck, 2010] for solving problem (1.1) with an upper-bound on its probability of mis-identification of the best arm. Given a $K-$armed (unvariate) bandit, the algorithm proceeds in $K-1$ phases. Initially, all the arms are set active. For each phase, the still active arms are uniformly sampled and the arm with the lowest empirical mean is deactivate at the end of the current phase. After $K-1$ phases, only one arm remains active, it is recommended as the best arm. Letting f be a real-valued function and $\nu(K, D)$ a $D-$variates bandit, algorithm 1 details the algorithm we call Successive Rejects (f) . When $D = 1$ and f is the identity function, this algorithm perfectly matches the one introduced in [Audibert and Bubeck, 2010]. Since Succes-

Algorithm 1: Successive Rejects(f).

Input : Bandit $\nu(K, D)$, budget $T \geq K$, function f
Initialize: $A_1 := \{1, \dots, K\}, n_0 = 0$
Define : $\overline{\log}(K) := \frac{1}{2} + \sum_{i=2}^{K} i^{-1}, n_k := \left\lceil \frac{1}{\overline{\log}(K)} \frac{T-K}{K+1-k} \right\rceil$
for $k \leftarrow 1$ **to** $K-1$ **do**
  foreach $a \in A_k$ **do**
    pull arm a for $n_k - n_{k-1}$ rounds;
  $A_{k+1} \leftarrow A_k \setminus \underset{a \in A_k}{\operatorname{argmin}} f(\widehat{\boldsymbol{\mu}}_{a,n_k})$;
$\widehat{a}(\nu, f, T) \leftarrow A_K$;
Output : $\widehat{a}(\nu, f, T)$

sive Rejects (f) is similar to *Successive Rejects*, except for the elimination condition, algorithm 1 does not exceed the budget $T \geq K$ (see [Audibert and Bubeck, 2010] for the proof). Additionally, Theorem 2.1 shows that when f satisfies the conditions below (Definition 2.1), the probability of error of Successive Rejects (f) can be upper bounded using a proof similar to the one in [Audibert and Bubeck, 2010].

Definition 2.1 (Class of function Γ)**.** *Let $\nu(K, D)$ be a bandit instance. Let $\Gamma(\nu, c_1, c_2)$ denotes the class of functions $f : \mathbb{R}^D \longrightarrow \mathbb{R}$ such that there exists $a^\star \in \mathbb{A}$ satisfying*

$$f(\boldsymbol{\mu}_{a^\star}) > \max_{k \in \mathbb{A} \setminus a^\star} f(\boldsymbol{\mu}_k),$$

and enjoying the following property,

$$\forall a \in \mathbb{A}\ , \forall s \geq 1,\ \mathbb{P}_\nu(f(\widehat{\boldsymbol{\mu}}_{a,s}) \geq f(\widehat{\boldsymbol{\mu}}_{a^\star,s})) \leq c_1 \exp\left(-\frac{s(\Delta_a^f)^2}{c_2}\right),$$

where c_1, c_2 *are positive (possibly dimension-dependent) constants and* $\Delta_a^f := f(\boldsymbol{\mu}_{a^\star}) - f(\boldsymbol{\mu}_a)$.

Theorem 2.1 (Bounding the probability of error of algorithm 1)**.** *Let* $\nu(K, D)$ *be a bandit,* $f : \mathbb{R}^D \longrightarrow \mathbb{R}$ *and* $T \geq K$*. Let* $c_1, c_2 > 0$*. If* $f \in \Gamma(\nu, c_1, c_2)$*, the probability of mis-identification of Successive Rejects* (f) *satisfies*

$$\mathbb{P}_\nu(a^\star \neq \widehat{a}^\star) \leq K\frac{K-1}{2}c_1 \exp\left(-\frac{1}{c_2}\frac{T-K}{H_2^f \overline{\log}(K)}\right).$$

Proof. The proof is inspired by the proof of Successive Rejects given in [Audibert and Bubeck, 2010].

Step 0: Setup and notations. Without loss of generality, assume $a^\star := a^\star(\nu, f) = 1$ is the unique optimal arm and $\Delta_1^f := \Delta_2^f \leq \cdots \leq \Delta_K^f$. Let $\widehat{a}^\star := \widehat{a}^\star(\nu, f, T)$ be the arm output by algorithm 1. Let $e_T(\nu, f) := \mathbb{P}_\nu(\widehat{a}^\star \neq a^\star)$. Let $\xi_k(a)$ be the event

"arm a is dismissed at the end of round k".

Finally, assume that the quantity $\widehat{\boldsymbol{\mu}}_{a,n_k}$ is always defined even if arm a has been dismissed before round k.

Step 1: A key event. We have

$$\{\widehat{a}^\star \neq a^\star\} = \bigcup_{k=1}^{K-1} \xi_k(a^\star).$$

Then,

$$\mathbb{P}_\nu(\widehat{a}^\star \neq a^\star) = \sum_{k=1}^{K-1} \xi_k(a^\star).$$

At the beginning of round k, the algorithm has discarded $k-1$ arms already. Therefore, it remains in the set A_k, one the k worst arms (ordered by the Δ_a^f's). Let i denotes this arm. If arm $a^\star$ is dismissed at the end of round k, then it should be empirically not better than i, that is

$$\xi_k(a^\star) \subset \left\{\exists\, i \in \{K+1-k, \ldots, K\} \text{ s.t } f(\widehat{\boldsymbol{\mu}}_{i,n_k}) \geq f(\widehat{\boldsymbol{\mu}}_{a^\star,n_k})\right\},$$

which yields,

$$e_T(\nu, f) \leq \sum_{k=1}^{K-1} \sum_{i=K+1-k}^{K} \mathbb{P}_\nu(f(\widehat{\boldsymbol{\mu}}_{i,n_k}) \geq f(\widehat{\boldsymbol{\mu}}_{a^\star,n_k})).$$

Since $f \in \Gamma(\nu, c_1, c_2)$,

$$\begin{aligned} e_T(\nu, f) &\leq \sum_{k=1}^{K-1} \sum_{a=K+1-k}^{K} c_1 \exp\left(-\frac{1}{c_2} n_k (\Delta_a^f)^2\right), \\ &\leq \sum_{k=1}^{K-1} k c_1 \exp\left(-\frac{1}{c_2} n_k (\Delta_{K+1-k}^f)^2\right), \end{aligned}$$

where the last inequality follows by $\Delta_{K+1-k}^f \leq \cdots \leq \Delta_K^f$.

Step 2: Conclusion. Recall that

$$H_2^f := \max_{k \in \mathbb{A}} k(\Delta_k^f)^{-2},$$

then,

$$\begin{aligned} n_k(\Delta_{K+1-k}^f)^2 &= \left\lceil \frac{1}{\overline{\log}(K)} \frac{T-K}{K+1-k} \right\rceil (\Delta_{K+1-k}^f)^2, \\ &\geq \frac{1}{\overline{\log}(K)} \frac{T-K}{(K+1-k)(\Delta_{K+1-k}^f)^{-2}}, \\ &\geq \frac{1}{\overline{\log}(K)} \frac{T-K}{H_2^f}. \end{aligned}$$

Putting things together,

$$e_T(\nu, f) \leq \sum_{k=1}^{K-1} k c_1 \exp\left(-\frac{1}{c_2} \frac{T-K}{H_2^f \overline{\log}(K)}\right),$$

which finally yields

$$e_T(\nu, f) \leq K \frac{K-1}{2} c_1 \exp\left(-\frac{1}{c_2} \frac{T-K}{H_2^f \overline{\log}(K)}\right). \tag{2.4}$$

□

Remark 2.1 (Univariate bounded bandit)**.** *When ν is $K-$armed univariate bandit*

bounded in $[0,1]$*, letting* f *be the identity function, Hoeffding's inequality yields*

$$\mathbb{P}_\nu(f(\widehat{\mu}_{a,n_k}) \geq f(\widehat{\mu}_{a^\star,n_k})) \leq \exp\left(-\frac{1}{2}n_k\Delta_a^2\right),$$

then

$$e_T(\nu, f) \leq K\frac{K-1}{2}\exp\left(-\frac{1}{2}\frac{T-K}{H_2^f\overline{\log}(K)}\right),$$

which is the result proved in [Audibert and Bubeck, 2010].

2.3 Sequential Halving (f)

This algorithm is built-upon Sequential Halving (introduced in [Karnin et al., 2013]), a well-known algorithm in the fixed-budget setting. Similarly to Successive Rejects , Sequential Halving proceeds by successive elimination phases. But differently to Successive Rejects , given a (univariate) $K-$armed bandit, the algorithm deactivate the worst half of the arms at the end of each phase, hence the word *halving*. Another difference to note is that at the beginning of each phase, Successive Rejects does not reuse the samples collected so far, that is the empirical means are computed by using only the pulls done during the current phase. Sequential Halving (f) builds upon this strategy by eliminating the worst half of the arms, in terms of the plugin values $f(\widehat{\boldsymbol{\mu}})$, f being a real-valued function. algorithm 2 depicts the strategy. While Sequential Halving has been initially analyzed for scalar $K-$armed bandit, we have extended the proof to any function $f : \mathbb{R}^D \longrightarrow \mathbb{R}$.

Algorithm 2: Sequential Halving (f).

Input : Bandit $\nu(K, D)$, budget $T \geq K$, function f
Initialize: $S_0 := \{1, \ldots, K\}$
Define : $n_r := \left\lfloor \frac{T}{|S_r|\lceil\log_2(K)\rceil} \right\rfloor$
for $r \leftarrow 0$ **to** $\lceil\log_2 K\rceil - 1$ **do**
  foreach $a \in S_r$ **do**
    pull arm a for n_r rounds;
    update $\widehat{\boldsymbol{\mu}}^r_{i,n_r} \leftarrow \frac{1}{n_r}\sum_{p=1}^{n_r}\mathbf{X}^r_{i,p}$;
  $S_{r+1} \leftarrow$ the set of $\left\lceil \frac{|S_r|}{2} \right\rceil$ arms in S_r with the largest value of $f(\widehat{\boldsymbol{\mu}}^r_{a,n_r})$;
$\widehat{a}^\star(\nu, f, T) \leftarrow S_{\lceil\log_2 K\rceil}$;
Output : $\widehat{a}^\star(\nu, f, T)$

Using the technique developed in [Karnin et al., 2013] (Theorem 4.1 therein), Theorem 2.2 upper-bounds the probability of mis-identification of the best arm of Sequential Halving (f) for functions satisfying Definition 2.1.

Theorem 2.2 (Bounding the probability of error of algorithm 2)**.** *Let $\nu(K, D)$ be a bandit, $f : \mathbb{R}^D \longrightarrow \mathbb{R}$, $K \geq 2$, a doubly power of 2 and $T \geq K$, a power of 2. Let $c_1, c_2 > 0$. If $f \in \Gamma(\nu, c_1, c_2)$, the probability of error of Sequential Halving (f) satisfies*

$$\mathbb{P}_\nu(a^\star \neq \widehat{a}^\star) \leq 3c_1 \log_2(K) \exp\left(-\frac{1}{4c_2} \frac{T}{H_2^f \log_2(K)}\right).$$

Proof. The proof is inspired by the one given in [Karnin et al., 2013].

Step 0: Setup and notations. We reuse some notations introduced in the proof of Theorem 2.1. Without loss of generality, assume $a^\star := a^\star(\nu, f) = 1$ is the unique optimal arm and $\Delta_1^f := \Delta_2^f \leq \cdots \leq \Delta_K^f$. Let $\widehat{a}^\star := \widehat{a}^\star(\nu, f, T)$ be the arm output by algorithm 2. Let $e_T(\nu, f) := \mathbb{P}_\nu(\widehat{a}^\star \neq a^\star)$. Let $\xi_k(a)$ be the event

$$\text{"arm } a \text{ is dismissed at the end of round } k\text{"}.$$

Additionally, for any arm a, let $\widehat{\boldsymbol{\mu}}_{a,n_r}^r$ denotes the empirical average computed on the n_r samples collected in round r. Finally, assume that the quantity $\widehat{\boldsymbol{\mu}}_{a,n_r}^r$ is always defined even if arm a has been dismissed before round r. We will later prove (Lemma 2.1) that $S_{\lceil \log_2 K \rceil}$ contains exactly one arm.

Step 1: A key event. Using the event $\xi_k(a^\star)$ defined earlier, we have

$$e_T(\nu, f) \leq \sum_{r=0}^{\lceil \log_2 K \rceil - 1} \mathbb{P}_\nu(\xi_r(a^\star)).$$

In particular, at the end of round r, arm $a^\star$ is dismissed if there exists at least $\lceil |S_r|/2 \rceil$ arms with larger empirical estimates. Let S_r' denotes the set of arms in S_r except the $\lfloor |S_r|/4 \rfloor$ best arms (ranked by the Δ_a^f's). Let N_r denotes the number of arms in S_r' with empirical estimate larger than $f(\widehat{\boldsymbol{\mu}}_{a^\star,n_r}^r)$. Since $\lceil |S_r|/2 \rceil \geq \lfloor |S_r|/4 \rfloor + \lceil |S_r|/4 \rceil$, if $a^\star$ is dismissed at the end of the round r, there should exists at least $\lceil |S_r|/4 \rceil$ arms in S_r' with empirical estimates larger than $f(\widehat{\boldsymbol{\mu}}_{a^\star,n_r}^r)$. Therefore,

$$\xi_r(a^\star) \subset \{N_r \geq \lceil |S_r|/4 \rceil\}.$$

Step 2: Upper-bounding the probability of $\xi_r(a^\star)$ by Markov's inequality. By

what precedes and by Markov's inequality,

$$e_T(\nu, f) \leq \sum_{r=0}^{\lceil \log_2 K \rceil - 1} \mathbb{P}_\nu(N_r \geq \lceil |S_r|/4 \rceil), \tag{2.5}$$

$$\leq \sum_{r=0}^{\lceil \log_2 K \rceil - 1} \frac{\mathbb{E}_\nu(N_r)}{\lceil |S_r|/4 \rceil}. \tag{2.6}$$

On the other hand,

$$N_r = \sum_{a \in S'_r} \mathbb{I}_{\{f(\widehat{\boldsymbol{\mu}}^r_{a,n_r}) \geq f(\widehat{\boldsymbol{\mu}}^r_{a^\star,n_r})\}}.$$

Since the means have been re-initialized at the beginning of episode r (see, algorithm 2), conditionally on S_r, $(\mathbf{X}_{i,s})_{i \in \mathbb{A}, s \in [T]}$ are still mutually independent, and, $f \in \Gamma(\nu, c_1, c_2)$, hence

$$\begin{aligned}\mathbb{E}_\nu(Nr|S_r) &= \sum_{a \in S'_r} \mathbb{P}_\nu(f(\widehat{\boldsymbol{\mu}}^r_{a,n_r}) \geq f(\widehat{\boldsymbol{\mu}}^r_{a^\star,n_r})|S_r), \\ &\leq \sum_{a \in S'_r} c_1 \exp\left(-\frac{1}{c_2} n_r (\Delta^f_a)^2\right).\end{aligned}$$

Since S'_r does not contain the $\lfloor |S_r|/4 \rfloor$ best arms, letting $i_r = \lceil |S_r|/4 \rceil$ we have (by the ordering)

$$\forall\, i \in S'_r\,, c_1 \exp\left(-\frac{1}{c_2} n_r (\Delta^f_i)^2\right) \leq c_1 \exp\left(-\frac{1}{c_2} n_r (\Delta^f_{i_r})^2\right).$$

Thus,

$$\begin{aligned}\mathbb{E}_\nu(\mathbb{E}_\nu(N_r|S_r)) &\leq \mathbb{E}_\nu(|S'_r| c_1 \exp\left(-\frac{1}{c_2} n_r (\Delta^f_{i_r})^2\right)), \\ &= c_1 \left(|S_r| - \lfloor |S_r|/4 \rfloor\right) \exp\left(-\frac{1}{c_2} n_r (\Delta^f_{i_r})^2\right).\end{aligned} \tag{2.7}$$

Step 3: Conclusion. By what precedes, and assuming that K is a power of 2, it comes

$$\left(|S_r| - \lfloor |S_r|/4 \rfloor\right) / \lceil |S_r|/4 \rceil = 3, \tag{2.8}$$

then, combining (2.7) and (2.6) yields

$$\mathbb{P}_\nu(N_r \geq \lceil |S_r|/4 \rceil) \leq 3 c_1 \exp\left(-\frac{1}{c_2} n_r (\Delta^f_{i_r})^2\right),$$

which by assuming that $T \geq K$ is also a power of 2 rewrites as

$$\mathbb{P}_\nu(N_r \geq \lceil |S_r|/4 \rceil) \leq 3c_1 \exp\left(-\frac{1}{c_2}\frac{T}{4\log_2(K)}\frac{1}{(\Delta^f_{i_r})^{-2} i_r}\right). \tag{2.9}$$

Since $H_2^f := \max_{k\in\mathbb{A}} k(\Delta_k^f)^{-2}$, it comes

$$\mathbb{P}_\nu(N_r \geq \lceil |S_r|/4 \rceil) \leq 3c_1 \exp\left(-\frac{1}{c_2}\frac{T}{4H_2^f \log_2(K)}\right),$$

therefore, using (2.5),

$$\begin{aligned} e_T(\nu, f) &\leq \sum_{r=0}^{\lceil \log_2 K\rceil - 1} 3c_1 \exp\left(-\frac{1}{c_2}\frac{T}{4H_2^f \log_2(K)}\right), \\ &\leq 3c_1 \lceil \log_2(K)\rceil \exp\left(-\frac{1}{c_2}\frac{T}{4H_2^f \log_2(K)}\right). \end{aligned}$$

□

The proof technique could be easily adapted to any value of K and T. In that case, equation (2.8) could be upper-bounded by 4, which holds for any value of K. However, we would need to modify the allocation strategy so that we do not exceed the budget and equation (2.9) still holds. Remark that when whatever T and K, the algorithm as defined in algorithm 2 does not exceed the budget (direct calculation). Lemma 2.1 shows that whatever the value of K, algorithm 2 outputs one arm, showing that the algorithm is well-defined. This was not proved in [Karnin et al., 2013] as it trivially holds when K is a power of 2.

Lemma 2.1 (algorithm 2 outputs exactly one arm)**.** *For any $K \geq 2$, $S_{\lceil \log_2(K)\rceil}$ as defined by algorithm 2 contains exactly one arm.*

Proof. We split the proof into two steps.

Step 1: Let us show that $\frac{1}{2} < |S_{\lceil \log_2 K\rceil}|$. Indeed, we have

$$\begin{aligned} |S_{r+1}| &= \left\lceil \frac{|S_r|}{2} \right\rceil, \\ &\geq \frac{|S_r|}{2}, \\ &\geq \frac{|S_0|}{2^{r+1}} = \frac{K}{2^{r+1}}. \qquad \text{(By induction)} \end{aligned}$$

Therefore,

$$|S_{\lceil \log_2 K \rceil}| \geq \frac{K}{2^{\lceil \log_2 K \rceil}} > \frac{K}{2^{\log_2(K)+1}} = \frac{1}{2}.$$

Step 2: Let us show that $|S_{\lceil \log_2 K \rceil}| \leq 1$. We have

$$\begin{aligned}
|S_{r+1}| &= \left\lceil \frac{|S_r|}{2} \right\rceil, \\
&< \frac{|S_r|}{2} + 1 = \frac{1}{2} \left\lceil \frac{|S_{r-1}|}{2} \right\rceil + 1, \\
&< \left\lceil \frac{|S_{r-1}|}{2^2} \right\rceil + 1, \\
&\leq \left\lceil \frac{|S_{r-1}|}{2^2} \right\rceil, \qquad (\text{since } \frac{1}{2} \lceil x \rceil \leq \left\lceil \frac{x}{2} \right\rceil) \\
&\leq \left\lceil \frac{|S_0|}{2^{r+1}} \right\rceil. \qquad (\text{By induction})
\end{aligned}$$

Therefore,

$$\begin{aligned}
|S_{\lceil \log_2 K \rceil}| &\leq \left\lceil \frac{K}{2^{\lceil \log_2 K \rceil}} \right\rceil, \\
&< \frac{K}{2^{\lceil \log_2 K \rceil}} + 1, \\
&< \frac{K}{2^{\log_2 K}} + 1 = 2, \\
&\leq 1.
\end{aligned}$$

Finally,

$$\frac{1}{2} < |S_{\lceil \log_2 K \rceil}| \leq 1,$$

which achieves the proof. □

Proposition 2.1 (Bounding the probability of error of Uniform allocation)**.** *Let $\nu(K, D)$ be a bandit, let $T \geq K$, $c_1, c_2 > 0$. Let $f : \mathbb{R}^D \longrightarrow \mathbb{R}$. If $f \in \Gamma(\nu, c_1, c_2)$, the algorithm which at each round randomly uniformly pull an arm and recommends the empirical best arm at the end of the $T-$th round satisfies*

$$\mathbb{P}_\nu(a^\star \neq \widehat{a}^\star) \leq (K-1)c_1 \exp\left(-\frac{1}{c_2} \frac{T}{H_2^f \log(2K)K} \right).$$

Proof. Recall that $a^\star := a^\star(\nu, f)$ and $\widehat{a}^\star := \widehat{a}^\star(\nu, f, T)$. We assume that each arm has

been played T/K times. We have

$$\begin{aligned}
\mathbb{P}(a^\star \neq \widehat{a}^\star) &= \sum_{a \neq a^\star} \mathbb{P}_\nu(f(\widehat{\boldsymbol{\mu}}_a(T)) \geq f(\widehat{\boldsymbol{\mu}}_{a^\star}(T))) \\
&\leq \sum_{a \neq a^\star} c_1 \exp(-\frac{1}{c_2}\frac{T(\Delta_a^f)^2}{K}) \qquad \text{since } f \in \Gamma(\nu, c_1, c_2) \\
&\leq (K-1)c_1 \exp(-\frac{1}{c_2}\frac{T}{H^f K}) \qquad (2.10) \\
&\leq (K-1)c_1 \exp(-\frac{1}{Kc_2}\frac{T}{H_2^f \log(2K)}),
\end{aligned}$$

where the last inequality from from (2.3). □

Compared to Successive Rejects (f) and Sequential Halving (f) , there is an additional $1/K$ in the exponential term of the upper-bound of the uniform sampling strategy. While these upper-bounds are not tight, they give an overview of the performance of each algorithm. We may expect Successive Rejects (f) and Sequential Halving (f) to outperform the uniform allocation strategy, at least when the sub-optimality gaps Δ_a^f's are very different. Indeed, if we consider a scenario where $\Delta_1^f \approx \cdots \approx \Delta_K^f$, the trivial (and large) upper-bound $(\Delta_a^f)^{-2} \leq H^f$ (which led to (2.10)) could be replaced by $(\Delta_a^f)^{-2} \approx H^f/K$ which would cancel the $1/K$ in the exponential. Finally, [Carpentier and Locatelli, 2016] proves that without additional information on the problem (e.g the complexity H^f, the mean of the best arm, nearly equal sub-optimality gaps, etc.), Successive Rejects is indeed optimal (see Theorem 1 therein).

So far, we have proved some upper-bounds on the probability of mis-identification of Successive Rejects (f) and Sequential Halving (f) for functions satisfying Definition 2.1. Proposition 2.2 and Proposition 2.3 shows a class of function for which this concentration inequality holds.

Proposition 2.2 (Concentration of weighted sum)**.** *Let $\nu(K, D)$ a $\sigma-$subgaussian bandit. Let $f_{\boldsymbol{\theta}} : \mathbb{R}^D \ni \boldsymbol{\mu} \mapsto \boldsymbol{\theta}^\top \boldsymbol{\mu}$. For any $\boldsymbol{\theta} \in \mathbb{R}^D$, $f_{\boldsymbol{\theta}} \in \Gamma(\nu, 1, 4\sigma^2\|\boldsymbol{\theta}\|_2^2)$.*

Proof. The proof follows by the subgaussian Hoeffding inequality. We have

$$\mathbb{P}_\nu(f_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\mu}}_{a,s}) \geq f_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\mu}}_{a^\star,s})) = \mathbb{P}_\nu(\boldsymbol{\theta}^\top(\widehat{\boldsymbol{\mu}}_{a,s} - \widehat{\boldsymbol{\mu}}_{a^\star,s}) - \boldsymbol{\theta}^\top(\boldsymbol{\mu}_a - \boldsymbol{\mu}_{a^\star}) \geq \Delta_a^{f_\theta}).$$

Since $(\widehat{\boldsymbol{\mu}}_{a,s} - \widehat{\boldsymbol{\mu}}_{a^\star,s})$ is $\sigma\sqrt{2/s}-$subgaussian (Lemma 1.5, sum of independent subgaussian variables), then, $\boldsymbol{\theta}^\top(\widehat{\boldsymbol{\mu}}_{a,s} - \widehat{\boldsymbol{\mu}}_{a^\star,s})$ is $\sigma\|\boldsymbol{\theta}\|_2\sqrt{2/s}-$subgaussian (Lemma 1.5). Therefore,

the subgaussian Hoeffding inequality yields

$$\mathbb{P}_\nu(f_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\mu}}_{a,s}) \geq f_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\mu}}_{a^\star,s})) \leq \exp\left(-\frac{(\Delta_a^{f_{\boldsymbol{\theta}}})^2 s}{4\|\boldsymbol{\theta}\|_2^2\sigma^2}\right).$$

□

Proposition 2.3 (Concentration of Lipschitz function)**.** *Let $f : (\mathbb{R}^D, \|\cdot\|_\infty) \mapsto (\mathbb{R}, |\cdot|)$ be a $\ell-$Lipschitz function. Let $\nu(K, D)$ be a $\sigma-$subgaussian bandit. Then, for any arm a,*

$$\mathbb{P}_\nu(f(\widehat{\boldsymbol{\mu}}_{a,s}) \geq f(\widehat{\boldsymbol{\mu}}_{a^\star,s})) \leq 4D\exp\left(-\frac{s(\Delta_a^f)^2}{8\ell^2\sigma^2}\right),$$

that is $f \in \Gamma(\nu, 4D, 8\ell^2\sigma^2)$.

Proof. In this proof we will show that through the Lipschitz property, this reduces to a concentration inequality on the arms empirical estimates.

Step 1: A useful event. We have

$$f(\widehat{\boldsymbol{\mu}}_{a,s}) \geq f(\widehat{\boldsymbol{\mu}}_{a^\star,s}) \Longleftrightarrow (f(\widehat{\boldsymbol{\mu}}_{a,s}) - f(\boldsymbol{\mu}_a)) + (f(\boldsymbol{\mu}_{a^\star}) - f(\widehat{\boldsymbol{\mu}}_{a^\star,s})) \geq \Delta_a^f.$$

Then, by the pigeonhole principle and since f is $\ell-$Lipschitz,

$$\begin{aligned}
\{f(\widehat{\boldsymbol{\mu}}_{a,s}) \geq f(\widehat{\boldsymbol{\mu}}_{a^\star,s})\} \subset &\{f(\widehat{\boldsymbol{\mu}}_{a,s}) - f(\boldsymbol{\mu}_a) \geq \frac{\Delta_a^f}{2}\} \cup \{f(\boldsymbol{\mu}_{a^\star}) - f(\widehat{\boldsymbol{\mu}}_{a^\star,s}) \geq \frac{\Delta_a^f}{2}\},\\
\subset &\{\|\widehat{\boldsymbol{\mu}}_{a,s} - \boldsymbol{\mu}_a\|_\infty \geq \frac{\Delta_a^f}{2\ell}\} \cup \{\|\boldsymbol{\mu}_{a^\star} - \widehat{\boldsymbol{\mu}}_{a^\star,s}\|_\infty \geq \frac{\Delta_a}{2\ell}\},\\
\subset &\{\exists\, d \in [D] \text{ s.t } |\widehat{\mu}_{a,s}^d - \mu_a^d| \geq \frac{\Delta_a^f}{2\ell}\} \cup \{\exists\, d \in [D] \text{ s.t } |\mu_{a^\star}^d - \widehat{\mu}_{a^\star,s}^d| \geq \frac{\Delta_a^f}{2\ell}\}
\end{aligned}$$

Step 2: Conclusion. By union bound and $\sigma-$subgaussian Hoeffding inequality,

$$\begin{aligned}
\mathbb{P}_\nu(f(\widehat{\boldsymbol{\mu}}_{a,s}) \geq f(\widehat{\boldsymbol{\mu}}_{a^\star,s})) &\leq \mathbb{P}_\nu(\exists\, d \in [D] \text{ s.t } |\widehat{\mu}_{a,s}^d - \mu_a^d| \geq \frac{\Delta_a^f}{2\ell}) + \mathbb{P}_\nu(\exists\, d \in [D] \text{ s.t } |\mu_{a^\star}^d - \widehat{\mu}_{a^\star,s}^d| \geq \frac{\Delta_a^f}{2\ell}),\\
&\leq \sum_{d=1}^{D}\{\mathbb{P}_\nu(|\widehat{\mu}_{a,s}^d - \mu_a^d| \geq \frac{\Delta_a^f}{2\ell}) + \mathbb{P}_\nu(|\mu_{a^\star}^d - \widehat{\mu}_{a^\star,s}^d| \geq \frac{\Delta_a^f}{2\ell})\},\\
&\leq \sum_{d=1}^{D} 4\exp(-\frac{s(\Delta_a^f)^2}{8\ell^2\sigma^2}),\\
&\leq 4D\exp(-\frac{s(\Delta_a^f)^2}{8\ell^2\sigma^2}),
\end{aligned}$$

which achieves the proof. □

Remark 2.2. *Since all norms are equivalent in $\mathbb{R}^D$, Proposition 2.3 can be extended to any Lipschitz function, whatever the norm, at the price of an additional constant in the exponential. For example if $f : (\mathbb{R}^D, \|\cdot\|_2) \longrightarrow \mathbb{R}$ is $\ell-$Lipschitz, since*

$$\forall\, \mathbf{x} \in \mathbb{R}^D, \|\mathbf{x}\|_2^2 \leq D\|\mathbf{x}\|_\infty^2,$$

f will be $\ell\sqrt{D}-$Lipschitz with respect to the norm $\|\cdot\|_\infty$.

The following corollary shows some example of Lipschitz functions and the corresponding concentration inequality. This example will appear more useful in the next section, making a link with the Pareto optimal set.

Corollary 2.3.1. *Let $f_1, \dots, f_D$ be $\mathbb{R} \longrightarrow \mathbb{R}$ functions such that for any k, f_k is ℓ_k-Lipschitz. Let $f : \mathbb{R}^D \ni \mathbf{X} := (X^1, \dots, X^D) \mapsto \max_d f_d(X^d)$. Let $\nu(K, D)$ a $\sigma-$sugaussian bandit. For any arm a,*

$$\mathbb{P}_\nu(f(\widehat{\boldsymbol{\mu}}_{a,s}) \geq f(\widehat{\boldsymbol{\mu}}_{a^\star,s})) \leq 4D \exp\left(-\frac{s(\Delta_a^f)^2}{8\|\boldsymbol{\ell}\|_\infty^2\sigma^2}\right),$$

where $\boldsymbol{\ell} := (\ell_1, \dots, \ell_D)$.

Proof. Let $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^D$, by reverse triangle inequality, we have

$$\begin{aligned}
|f(\mathbf{X}) - f(\mathbf{Y})| &= |\max_d f_d(X^d) - \max_d f_d(Y^d)| \\
&\leq \max_d |f_d(X^d) - f_d(Y^d)| \\
&\leq \max_d \ell_d |X^d - Y^d| \\
&\leq \|\boldsymbol{\ell}\|_\infty \|\mathbf{X} - \mathbf{Y}\|_\infty.
\end{aligned}$$

The results simply follows by Proposition 2.3. □

Before closing this section, we describe an application where Successive Rejects (f) and Sequential Halving (f) can be used with proved upper-bound. In the context of early-stage clinical trials, one wishes to determine the maximum tolerated dose of a candidate vaccine. Given a maximum tolerated toxicity θ and K candidate dose levels (or arms), one is interested in finding the arm with toxicity closest to this level (consider that both efficiency and toxicity increase with the dose level). This problem is studied in [Aziz et al., 2019] (among others). This can be formulated as a bandit problem where

the goal is to find the arm minimizing $f(\mu_a) := |\mu_a - \theta|$, that is the arm maximizing $\mu \mapsto -|\mu - \theta|$. The following remark shows that Proposition 2.3 can be used.

Remark 2.3. *Let $\theta \in \mathbb{R}$, $f_\theta : \mathbb{R} \ni \mu \mapsto -|\mu - \theta|$, f is $1-$Lipschitz, since*

$$|f_\theta(\mu) - f_\theta(\mu')| = \big||\mu' - \theta| - |\mu - \theta|\big| \leq |\mu - \mu'|.$$

Letting $\nu(K, 1)$ be a $\sigma-$subgaussian bandit, by Proposition 2.3, $f_\theta \in \Gamma(\nu, 4, 8\sigma^2)$, therefore the probability of error of Sequential Halving (f_θ) satisfies

$$\mathbb{P}_\nu(a^\star \neq \widehat{a}^\star) \leq 12 \exp\left(-\frac{T}{32\sigma^2 H_2^f \log_2(K)}\right),$$

which is up to a constant (12 instead of 9) the result proved in [Aziz et al., 2019] (Theorem 3 therein) for $\sigma^2 = 1/4$.

2.4 Pareto-increasing functions

In the previous section, we have shown that as long as a real-valued function f satisfies a given concentration inequality (Definition 2.1), some well-known algorithms in the fixed-budget setting can be adapted to the the problem of finding

$$\operatorname*{argmax}_{a \in \mathbb{A}} f(\boldsymbol{\mu}_a),$$

for a bandit with means $(\boldsymbol{\mu}_i)_{(i \in \mathbb{A})}$. While this problem may seem independent from finding the Pareto optimal set, this section shows that both can actually be related for some functions.

Definition 2.2 (Pareto-increasing function)**.** *Let $f : \mathbb{R}^p \longrightarrow \mathbb{R}^q$. f is Pareto-increasing if for $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathbb{R}^p$,*

$$\boldsymbol{\mu} \prec \boldsymbol{\mu}' \Longrightarrow f(\boldsymbol{\mu}) \prec f(\boldsymbol{\mu}').$$

In particular, when $q = 1$, $f : \mathbb{R}^p \longrightarrow \mathbb{R}$ is said Pareto-increasing if $f(\boldsymbol{\mu}) < f(\boldsymbol{\mu}')$ for $\boldsymbol{\mu} \prec \boldsymbol{\mu}' \in \mathbb{R}^p$.

The following lemma, makes the link between the Pareto-optimal set and Pareto-increasing functions.

Lemma 2.2 (Maximum of a Pareto-increasing function)**.** *Let $\mathcal{S} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$. If*

$f : \mathbb{R}^d \longrightarrow \mathbb{R}$ is a Pareto-increasing function, then

$$\operatorname*{argmax}_{\boldsymbol{\mu}\in\mathcal{S}} f(\boldsymbol{\mu}) \subset a^\star(\mathcal{S}).$$

Proof. Let $\boldsymbol{\mu} \in \mathcal{S}$ be a sub-optimal vector. By Lemma 1.2, there exists $\boldsymbol{\mu}'$ a Pareto-optimal vector such that $\boldsymbol{\mu} \preceq \boldsymbol{\mu}'$. Since f is Pareto increasing, $f(\boldsymbol{\mu}) < f(\boldsymbol{\mu}')$, which achieves the proof. □

The following proposition gives some example of Pareto-increasing functions.

Example 2.1. *Let $f_1, \dots, f_d$ be strictly increasing $\mathbb{R} \longrightarrow \mathbb{R}$ functions. $f : \mathbb{R}^d \ni \mathbf{x} = (x_1, \dots, x_d) \mapsto \sum_i f_i(x_i)$ is Pareto-increasing (follows from the definition). However, $g : \mathbb{R}^d \ni \mathbf{x} = (x_1, \dots, x_d) \mapsto \max_i f_i(x_i)$ is not Pareto-increasing but has the nice property that if*

$$\operatorname*{argmax}_{k\in[K]} g(\boldsymbol{\mu}_k)$$

is unique, then $a^\star(\nu, g)$ is a Pareto optimal arm (see [Kaisa, 1999], corollary 3.4.4 therein)

Remark 2.4. *Letting $\mathcal{S} = \{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K\}$, $\mathbf{w} \in \mathbb{R}^D$ and $\boldsymbol{z} \in \mathbb{R}^D$. The function $f_{\mathbf{w}} : \mathbf{x} \mapsto \mathbf{w}^\top \mathbf{x}$ is mostly called weighted sum or linear scalarization in the multi-objective optimization literature and $g_{\mathbf{w},\boldsymbol{z}} : \mathbf{x} \mapsto \max_d w^d(x^d - z^d)$ is called weighted Chebyschev or Chebyshev scalarization, see e.g [Drugan and Nowe, 2013, Kaisa, 1999]. When $\mathbf{w} > 0$ (component-wise), both are particular cases of Example 2.1 and, Lemma 2.2 yields*

$$\operatorname*{argmax}_{\boldsymbol{\mu}\in\mathcal{S}} f_{\mathbf{w}}(\boldsymbol{\mu}) \subset P^\star(\mathcal{S}),$$

and

$$\operatorname*{argmax}_{\boldsymbol{\mu}\in\mathcal{S}} g_{\mathbf{w},\boldsymbol{z}}(\mathbf{x}) \subset P^\star(\mathcal{S}),$$

where $P^\star(\mathcal{S})$ is the Pareto optimal set of $\mathcal{S}$. This particular result is also well-known in multi-objective optimization [Kaisa, 1999].

Letting $\nu(K, D)$ be a $\sigma-$ subgaussian bandit, Corollary 2.3.1 yields that $g_{\mathbf{w},\boldsymbol{z}} \in \Gamma(\nu, 4D, 8\|\mathbf{w}\|_\infty^2\sigma^2)$. Recall that Proposition 2.2 shows that $f_{\mathbf{w}} \in \Gamma(\nu, 4\sigma^2\|\mathbf{w}\|_2^2)$. Therefore, one can use Theorem 2.2 and Theorem 2.1 to upper-bound the probability of error of Successive Rejects (f) and Sequential Halving (f) for both $g_{\mathbf{w},\boldsymbol{z}}$ and $f_{\mathbf{w}}$.

The latter result shows how the problem of finding the Pareto optimal set can indeed

be related to the problem of finding the arm which maximizes a given real-valued function. However, as described earlier, both Successive Rejects (f) and Sequential Halving (f) only recommend one arm. That is, using these algorithms with a Pareto-increasing function, we cannot find the entire Pareto optimal set at a time (since most of the time there is more than one Pareto optimal arm). An idea could be to run parallel instances of Successive Rejects (f) or Sequential Halving (f) with different Pareto-increasing functions. But is not clear how theses functions should be chosen, or how many they should be (see [Drugan and Nowe, 2013], end of section 5.A for a discussion on a related topic in *regret minimization*). More importantly, it worth noting that some Pareto optimal vectors are "unreachable" for some Pareto-increasing functions. A well known example is that when we consider functions $f_{\mathbf{w}} : \mathbf{x} \mapsto \mathbf{w}^\top \mathbf{x}$, there exists some set $\mathcal{S}$ for which there is a Pareto optimal vector $\boldsymbol{\mu}$ such that

$$\forall\, \mathbf{w} > 0, f_{\mathbf{w}}(\boldsymbol{\mu}) < \max_{\mathbf{x} \in \mathcal{S}} f_{\mathbf{w}}(\mathbf{x}),$$

that is the Pareto optimal vector $\boldsymbol{\mu}$ can not be the maximizer of any function $f_{\mathbf{w}}$, $\mathbf{w} > 0$ (component-wise). While this result is well-known (see [Drugan and Nowe, 2013]), we did not found ts proof. So, Proposition 2.4 provides a geometric proof for a similar scenario.

Proposition 2.4 (Unreachable optimal vectors)**.** *Let $K \geq 3$, $\mathcal{S} := \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\} \subset \mathbb{R}^D$. Let $\mathbf{Y}_1, \mathbf{Y}_2 \in \mathcal{S}$ be two Pareto optimal vectors. Any Pareto optimal vector $\mathbf{X} \in]\mathbf{Y}_1, \mathbf{Y}_2[$, the interior convex hull of $\mathbf{Y}_1, \mathbf{Y}_2$ satisfies*

$$\forall\, \mathbf{w} > 0, \mathbf{w}^\top \mathbf{X} < \max_{\mathbf{x} \in \mathcal{S}} \mathbf{w}^\top \mathbf{x}.$$

Proof. Let $\vec{\mathbf{w}} > 0$. Let $\vec{u} := \frac{\vec{\mathbf{w}}}{\|\vec{\mathbf{w}}\|}$ and let $D_{\vec{\mathbf{w}}}$ be the line directed by $\vec{\mathbf{w}}$ equiped with $(O, \vec{u})$. Let I be the segment $[\mathbf{Y}_1, \mathbf{Y}_2]$. The orthogonal projection of I on $D_{\vec{\mathbf{w}}}$ is convex,

$$\vec{u} \cdot \sum_i \alpha_i \mathbf{Y}_i = \sum_i \alpha_i \vec{u} \cdot \mathbf{Y}_i = \sum_i \alpha_i \overline{OH_i},$$

where H_i is the orthogonal projection of $\mathbf{Y}_i$ on $D_{\vec{\mathbf{w}}}$. Since $\mathbf{X} \in]\mathbf{Y}_1, \mathbf{Y}_2[$, there exists $\alpha \in (0, 1)$ such that $\mathbf{X} = \alpha \mathbf{Y}_1 + (1 - \alpha)\mathbf{Y}_2$. Then, letting H be the projection of $\mathbf{X}$ on $D_{\vec{\mathbf{w}}}$,

$$\begin{aligned} \vec{\mathbf{w}} \cdot \mathbf{X} &= \|\vec{\mathbf{w}}\| \vec{u} \cdot \mathbf{X} = \alpha \|\vec{\mathbf{w}}\| \vec{u} \cdot \mathbf{Y}_1 + (1 - \alpha) \|\vec{\mathbf{w}}\| \vec{u} \cdot \mathbf{Y}_2, \\ &= \|\vec{\mathbf{w}}\| \overline{OH} = \alpha \|\vec{\mathbf{w}}\| \overline{OH}_1 + (1 - \alpha) \|\vec{\mathbf{w}}\| \overline{OH}_2. \end{aligned}$$

Hence $\overline{OH} = \alpha\overline{OH_1} + (1-\alpha)\overline{OH_2}$, which yields

$$\overrightarrow{OH} = \alpha\overrightarrow{OH_1} + (1-\alpha)\overrightarrow{OH_2},$$

where $\overrightarrow{OH} = \overline{OH}\vec{u}$. That is,

$$\overrightarrow{OH} = \overrightarrow{OH_2} + \overrightarrow{H_2H} = \alpha\overrightarrow{OH_1} + (1-\alpha)\overrightarrow{OH_2}.$$

By which precedes, $\overrightarrow{H_2H} = \alpha\overrightarrow{H_2H_1}$, with $\alpha \in (0,1)$. Therefore, whether $\mathbf{w}^\top\mathbf{Y}_1 > \mathbf{w}^\top\mathbf{X}$ or $\mathbf{w}^\top\mathbf{Y}_2 > \mathbf{w}^\top\mathbf{X}$. The following paragraph shows an example where a Pareto optimal vector is never reached by *linear scalarization.* □

Example 2.2. *In $\mathbb{R}^2$, such examples of Pareto optimal points can be simply generated by taking points on a decreasing line e.g with $\mathbf{X}_1 = (1,0)^\top$, $\mathbf{X}_2 = (\frac{1}{2},\frac{1}{2})^\top$, $\mathbf{X}_3 = (0,1)^\top$ and $\mathcal{S} := \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}$. $\mathbf{X}_2$ could not maximizes any $f_\mathbf{w} : \mathbf{x} \mapsto \mathbf{w}^\top\mathbf{x}$, $\mathbf{w} > 0$. Similar sets can be simply generated in higher dimensions.*

This motivates the need to explore different strategies in order to find all the Pareto optimal set at a time.

3

Pareto Optimal Set Identification

In the previous chapter, we have analyzed two algorithms which when appropriately tuned can be used to find one of the Pareto optimal arms. The way to choose the functions so as to identify all the Pareto optimal set being not obvious. In this chapter, we will suggest and analyze two other algorithms, which instead of finding one optimal vector at a time are expected to identify the entire Pareto optimal set. As we did in the previous chapter, these algorithms will be compared to the uniform allocation strategy.

3.1 Problem setup

Given a bandit $\nu(K, D)$, letting $\mathbb{A}$ denotes its set of arms, and $(\boldsymbol{\mu}_i)_{i\in\mathbb{A}}$ the arms' means, we are interested in identifying the Pareto optimal set of arms, that is the arms whose means belong the Pareto front of the set

$$\mathcal{S} := \{\boldsymbol{\mu}_i : i \in \mathbb{A}\}.$$

In particular, we are interested in the problem described by (1.3), that is given $T \in \mathbb{N}$, identify with high probability the Pareto optimal set of the bandit ν after T sequential interactions (as described in chapter 1 and section 2.1). In the sequel, we say that an arm i is optimal if its mean $\boldsymbol{\mu}_i$ is a Pareto optimal vector. In particular, the notation

introduced in section 2.1 still holds.

3.2 Sub-optimality metrics

We introduce some functions that would be used extensively in the sequel. The notions we will introduce are taken from [Auer et al., 2016] which introduce them for a *fixed-confidence* algorithm, having also the goal to identify the Pareto optimal set of a bandit.

Since optimality is defined regarding $a^\star(\nu)$, it makes sense to define the sub-optimality of any arm i by a kind of "distance" to $a^\star(\nu)$. Let $m(i,j)$ denote the minimum non-negative quantity s such that $\boldsymbol{\mu}_i + s$ (added component-wise uniformly) will not be strictly dominated by $\boldsymbol{\mu}_j$ (i.e $\boldsymbol{\mu}_i + s \not\prec \boldsymbol{\mu}_j$). That is

$$\begin{aligned} m(i,j) &= \min\{s \geq 0 : \boldsymbol{\mu}_i + s \not\prec \boldsymbol{\mu}_j\} \\ &= \min\{s \geq 0 : \exists\, 1 \leq d \leq D, \mu_i^d + s \geq \mu_j^d\} \\ &= \max(0, \min_{1\leq d\leq D} (\mu_j^d - \mu_i^d)) \\ &= \min_{d\in[D]} \max(0, \mu_j^d - \mu_i^d) \\ &= \min_{d\in[D]} (\mu_j^d - \mu_i^d)^+. \end{aligned} \tag{3.1}$$

By definition of $m(i,j)$, it holds that for any sub-optimal arm i, if we define

$$\alpha_i := \max_{j\in a^\star(\nu)} m(i,j),$$

then, for any $\sigma > 1$, no vector in $a^\star(\nu)$ will Pareto dominate $\boldsymbol{\mu}_i + \sigma\alpha_i$, making α_i a key quantity; which motivates the following definition.

Definition 3.1 (Sub-optimality gap of non Pareto optimal arms)**.** *For any non Pareto optimal arm i (i.e $i \notin a^\star(\nu)$),the sub-optimality gap of arm i is*

$$\Delta_i^{sub} := \max_{j\in a^\star(\nu)} m(i,j).$$

Remark 3.1. *For any bandit $\nu(K,1)$, Definition 3.1 matches the classical definition of sub-optimality gap for bandits in best arm identification (see e.g [Kaufmann et al., 2014, Audibert and Bubeck, 2010]).*

The following simple lemma is of particular interest as it will be used to provide empirical estimates of Δ_i^{sub}.

Lemma 3.1. *For any $a^\star(\nu) \subset \mathcal{S} \subset \mathbb{A}$, it holds that*

$$\Delta_i^{sub} = \max_{j \in \mathcal{S}} m(i,j).$$

Proof. The proof is as follows. Assume that i is a non Pareto optimal arm, otherwise the result holds trivially. For any non Pareto optimal arm j, there exists a Pareto optimal arm $j^\star$ such that $j \prec j^\star$ (see Lemma 1.2). If $m(i,j) > m(i,j^\star)$, then

$$m(i,j^\star) + \boldsymbol{\mu}_i \prec \boldsymbol{\mu}_j \prec \boldsymbol{\mu}_{j^\star},$$

that is

$$m(i,j^\star) + \boldsymbol{\mu}_i < \boldsymbol{\mu}_{j^\star},$$

which is impossible by definition of $m(i,j^\star)$, hence $m(i,j) \leq m(i,j^\star)$ and

$$\max_{j \in \mathcal{S}} m(i,j) = \max_{j \in a^\star(\nu)} m(i,j).$$

□

It worth noting that for any Pareto optimal arm i,

$$\Delta_i^{\text{sub}} = 0,$$

which is even an equivalence (by applying the definition).

While the definition of the sub-optimality gap of non optimal arms is clear (at least for us) and matches the classical definition for one dimensional bandits, defining the sub-optimality gap of Pareto optimal arms is more tedious. Even if they all share the same status of being "optimal", they're not equally difficult to identify whence the need to define appropriate gaps. [Auer et al., 2016] suggest a gap in the fixed-confidence setting. We will not use this gap in this report but we introduce another function that will be of interest in the sequel.

For any arm i, j, let $M(i,j)$ denote the minimum non-negative quantity s such that $\boldsymbol{\mu}_j + s$ (added component-wise uniformly) weakly dominates $\boldsymbol{\mu}_i$ (i.e $\boldsymbol{\mu}_i \preceq \boldsymbol{\mu}_j + s$). That

is,

$$\begin{aligned} M(i,j) &= \min\{s \geq 0 : \boldsymbol{\mu}_i \preceq \boldsymbol{\mu}_j + s\} \\ &= \min\{s \geq 0 : \forall\, 1 \leq d \leq D, \mu_i^d \leq \mu_j^d + s\} \\ &= \max(0, \max_{1 \leq d \leq D} (\mu_i^d - \mu_j^d)) \\ &= \max_{1 \leq d \leq D} \max(\mu_i^d - \mu_j^d, 0) \\ &= \|(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^+\|_\infty. \end{aligned}$$

This section ends with the following lemma, which is of important interest as it bounds the deviation of the plugin estimate of $m(i,j)$ from its true value for any arms i, j. While this result was already given in [Auer et al., 2016] (in a slightly different form), the proof is a personal work since the authors didn't describe the proof.

Lemma 3.2 (Concentration of the plugin estimate)**.** *For any time t, letting $\boldsymbol{\mu}_i(t)$ denotes the empirical mean of arm i at time t, and any d, and for any arms i, j, it holds that*

$$|\widehat{m}(i,j,t) - m(i,j)| \leq \|\widehat{\boldsymbol{\mu}}_i(t) - \boldsymbol{\mu}_i\|_\infty + \|\widehat{\boldsymbol{\mu}}_j(t) - \boldsymbol{\mu}_j\|_\infty,$$

where $\widehat{m}(i,j,t)$ is the evaluation of (3.1) using the empirical means $\widehat{\boldsymbol{\mu}}_i(t)$ and $\widehat{\boldsymbol{\mu}}_j(t)$.

Proof. From the definition, there exists $\widehat{d}_t$ and d such that

$$\widehat{m}(i,j,t) = \max(0, \widehat{\mu}_j^{\widehat{d}_t}(t) - \widehat{\mu}_i^{\widehat{d}_t}(t)),$$

and

$$m(i,j) = \max(0, \mu_j^d - \mu_i^d),$$

which (by reverse triangle inequality) yields

$$|\widehat{m}(i,j,t) - m(i,j)| \leq |\widehat{\mu}_j^{\widehat{d}_t}(t) - \widehat{\mu}_i^{\widehat{d}_t}(t) - (\mu_j^d - \mu_i^d)|.$$

By definition of d and $\widehat{d}_t$, it follows that

$$\widehat{\mu}_j^{\widehat{d}_t}(t) - \widehat{\mu}_i^{\widehat{d}_t}(t) \leq \widehat{\mu}_j^d(t) - \widehat{\mu}_i^d(t),$$

therefore,

$$\widehat{\mu}_j^{\widehat{d}_t}(t) - \widehat{\mu}_i^{\widehat{d}_t}(t) - (\mu_j^d - \mu_i^d) \leq \widehat{\mu}_j^d(t) - \widehat{\mu}_i^d(t) - (\mu_j^d - \mu_i^d). \tag{3.2}$$

On the other side, it holds that

$$\mu_j^d - \mu_i^d \leq \mu_j^{\widehat{d}_t} - \mu_i^{\widehat{d}_t},$$

which yields

$$\widehat{\mu}_j^{\widehat{d}_t}(t) - \widehat{\mu}_i^{\widehat{d}_t}(t) - (\mu_j^d - \mu_i^d)$$

$$\widehat{\mu}_j^{\widehat{d}_t}(t) - \mu_i^{\widehat{d}_t}(t) - (\mu_j^d - \mu_i^d) \geq \widehat{\mu}_j^{\widehat{d}_t}(t) - \widehat{\mu}_i^{\widehat{d}_t}(t) - (\mu_j^{\widehat{d}_t} - \mu_i^{\widehat{d}_t}), \tag{3.3}$$

then, (3.2) and (3.3) yields

$$\widehat{\mu}_j^{\widehat{d}_t}(t) - \widehat{\mu}_i^{\widehat{d}_t}(t) - (\mu_j^{\widehat{d}_t} - \mu_i^{\widehat{d}_t}) \leq \widehat{\mu}_j^{\widehat{d}_t}(t) - \widehat{\mu}_i^{\widehat{d}_t}(t) - (\mu_j^d - \mu_i^d) \leq \widehat{\mu}_j^d(t) - \widehat{\mu}_i^d(t) - (\mu_j^d - \mu_i^d),$$

that is

$$\begin{aligned} |\widehat{m}(i,j,t) - m(i,j)| &\leq |\widehat{\mu}_j^d(t) - \widehat{\mu}_i^d(t) - (\mu_j^d - \mu_i^d)| \\ &\leq |\widehat{\mu}_j^d(t) - \mu_j^d| + |\widehat{\mu}_i^d(t) - \mu_i^d| \\ &\leq \|\widehat{\boldsymbol{\mu}}_i(t) - \boldsymbol{\mu}_i\|_\infty + \|\widehat{\boldsymbol{\mu}}_j(t) - \boldsymbol{\mu}_j\|_\infty, \end{aligned}$$

which achieves the proof. □

3.3 UCB-E

This algorithm is inspired from [Katz-Samuels and Scott, 2018] which itself can be traced back to the seminal paper of [Audibert and Bubeck, 2010]. The idea is to use an approximation of the complexity term H of the problem to tune the level of exploration of an algorithm (see e.g [Gabillon et al., 2012, Audibert and Bubeck, 2010]). While this seems not realistic in practice, algorithm 3 can at least be evaluated on simulated instances. Before deriving the algorithm and its proof, we need to introduce a new gap and the associated complexity term. Let $\nu(K, D)$ a bandit, $\varepsilon > 0$, $a \in \mathbb{A}$ and $\Delta_{a,\varepsilon}$ the novel sub-optimality gap defined as

$$\Delta_{a,\varepsilon} := |\Delta_a^{\text{sub}} - \varepsilon| = |\max_{i \in \mathbb{A}} m(a,i) - \varepsilon|,$$

which is associated to the complexity term

$$H_\varepsilon := \sum_a \Delta_{a,\varepsilon}^{-2}.$$

Let $\widehat{\Delta}_{a,\varepsilon}$ denotes the empirical estimate of $\Delta_{a,\varepsilon}$, that is

$$\widehat{\Delta}_{a,\varepsilon}(t) := |\widehat{\Delta}_a^{\text{sub}}(t) - \varepsilon| = |\max_{i \in \mathbb{A}} \widehat{m}(a,i,t) - \varepsilon|.$$

Assuming that for any arm a, $\Delta_{a,\varepsilon} > 0$, let

$$h := \max_{i,j \in \mathbb{A}} \frac{\Delta_{i,\varepsilon}}{\Delta_{j,\varepsilon}}.$$

The following lemma upper-bounds the probability of error of algorithm 3.

Algorithm 3: UCB-E.

Input : Bandit $\nu(K, D)$, parameter ε, c
pull each arm once ;
for $t = K + 1$ **to** T **do**
  pull arm $a_t := \underset{a \in \mathbb{A}}{\text{argmin}}\, \widehat{\Delta}_{a,\varepsilon}(t) - \sqrt{c/T_a(t)}$;
  observe reward $X_{a_t,t} \sim \nu_{a_t}$;
Output : $\widehat{a}_\varepsilon^\star(\nu, T) := \left\{k : \widehat{\Delta}_k^{\text{sub}}(T) \leq \varepsilon\right\}$

Theorem 3.1 (Bounding the probability of error of algorithm 3)**.** *Let $\varepsilon > 0$, budget $T \geq K$. Suppose $0 \leq c \leq \frac{49}{64}\frac{T}{H_\varepsilon}$. algorithm 3 has a probability of error upper-bounded as*

$$\mathbb{P}_\nu(a_\epsilon^\star(\nu) \neq \widehat{a}_\epsilon^\star(\nu)) \leq 2KD(1 + \log T) \exp\left(-\frac{c}{4\sigma^2(2h(1+h)+3)^2}\right).$$

Proof. The idea is to define a favorable event and show that on this event, given that the arms are sampled enough, the algorithm makes no error. The proof borrows from [Katz-Samuels and Scott, 2018] which extends the proof of APT in [Locatelli et al., 2016].

Step 0: Setting and notations. Let $\varepsilon > 0$, $\nu(K, D)$ a bandit instance and $T \geq K$. For any arm $a \in \mathbb{A}$, let $\beta_a(t) := \sqrt{\frac{c\theta^2}{T_a(t)}}$ and let

$$I_a(t) := \widehat{\Delta}_{a,\varepsilon}(t) - \sqrt{c/T_a(t)}.$$

Step 1: A good event. Let θ be a parameter to be set latter. Let us define the event

$$\Xi_\theta := \left\{\forall a \in \mathbb{A}, \forall d \in [D], \forall 1 \leq s \leq T, |\widehat{\mu}_{a,s}^d - \mu_a^d| \leq \sqrt{\frac{c\theta^2}{s}}\right\}.$$

The probability of the complementary event can be upper-bounded as

$$\begin{aligned}\mathbb{P}_\nu(\bar{\Xi}_\theta) &\leq \sum_{a=1}^{K}\sum_{d=1}^{D}\mathbb{P}_\nu\left(\exists\, s\leq T : |\widehat{\mu}_{a,s}^d-\mu_a^d|>\sqrt{\frac{c\theta^2}{s}}\right),\\ &\leq \sum_{a=1}^{K}\sum_{d=1}^{D}\sum_{u=0}^{\lceil\log T\rceil-1}\mathbb{P}_\nu\left(\exists\, s\in[2^u,2^{u+1}] : s|\widehat{\mu}_{a,s}^d-\mu_a^d|>\sqrt{s\theta^2 c}\right),\\ &\leq \sum_{a=1}^{K}\sum_{d=1}^{D}\sum_{u=0}^{\lceil\log T\rceil-1}\mathbb{P}_\nu\left(\exists\, s\leq 2^{u+1} : s|\widehat{\mu}_{a,s}^d-\mu_a^d|\geq\sqrt{2^u\theta^2 c}\right).\end{aligned}$$

Then, using the subgaussian martingale inequality(Lemma 1.7), it follows

$$\begin{aligned}\mathbb{P}_\nu(\bar{\Xi}_\theta) &\leq \sum_{a=1}^{K}\sum_{d=1}^{D}\sum_{u=0}^{\lceil\log T\rceil-1}2\exp\left(-\frac{c\theta^2}{4\sigma^2}\right),\\ &\leq 2KD\lceil\log T\rceil\exp\left(-\frac{c\theta^2}{4\sigma^2}\right),\\ &< 2KD(1+\log T)\exp\left(-\frac{c\theta^2}{4\sigma^2}\right).\end{aligned}$$

By what precedes, the probability of Ξ_θ can be lower-bounded as

$$\mathbb{P}_\nu(\Xi_\theta)\geq 1-2KD(1+\log T)\exp\left(-\frac{c\theta^2}{4\sigma^2}\right). \tag{3.4}$$

Using Lemma 3.1 and Lemma 3.2 it comes (by reverse triangle inequality) that on the event Ξ_θ, for any arm a,

$$|\widehat{\Delta}_{a,\varepsilon}(t)-\Delta_{a,\varepsilon}|\leq\beta_a(t)+\max_{k\in\mathbb{A}}\beta_k(t). \tag{3.5}$$

Step 2: Lower bound on the number of pulls of any arm. By the pigeonhole principle, there exists an arm $k\in\mathbb{A}$ such that $T_k(T)\geq\frac{T}{H_\varepsilon\Delta_{k,\varepsilon}^2}$. Let t be the last time arm k has been pulled by algorithm 3. Let a be one of the least played arm up to time t, Since k has been pulled at time t, for any arm j, we have $I_k(t)\leq I_j(t)$, then,

$$\widehat{\Delta}_{k,\varepsilon}(t)-\sqrt{\frac{c}{T_k(t)}}\leq\widehat{\Delta}_{a,\varepsilon}(t)-\sqrt{\frac{c}{T_a(t)}},$$

then, using (3.5) yields

$$\Delta_{k,\varepsilon} - \beta_k(t) - \sqrt{\frac{c}{T_k(t)}} \leq \Delta_{a,\varepsilon} + 3\beta_a(t) - \sqrt{\frac{c}{T_a(t)}},$$

which rewrites as

$$\Delta_{k,\varepsilon} - (\theta+1)\sqrt{\frac{c}{T_k(t)}} \leq \Delta_{a,\varepsilon} + (3\theta-1)\sqrt{\frac{c}{T_a(t)}},$$

using $T_k(t) \geq T/H_\varepsilon \Delta_{k,\varepsilon}^2$ and $0 \leq c \leq \delta^2 T/H_\varepsilon$ yields

$$\Delta_{k,\varepsilon} - (\theta+1)\Delta_{k,\varepsilon}\delta \leq \Delta_{a,\varepsilon} + (3\theta-1)\sqrt{\frac{c}{T_a(t)}}.$$

Setting $\delta \leq 1/(\theta+1)$ (discussed later) yields

$$0 \leq \Delta_{a,\varepsilon} + (3\theta-1)\sqrt{\frac{c}{T_a(t)}},$$

then,

$$T_a(t) \geq \frac{cq(\theta)^2}{\Delta_{a,\varepsilon}^2}, \tag{3.6}$$

where $q(\theta) := 1-3\theta$ and we assume θ is chosen such that $q(\theta)$ is non-negative. Therefore, for any $j \in \mathbb{A}$,

$$T_j(T) \geq T_j(t) \geq \frac{cq(\theta)^2}{\Delta_{j,\varepsilon}^2 h^2}. \tag{3.7}$$

Step 3: On the event Ξ_θ, for a well-chosen θ, algorithm 3 makes no error. For any arm $k \in \mathbb{A}$, we have (Lemma 3.2)

$$\begin{aligned}
|\widehat{\Delta}_k^{\text{sub}}(T) - \Delta_k^{\text{sub}}| &\leq \beta_k(T) + \max_{j\in\mathbb{A}} \beta_j(T) \\
&\leq q(\theta)^{-1}\theta h(\Delta_{k,\varepsilon} + \max_{j\in K} \Delta_{j,\varepsilon}) \qquad \text{using (3.7)} \\
&\leq q(\theta)^{-1}\theta(1+h)h\Delta_{k,\varepsilon}
\end{aligned}$$

Assume $\tilde{\theta}$ is chosen such that $q(\tilde{\theta})^{-1}\tilde{\theta}(1+h)h = \frac{1}{2}$ (discussed later), then, $|\widehat{\Delta}_k^{\text{sub}}(T) - \Delta_k^{\text{sub}}| \leq \frac{1}{2}\Delta_{k,\varepsilon}$. Let an arm $k \in \mathbb{A}$. If $\Delta_k^{\text{sub}} \leq \varepsilon$. By what precedes,

$$\widehat{\Delta}_k^{\text{sub}}(T) - \Delta_k^{\text{sub}} \leq \frac{1}{2}\Delta_{k,\varepsilon} = \frac{1}{2}(\varepsilon - \Delta_k^{\text{sub}}),$$

then,

$$\begin{aligned}\widehat{\Delta}_k^{\text{sub}}(T) - \varepsilon &\leq \frac{1}{2}(\varepsilon - \Delta_k^{\text{sub}}) - \varepsilon + \Delta_k^{\text{sub}}, \\ &= \frac{1}{2}(\Delta_k^{\text{sub}} - \varepsilon) \leq 0.\end{aligned}$$

On the other side, if $\Delta_k^{\text{sub}} \geq \varepsilon$,

$$\widehat{\Delta}_k^{\text{sub}}(T) - \varepsilon \geq \frac{1}{2}(\varepsilon - \Delta_k^{\text{sub}}) + \Delta_k^{\text{sub}} - \varepsilon = \frac{1}{2}(\Delta_k^{\text{sub}} - \varepsilon) \geq 0,$$

that is, the algorithm makes no error on $\Xi_{\tilde{\theta}}$.

Step 4: Conclusion. Let us now discuss the choice of θ. We need $\tilde{\theta}$ such that

$$2\tilde{\theta}(1+h)h = 1 - 3\tilde{\theta},$$

which yields,

$$\tilde{\theta}\,(2(1+h)h + 3) = 1$$

then,

$$\tilde{\theta}^2 := \frac{1}{(2h(1+h)+3)^2}.$$

We check that $q(\tilde{\theta}) = \frac{2h(1+h)}{2h(1+h)+3} \geq 0$ and taking $\delta = \frac{7}{8}$, we have $\delta \leq 1/(\tilde{\theta}+1) = \frac{2h(1+h)+3}{2h(1+h)+4}$ for $h \geq 1$. Finally, the probability of error is upper-bounded by the probability of $\tilde{\Xi}_{\tilde{\theta}}$ which by (3.4) achieves the proof. □

The are many limitations to this algorithm. The first being to tune c appropriately. Another point is that h is problem-dependent and can be considerably large. The last point we would like to mention is that when $\varepsilon \approx 0$ (so $c \approx 0$ for T small enough), the complexity term H_ε explodes and algorithm 3 very likely reduces to sample for each round the least sampled arm of the empirical Pareto optimal set. Which will probably result in empirical sub-optimal arms not explored enough, making this strategy inefficient ([Audibert et al., 2009]) in this scenario.

3.4 Successive Rejects (q)

Based on the elimination strategy of Successive Rejects and Successive Rejects (f) we suggest a new algorithm called Successive Rejects (q) , q being an integer. Recall that given a $K-$armed bandit, Successive Rejects proceeds by $K-1$ elimination rounds,

where at the end of each round, the arm with the lowest empirical mean is eliminated, the ultimately surviving arm being recommended by the algorithm as the best arm. Successive Rejects (f) builds upon this idea by using a function real-value f and by eliminating at the end of each round, the arm whose empirical mean minimizes f. This allows to use bandits with multivariate arms, and, under some assumptions on f, we give an upper-bound on the probability of error of this algorithm. The idea behind Successive Rejects (q) is quite similar. Suppose we are given the size of the Pareto optimal set, q. Let $\nu(K, D)$ be a bandit. Having at hand a function f such that

$$\forall\, k \in \mathbb{A} \setminus a^\star(\nu), f(\boldsymbol{\mu}_k) < \min_{a \in a^\star(\nu)} f(\boldsymbol{\mu}_a), \tag{3.8}$$

one could run Successive Rejects (f) for $K - q$ rounds and expect the still alive q arms to be the Pareto optimal arms. On the other hand, recall that

$$\forall k \in \mathbb{A} \setminus \alpha^\star(\nu), \Delta_k^{\text{sub}} > 0 \text{ and } \forall\, a \in a^\star(\nu), \Delta_a^{\text{sub}} = 0,$$

making $k :\mapsto -\Delta_k^{\text{sub}}$ a function satisfying (3.8). algorithm 4 details the strategy of Successive Rejects (q) with some new constants so as to use all (at least the maximum) the budget in $K - q$ rounds. At step k, letting A_k be the set of still active arms, for any arm a, we will estimate Δ_k^{sub} by the empirical quantity

$$\widehat{\Delta}_{a,n_k} := \max_{i \in A_k} \widehat{m}(a, i, k),$$

where $\widehat{m}(a, i, k)$ is the plugin estimate of $m(a, i)$ using the empirical means $\widehat{\boldsymbol{\mu}}_{a,n_k}$ and $\widehat{\boldsymbol{\mu}}_{i,n_k}$. We define the sub-optimality gap of any sub-optimal arm a as

$$\Delta_a := \Delta_a^{\text{sub}},$$

and all the Pareto optimal arms are assigned the same gap

$$\min_{a \in \mathbb{A} \setminus a^\star(\nu)} \Delta_a.$$

Assuming that the arms are ordered as $\Delta_1 = \cdots = \Delta_{q+1} \leq \cdots \leq \Delta_K$, the complexity associated to this algorithm will be measure by

$$H_2^q := \max_{k \in \mathbb{A}} k \Delta_k^{-2},$$

and

$$H^q := \sum_{k \in \mathbb{A}} \Delta_k^{-2}.$$

Using equation 1. from [Audibert and Bubeck, 2010] yields

$$H_2^q \leq H^q \leq \log(2K) H_2^q. \tag{3.9}$$

Before providing an upper-bound on the probability of error of Successive Rejects (q) ,

Algorithm 4: Successive Rejects(q).

Input : Bandit $\nu(K, D)$, budget $T \geq K$, $q := |a^\star(\nu)| < K$
Initialize: $A_1 := \{1, \ldots, K\}, n_0 = 0$
Define : $\overline{\log^q}(K) := \frac{q}{q+1} + \sum_{i=q+1}^{K} i^{-1}, n_k := \left\lceil \frac{1}{\overline{\log^q}(K)} \frac{T-K}{K+1-k} \right\rceil$
for $k \leftarrow 1$ **to** $K - q$ **do**
  foreach $a \in A_k$ **do**
    pull arm a for $n_k - n_{k-1}$ rounds;
  $A_{k+1} \leftarrow A_k \setminus \operatorname{argmax}_{a \in A_k} \widehat{\Delta}_{a,n_k}$;
$\widehat{a}^\star(\nu, T) \leftarrow A_{K-q+1}$;
Output : $\widehat{a}^\star(\nu, T)$

Lemma 3.3 shows that the algorithm does not exceed the budget.

Lemma 3.3 (Successive Rejects (q) does not exceed the budget)**.** *Let $\nu(K, D)$ a bandit with $q = |a^\star(\nu)|$. Let a budget $T \geq K$. The number of pulls of algorithm 4 does not exceed T.*

Proof. The number of pulls of algorithm 4 is

$$N_T := \sum_{k=1}^{K-q} (n_k - n_{k-1})(K - k + 1),$$

which by telescoping rewrites as

$$\begin{aligned} N_T &= (K+1)n_{K-q} - \sum_{k=1}^{K-q}(kn_k - (k-1)n_{k-1} - n_{k-1}), \\ &= (K+1)n_{K-q} - (K-q)n_{K-q} + \sum_{k=1}^{K-q-1} n_k, \\ &= qn_{K-q} + \sum_{k=1}^{K-q} n_k. \end{aligned}$$

Then, using the expression of n_k yields,

$$\begin{aligned} N_T &< q\left(\frac{T-K}{\overline{\log^q}(K)}\frac{1}{q+1}+1\right) + (K-q) + \sum_{k=1}^{K-q}\frac{T-K}{\overline{\log^q}(K)}\frac{1}{K+1-k}, \\ &< \frac{T-K}{\overline{\log^q}(K)}\frac{q}{q+1} + K + \frac{T-K}{\overline{\log^q}(K)}\sum_{k=q+1}^{K} k^{-1}, \\ &= \frac{T-K}{\overline{\log^q}(K)}\underbrace{\left(q/(q+1) + \sum_{k=q+1}^{K} k^{-1}\right)}_{\overline{\log^q}(K)} + K, \\ &= T. \end{aligned}$$

□

Therefore, algorithm 4 is well-defined and the following lemma upper-bounds its probability of error.

Theorem 3.2 (Upper-bounding the probabilitiy of error of algorithm 4)**.** *Let $\nu(K, D)$ be a $\sigma-$subgaussian bandit. Let $q := |a^\star(\nu)| < K$, and $T \geq K$. The probability of error of Successive Rejects (q) satisfies*

$$\mathbb{P}_\nu(a^\star \neq \widehat{a}^\star) \leq 2KDq(K-q)\frac{K-q+1}{2}\exp\left(-\frac{T-K}{32\sigma^2 H_2^q \overline{\log^q}(K)}\right).$$

Proof. The proof borrows from the proof given in [Audibert and Bubeck, 2010].
Step 0: Setup and notations. Let $q \geq 1$, $\overline{\log^q}(K) := q/(q+1) + \sum_{i=q+1}^{K} i^{-1}$. Without loss of generality, assume $a^\star(\nu) := \{1, \ldots, q\}$ and the arms are ordered as $\Delta_1 = \Delta_2 = \ldots \Delta_q := \Delta_{q+1} \leq \cdots \leq \Delta_K$. Let $\widehat{a}^\star := \widehat{a}^\star(\nu, T)$ be the set output by

algorithm 4. Let $e_T(\nu, f) := \mathbb{P}_\nu(\widehat{a}^\star \neq a^\star)$. Let ξ_k be the event

"an optimal an is dismissed at the end of round k".

Let ζ_k be the event

$$\bigcap_{q=1}^{k-1} \bar{\xi}_q,$$

that is, no error occurs before (the beginning of) round k. Let ϑ_k be the event

"the first error occurs at the end of round k"

Moreover, assume that the quantity $\widehat{\boldsymbol{\mu}}_{a,n_k}$ is always defined even if arm a has been dismissed before round k.

Step 1: Upper-bounding the probability that an optimal arm is empirically worse than a sub-optimal arm. Recall that $\nu(K, D)$ is a $\sigma-$subgaussian bandit. Let $i \in a^\star$, an optimal arm. Let $j \in \mathbb{A}$ be a sub-optimal arm. We have

$$\begin{aligned}\widehat{\Delta}_{i,n_k} \geq \widehat{\Delta}_{j,n_k} \Longleftrightarrow & \widehat{\Delta}_{i,n_k} + (\Delta_j - \widehat{\Delta}_{j,n_k}) \geq \Delta_j, \\ \Longrightarrow & \widehat{\Delta}_{i,n_k} \geq \frac{\Delta_j}{2} \text{ or } \Delta_j - \widehat{\Delta}_{j,n_k} \geq \frac{\Delta_j}{2}. \end{aligned} \tag{3.10}$$

Let the event

$$\Pi_{j,k}(A) := \{\forall\, d \in [D], \forall\, a \in A\ , |\widehat{\mu}^d_{a,n_k} - \mu^d_a| < \frac{\Delta_j}{4}\}.$$

On the event ζ_k, no error occurs before round k, hence $a^\star(\nu) \subset A_k$. Therefore, for any arm a, by Lemma 3.1,

$$\Delta_a := \max_{a' \in a^\star(\nu)} m(a, a') = \max_{a' \in A_k} m(a, a').$$

On the event $\zeta_k \cap \Pi_{j,k}(A_k)$, we have for any arm $a \in A_k$,

$$\begin{aligned}|\widehat{\Delta}_{a,n_k} - \Delta_a| :=& |\max_{a' \in A_k} \widehat{m}(a, a', k) - \max_{a' \in A_k} m(a, a')|, \\ \leq& \max_{a' \in A_k} |\widehat{m}(a, a', k) - m(a, a')|, \\ <& \max_{a' \in A_k} \frac{\Delta_j}{4} + \frac{\Delta_j}{4}, \quad \text{by Lemma 3.2, since } \Pi_{j,k}(A_k) \text{ holds} \\ =& \frac{\Delta_j}{2}.\end{aligned}$$

The latter result implies by contraposition of (3.10) that on the event ζ_k, if $\widehat{\Delta}_{i,n_k} \geq \widehat{\Delta}_{j,n_k}$ then $\Pi_{j,k}(A_k)$ does not hold. Therefore,

$$\begin{aligned}
\mathbb{P}_\nu(\widehat{\Delta}_{i,n_k} \geq \widehat{\Delta}_{j,n_k}, \zeta_k) &\leq \mathbb{P}_\nu(\Pi_{j,k}(A_k)^{\mathsf{c}}), \\
&\leq \mathbb{P}_\nu(\Pi_{j,k}(\mathbb{A})^{\mathsf{c}}), \qquad \text{since } \Pi_{j,k}(\mathbb{A}) \subset \Pi_{j,k}(A_k) \\
&\leq \sum_{a=1}^{K}\sum_{d=1}^{D} \mathbb{P}_\nu(|\widehat{\mu}^d_{a,n_k} - \mu^d_a| \geq \frac{\Delta_j}{4}), \\
&\leq \sum_{a=1}^{K}\sum_{d=1}^{D} 2\exp(-\frac{\Delta_j^2 n_k}{32\sigma^2}), \qquad \text{by } \sigma-\text{subgaussian Hoeffdding, Lemma 1.6} \\
&= 2KD\exp(-\frac{\Delta_j^2 n_k}{32\sigma^2}). \qquad (3.11)
\end{aligned}$$

Step 2: Upper-bounding the probability of the event ϑ_k. We have

$$\vartheta_k := \zeta_k \cap \xi_k,$$

then

$$\begin{aligned}
\mathbb{P}_\nu(\vartheta_k) &= \mathbb{P}_\nu(\xi_k, \zeta_k) \\
&\leq \sum_{a \in a^\star(\nu)} \mathbb{P}_\nu(\exists\, i \in A_k : \widehat{\Delta}_{a,n_k} \geq \widehat{\Delta}_{i,n_k}, \zeta_k). \qquad (3.12)
\end{aligned}$$

In particular, if an optimal a is removed at the end of round k, then there exists an arm i, among the k worst arms (ordered by the Δ_a's, larger is worse) such that $\widehat{\Delta}_{a,n_k} \geq \widehat{\Delta}_{i,n_k}$. Since $1 \geq k \geq K - q$, this arm i is a sub-optimal arm. In the sequel of (3.12) and by

what precedes,

$$\begin{aligned}
\mathbb{P}_\nu(\vartheta_k) &\leq \sum_{a\in a^\star(\nu)} \mathbb{P}_\nu(\exists\, i \in \{K+1-k,\dots,K\} : \widehat{\Delta}_{a,n_k} \geq \widehat{\Delta}_{i,n_k}, \zeta_k),\\
&\leq \sum_{a\in a^\star(\nu)} \sum_{i=K+1-k}^{K} \mathbb{P}_\nu(\widehat{\Delta}_{a,n_k} \geq \widehat{\Delta}_{i,n_k}, \zeta_k),\\
&\leq \sum_{a\in a^\star(\nu)} \sum_{i=K+1-k}^{K} 2KD\exp(-\frac{\Delta_i^2 n_k}{32\sigma^2}), \qquad \text{by (3.11)}\\
&\leq \sum_{i=K+1-k}^{K} 2KDq\exp(-\frac{\Delta_i^2 n_k}{32\sigma^2}),\\
&\leq 2KDqk\exp(-\frac{\Delta_{K+1-k}^2 n_k}{32\sigma^2}),
\end{aligned}$$

where the last inequality follows by $\Delta_{K+1-k} \leq \cdots \leq \Delta_K$.

Step 3: Conclusion. Recall that

$$H_2^q := \max_{a\in\mathbb{A}} a\Delta_a^{-2},$$

and,

$$\begin{aligned}
n_k\Delta_{K+1-k}^2 &= \left\lceil \frac{1}{\overline{\log^q}(K)} \frac{T-K}{K+1-k} \right\rceil \Delta_{K+1-k}^2,\\
&\geq \frac{1}{\overline{\log^q}(K)} \frac{T-K}{\Delta_{K+1-k}^{-2}(K+1-k)},\\
&\geq \frac{1}{\overline{\log^q}(K)} \frac{T-K}{H_2^q}.
\end{aligned}$$

Putting things together,

$$\begin{aligned}
\mathbb{P}_\nu(\widehat{a}^\star \neq a^\star) &= \sum_{k=1}^{K-q} \mathbb{P}_\nu(\vartheta_k),\\
&\leq \sum_{k=1}^{K-q} 2KDqk\exp(-\frac{T-K}{32\sigma^2 H_2^q \overline{\log^q}(K)}),\\
&\leq 2KDq(K-q)\frac{K-q+1}{2}\exp(-\frac{T-K}{32\sigma^2 H_2^q \overline{\log^q}(K)}),
\end{aligned}$$

which achieves the proof. □

In the next section, we provide an upper-bound of the mis-identification probability of the uniform allocation strategy which will be compared to algorithm 4 and algorithm 3.

3.5 Uniform allocation

Given a bandit $\nu(K, D)$ and a budget $T \geq K$, the uniform allocation strategy randomly uniformly selects one of the K arms at each round and ultimately recommends the empirical Pareto optimal set. We define the sub-optimality gap of any sub-optimal arm a as

$$\Delta_a := \Delta_a^{\text{sub}},$$

and for any optimal arm i,

$$\Delta_i := \min_{j \neq i} M(i, j).$$

Remark that when $D = 1$, the latter gap also matches the gap of the (unique) optimal arm as introduced in [Audibert and Bubeck, 2010]. Assuming that the arms are ordered as $\Delta_1 = \cdots = \Delta_{q+1} \leq \cdots \leq \Delta_K$, the complexity associated to this algorithm will be measured by

$$H_2^u := \max_{k \in \mathbb{A}} k\Delta_k^{-2}, \qquad \text{and} \qquad H^u := \sum_{k \in \mathbb{A}} \Delta_k^{-2}.$$

Similarly to (3.9), it holds that

$$H_2^u \leq H^u \leq \log(2K) H_2^u. \tag{3.13}$$

algorithm 5 describes the uniform allocation strategy and Theorem 3.3 upper-bounds its probability of error.

Algorithm 5: Uniform Allocation.

Input : Bandit $\nu(K, D)$,budget T
pull each arm once ;
for $t = K + 1$ **to** T **do**
  sample $l_t \sim \mathcal{U}(\mathbb{A})$;
  pull arm ℓ_t;
  observe reward $X_{\ell_t, t} \sim \nu_{\ell_t}$;
Output : $\widehat{a}^\star(\nu, T)$, the empirical Pareto set

Theorem 3.3 (Bounding the probability of error of algorithm 5)**.** *algorithm 5 run with a budget $T \geq K$ on the bandit instance $\nu(K, D)$ has a probability of error upper-bounded*

as

$$\mathbb{P}_\nu(a^\star \neq \widehat{a}^\star) \leq K(K+D)\exp\left(-\frac{T}{4\sigma^2 K\log(2K)H_2^u}\right).$$

Proof. The proof is split into two parts. We upper-bound the probability that an optimal arm is empirically sub-optimal and vice-versa. Without loss of generality, we assume that each arm is pulled T/K times.

Step 1: Upper-bounding the probability that a sub-optimal arm is empirically optimal. Let a be a sub-optimal arm. There exist $k \in a^\star(\nu)$ such that $a \preceq k$ (Lemma 1.2) and $\Delta_a = m(a,k)$. If $a \in \widehat{a}^\star(\nu,T)$, then $\widehat{\boldsymbol{\mu}}_a(T) \not\preceq \widehat{\boldsymbol{\mu}}_k(T)$, that is, there exist $d \in [D]$ such that $\widehat{\mu}_a^d(T) > \widehat{\mu}_k^d(T)$. Therefore,

$$\begin{aligned}
\mathbb{P}_\nu(a \in a^\star(\nu)^{\mathsf{c}} \cap \widehat{a}^\star(\nu,T)) &\leq \mathbb{P}_\nu(\widehat{\mu}_a(T) \not\preceq \widehat{\mu}_k(T)),\\
&\leq \mathbb{P}_\nu(\exists\, d \in [D] \text{ s.t } \widehat{\mu}_a^d(T) > \widehat{\mu}_k^d(T)),\\
&\leq \mathbb{P}_\nu(\exists\, d \in [D] \text{ s.t } \widehat{\mu}_a^d(T) - \widehat{\mu}_k^d(T) - (\mu_a^d(T) - \mu_k^d(T)) > (\mu_k^d(T) - \mu_a^d(T)) \geq \Delta_a),\\
&\leq D\exp\left(-\frac{T\Delta_a^2}{4\sigma^2 K}\right). \quad \text{(union bound and } \sigma\text{− subgaussian Hoeffding)}
\end{aligned}$$

Step 2: Upper-bounding the probability that an optimal arm is empirically sub-optimal. Let $a \in a^\star(\nu) \cap \widehat{a}^\star(\nu,T)^{\mathsf{c}}$. Let for any $k \in \mathbb{A}$, $d_{a,k} := \text{argmax}_{d\in[D]}(\mu_a^d - \mu_k^d)$. Since $a \in a^\star(\nu)$, for any k, $\mu_a^{d_{a,k}} - \mu_k^{d_{a,k}} = M(a,k) > 0$. Since $a \notin \widehat{a}^{\star\mathsf{c}}$, there exists $k \in \mathbb{A}$ such that $\widehat{\boldsymbol{\mu}}_a(T) \preceq \widehat{\boldsymbol{\mu}}_k(T)$ and, $M(a,k) > 0$ since $a \in a^\star(\nu)$.

$$\begin{aligned}
\mathbb{P}_\nu(a \in a^\star(\nu) \cap \widehat{a}^\star(\nu,T)^{\mathsf{c}}) &\leq \mathbb{P}_\nu(\exists\, k \in \mathbb{A}\setminus\{a\} \text{ s.t } \widehat{\boldsymbol{\mu}}_a(T) \preceq \widehat{\boldsymbol{\mu}}_k(T)),\\
&\leq \mathbb{P}_\nu(\exists\, k \in \mathbb{A}\setminus\{a\} \text{ s.t } \forall\, d \in [D], \widehat{\mu}_k^d(T) - \widehat{\mu}_a^d(T) - (\mu_k^d - \mu_a^d) \geq (\mu_a^d - \mu_k^d)),\\
&\leq \sum_{k\in\mathbb{A}\setminus\{a\}} \mathbb{P}_\nu(\forall\, d \in [D], \widehat{\mu}_k^d(T) - \widehat{\mu}_a^d(T) - (\mu_k^d - \mu_a^d) \geq (\mu_a^d - \mu_k^d)),\\
&\leq \sum_{k\in\mathbb{A}\setminus\{a\}} \mathbb{P}_\nu(\widehat{\mu}_k^{d_{a,k}}(T) - \widehat{\mu}_a^{d_{a,k}}(T) - (\mu_k^{d_{a,k}} - \mu_a^{d_{a,k}}) \geq \underbrace{(\mu_a^{d_{a,k}} - \mu_k^{d_{a,k}})}_{=M(a,k)} \geq \Delta_a),\\
&\leq (K-1)\exp\left(-\frac{T\Delta_a^2}{4\sigma^2 K}\right). \quad \text{(union bound and } \sigma\text{− subgaussian Hoeffding)}
\end{aligned}$$

Step 3: Conclusion. Putting the two preceding steps together yields

$$\begin{aligned}\mathbb{P}_\nu(a^\star \neq \widehat{a}^\star) &\leq \sum_{a\in a^\star(\nu)} K\exp\left(-\frac{T\Delta_a^2}{4\sigma^2 K}\right) + \sum_{a\in\mathbb{A}\setminus a^\star(\nu)} D\exp\left(-\frac{T\Delta_a^2}{4\sigma^2 K}\right),\\ &\leq K(K+D)\exp\left(-\frac{T}{4\sigma^2 K H^u}\right),\\ &\leq K(K+D)\exp\left(-\frac{T}{4\sigma^2 K\log(2K)H_2^u}\right),\end{aligned}$$

where the last inequality derives from (3.13). □

As explained in the previous chapter, for the same reasons, when $\Delta_1 \approx \cdots \approx \Delta_K$, the $1/K$ in the exponential can be removed. We would like to recall that, as mentioned in section 3.3 when $\varepsilon \approx 0$, algorithm 3 is not efficient. Indeed, in this scenario, if K is not too large, on reasonably not too difficult instances, we may write

$$H_{\varepsilon\approx 0} \gg KH^u,$$

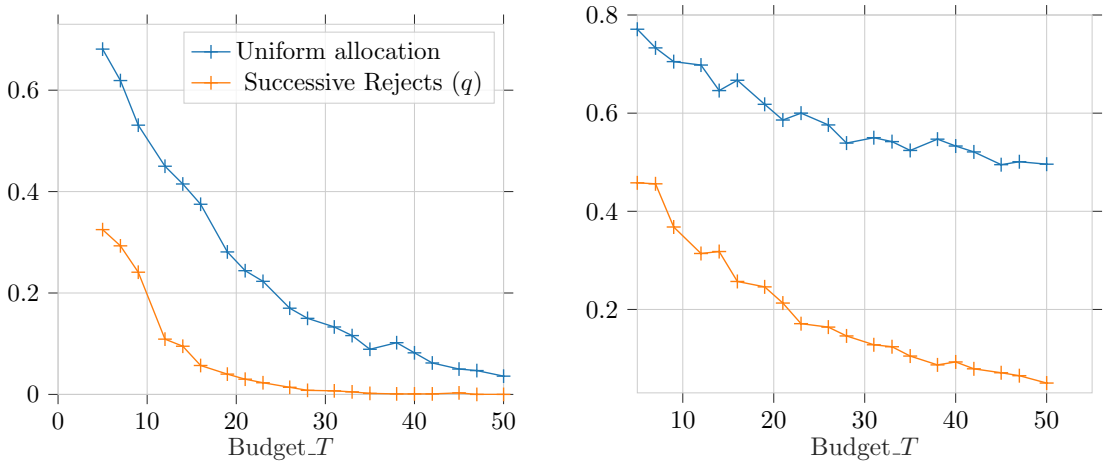
making the uniform allocation a better strategy than algorithm 3. On the contrary, if ε is not too small and K is large, we expect algorithm 3 to outperform the uniform allocation strategy. Finally, compared to Successive Rejects (q) , the uniform allocation has an additional $1/K$ in the exponential. The gaps are the same for sub-optimal arms. However, should we always guarantee that that $H^q \leq H^u$? Indeed, this does not always hold. For example letting $\theta > 0$, consider a bandit with arms' means

$$\boldsymbol{\mu}_1 := (1,1)^\top, \qquad \boldsymbol{\mu}_2 := (2,2)^\top \qquad \text{and} \qquad \boldsymbol{\mu}_3 := (2+\theta, 1-\theta)^\top,$$

direct computation yields

$$H^q = 3, \qquad \text{and} \qquad H^u = 1 + \theta^{-2} + (1+\theta)^{-2}.$$

When $\theta \approx 0$ (difficult instance), $H^q \ll H^u$ on the contrary, when θ is large (easy instance), we may have $H^u \leq H^q$ but still $H^q \leq KH^u$, that is Successive Rejects (q) still outperforms the uniform allocation strategy. This is what we expect, in particular when K is large. Figure 3.1 shows the empirical mis-identification error for this bandit for $\theta_1 = 1$ and $\theta_2 = 0.01$, showing that Successive Rejects (q) outperforms uniform allocation in both cases.



(a) $\theta = 1$ (b) $\theta = 0.01$

FIGURE 3.1: Empirical mis-identification rate $\mathbb{P}_\nu(\widehat{a}^\star(\nu) \neq a^\star(\nu))$. The x-axis represents 20 equally spaced values between 5 and 50. The results are averages overs 1000 runs for each value of budget.

4

Application To Clinical Trials

To give some context, imagine a drug development process in which one is trying to identify the appropriate dose that patients should receive. One is given $K = 5$ candidate dose levels (also called arms) and the goal is to identify the dose minimizing some (possibly conflicting) indicators. In the simulated vaccine campaign, T patients are recruited and some indicators are measured on a regular basis (see Figure 4.1). In our problem, we consider that only the measure at day 42 is of interest. The datasets are provided by the joint *Inria-Inserm* team *SISTM* with which we have collaborating during the internship.

We are given two datasets corresponding to 2 different scenarios but sharing the same structure. Each arm is assigned to 326 patients and interim analyses are conducted on a

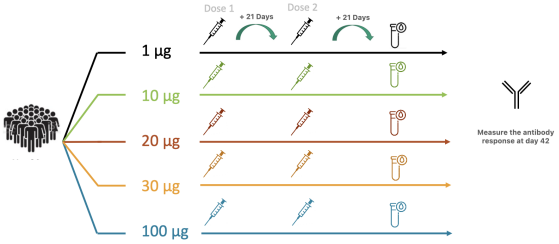


FIGURE 4.1: Illustration of the simulated vaccine protocol.

daily basis from day 0 (day of injection) to day 42. Six indicators (level of different antibodies) are recorded. Each dataset is parsed as a `NumPy` array of shape $(5, 326, 43, 6)$. The indicators of interest for our applications are the records at day 42. Letting ν denotes this tabular bandit, the sequential interaction described so far still holds, at round t, the forecaster selects an arm $A_t \in \mathbb{A} := \{1, 2, 3, 4, 5\}$ (corresponding resp. to the dose levels $1, 10, 20, 30, 100\mu g$) and observes the $6-$ dimensional reward at index $(A_t, T_{A_t}(t-1) + 1, 43)$ of the table, where $T_{A_t}(t)$ corresponds to the number of times arm A_t has been selected up to arm t. For each dataset (in superscript), we have

$$a^\star(\nu^1) = \{5\} \qquad \text{and} \qquad a^\star(\nu^2) = \{2, 4, 5\},$$

and the maximal budget allowed is $T = 326$ (the sum of the number of pulls of the arms must not exceed this value) for both datasets. For the first dataset, we would expect Successive Rejects (f) or Sequential Halving (f) to correctly identify the optimal arm when f is a Pareto-increasing function (Lemma 2.2). For this dataset, we will test six different function $f_{e_1}, \dots f_{e_6}$ where

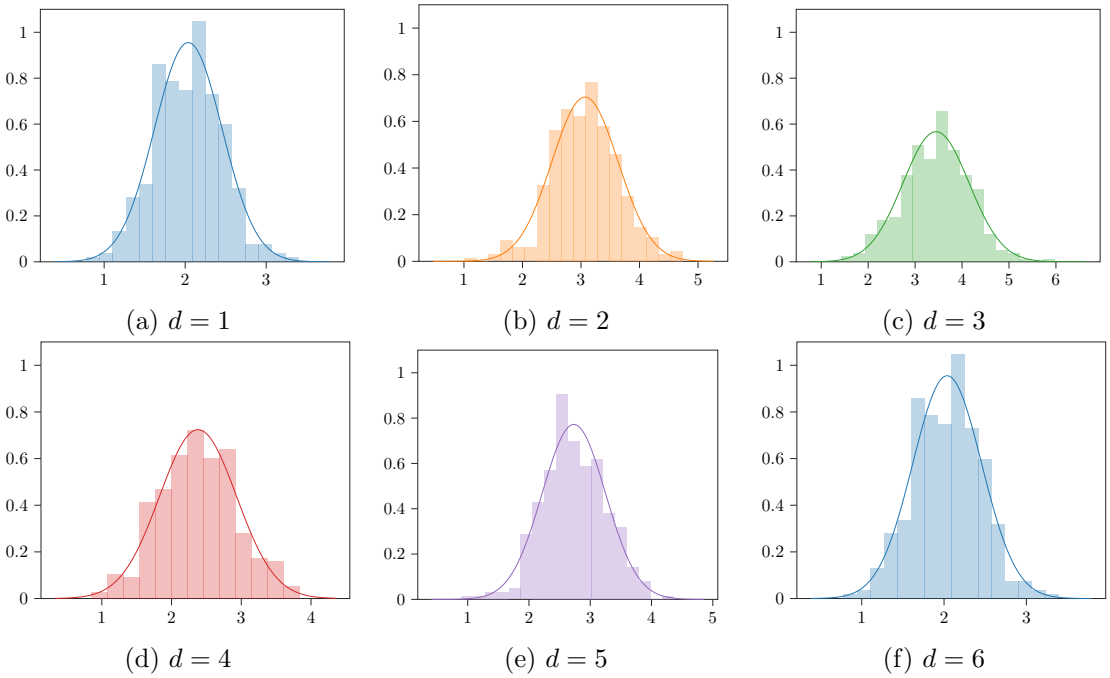
$$f_{e_i} : \mathbf{x} \in \mathbb{R}^D \mapsto e_i^\top \mathbf{x},$$

and e_i is the vector with zeros everywhere except the $i-$th coordinate which is 1. Table 4.1 summarizes the complexity of the bandit problems corresponding to each function (computed using the arms empirical means on the entire dataset). Figure 4.2 shows the antibody response of arm 1 measured at the 43rd endpoint for the six (06) types of antibodies. The shapes of the histograms are similar for the others arms.

f_{e_i}	H^f
e_1	486.29
e_2	3481.08
e_3	17738.51
e_4	3815.82
e_5	2918.42
e_6	34796.49

TABLE 4.1: H^f for different function f.

Figure 4.4 shows the result for each function f. The results are averaged over 5000 trials, each trial corresponding to a randomly shuffled copy of the dataset (shuffle the second dimension, that is the order in which patients are recruited). We can observe



(a) $d = 1$ (b) $d = 2$ (c) $d = 3$

(d) $d = 4$ (e) $d = 5$ (f) $d = 6$

FIGURE 4.2: Fitted-histograms of the six indicators of arm 1 measured at day 42.

that Successive Rejects (f) consistently outperforms Uniform allocation, and, the mis-identification error globally increases as H^f increases, that is the problem becomes harder. Finally, Figure 4.3 shows the mis-identification of the Pareto optimal set for the second dataset whose complexity (as defined in [Auer et al., 2016]) is $H(\nu^2) = 12.47 \times 1e6$, (that is a very hard instance).

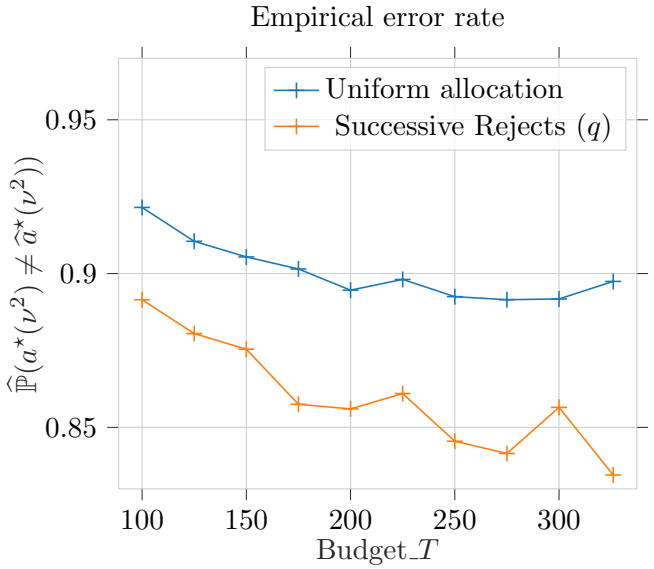


FIGURE 4.3: Mis-identification rate on the second dataset.

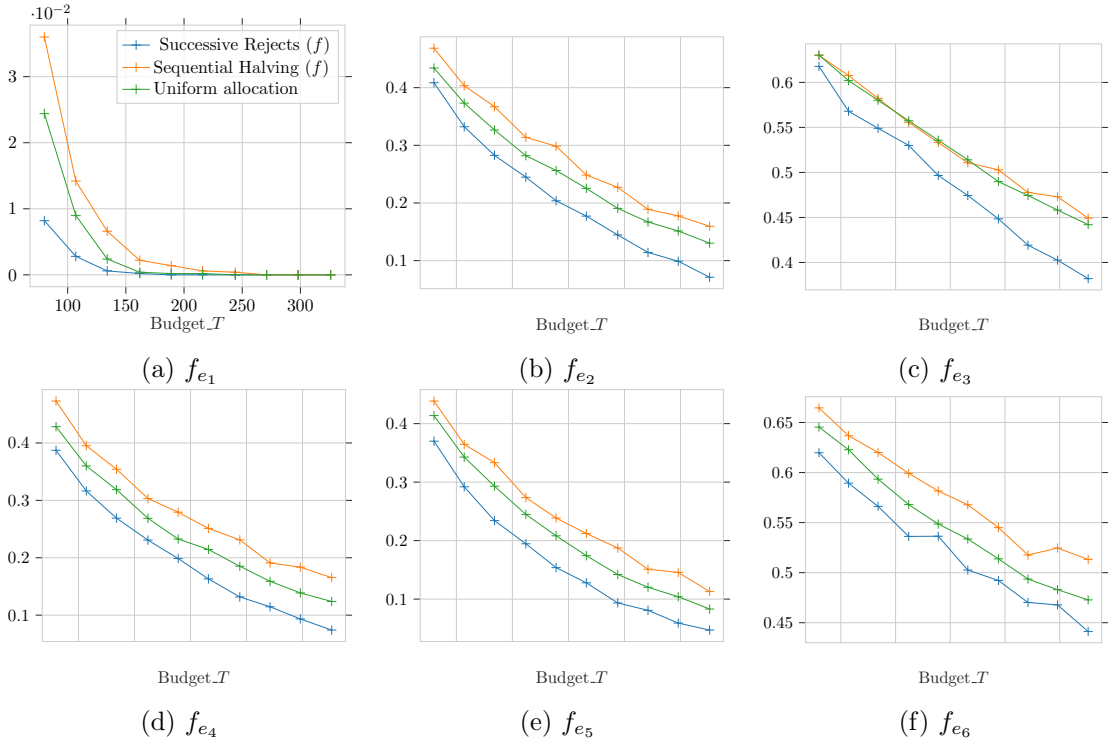


FIGURE 4.4: Empirical mis-identification rate $\mathbb{P}_{\nu^1}(\widehat{a}^\star(\nu^1, f) \neq a^\star(\nu^1, f))$ for the different functions f. The x-axis represents 10 equally spaced values between 80 and 326.

5

Conclusion and Final Remarks

In this internship, we have seen how to design some algorithms in the *fixed-budget* setting and prove some upper-bounds on their probability of error. We have been familiarizing with statistical tools extensively used in the bandit literature, such concentration inequalities, likelihood ratio test, confidence intervals, change of distributions etc.

Our algorithmic contributions include Successive Rejects (f) , Sequential Halving (f) , algorithm 3, Successive Rejects (q) with the accompanying theorems and the analysis of the uniform allocation strategy for Pareto optimal set identification. The lemmas and propositions proved in this report are also personal work. Whenever we knew the existing proofs they borrow from, we mentioned it in the report. Those, whose proofs are omitted are taken from references given in their statements.

Future research directions include providing an optimal algorithm in the *fixed-budget* setting, providing some *adaptive* algorithms in the *fixed-confidence* that could outperform [Auer et al., 2016]. Both challenges require to understand precisely what are the true sub-optimality gaps and complexity terms in the multidimensional case, going from those suggested in [Auer et al., 2016].

References

[Audibert and Bubeck, 2010] Audibert, J.-Y. and Bubeck, S. (2010). Best Arm Identification in Multi-Armed Bandits. In *COLT - 23th Conference on Learning Theory - 2010*, page 13 p., Haifa, Israel. 1, 2, 10, 11, 12, 14, 28, 31, 37, 38, 42

[Audibert et al., 2009] Audibert, J.-Y., Munos, R., and Szepesvári, C. (2009). Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902. 35

[Auer et al., 2016] Auer, P., Chiang, C.-K., Ortner, R., and Drugan, M. (2016). Pareto Front Identification from Stochastic Bandit Feedback. In Gretton, A. and Robert, C. C., editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 939–947, Cadiz, Spain. PMLR. 3, 28, 29, 30, 49, 51

[Aziz et al., 2019] Aziz, M., Kaufmann, E., and Riviere, M.-K. (2019). On Multi-Armed Bandit Designs for Dose-Finding Clinical Trials. 1, 21, 22

[Carpentier and Locatelli, 2016] Carpentier, A. and Locatelli, A. (2016). Tight (Lower) Bounds for the Fixed Budget Best Arm Identification Bandit Problem. 19

[Drugan and Nowe, 2013] Drugan, M. and Nowe, A. (2013). Designing multi-objective multi-armed bandits algorithms: A study. 3, 4, 23, 24

[Gabillon et al., 2012] Gabillon, V., Ghavamzadeh, M., and Lazaric, A. (2012). Best Arm Identification: A Unified Approach to Fixed Budget and Fixed Confidence. In Pereira, F., Burges, C. J., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc. 2, 31

[Garivier and Kaufmann, 2016] Garivier, A. and Kaufmann, E. (2016). Optimal Best Arm Identification with Fixed Confidence. Technical Report arXiv:1602.04589, arXiv. arXiv:1602.04589 [cs, math, stat] type: article. 2

[Jamieson et al., 2013] Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. (2013). lil' UCB : An Optimal Exploration Algorithm for Multi-Armed Bandits. 2

[Kaisa, 1999] Kaisa, M. (1999). *Nonlinear Multiobjective Optimization*, volume 12 of *International Series in Operations Research & Management Science*. Kluwer Academic Publishers, Boston, USA. 23

[Karnin et al., 2013] Karnin, Z., Koren, T., and Somekh, O. (2013). Almost Optimal Exploration in Multi-Armed Bandits. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pages III–1238–III–1246. JMLR.org. event-place: Atlanta, GA, USA. 1, 2, 14, 15, 17

[Katz-Samuels and Scott, 2018] Katz-Samuels, J. and Scott, C. (2018). Feasible Arm Identification. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2535–2543. PMLR. 1, 31, 32

[Kaufmann et al., 2014] Kaufmann, E., Cappé, O., and Garivier, A. (2014). On the Complexity of A/B Testing. In Balcan, M. F., Feldman, V., and Szepesvári, C., editors, *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 461–481, Barcelona, Spain. PMLR. 28

[Kaufmann and Kalyanakrishnan, 2013] Kaufmann, E. and Kalyanakrishnan, S. (2013). Information Complexity in Bandit Subset Selection. In *Conference On Learning Theory*, volume 30, Princeton, United States. JMLR: Workshop and Conference Proceedings. 2

[Lattimore and Szepesvári, 2020] Lattimore, T. and Szepesvári, C. (2020). *Bandit Algorithms*. Cambridge University Press. 5, 6, 7

[Locatelli et al., 2016] Locatelli, A., Gutzeit, M., and Carpentier, A. (2016). An optimal algorithm for the Thresholding Bandit Problem. 7, 32

A
List of Symbols

$\mathbb{A}$	a set of arms $\{1, \ldots, K\}$
$\nu(K, D)$	a K-armed D-dimensional bandit instance
$[K]$	the set of integers $\{1, \ldots, K\}$
$a^\star(\nu)$	the set of Pareto optimal arms of bandit ν
$a^\star_\epsilon(\nu)$	the set of ϵ-Pareto optimal arms of bandit ν
$a^\star(\nu, f)$	the optimal arms w.r.t function f
Δ_a	the sub-optimality gap of arm a
Δ^f_a	the sub-optimality gap of arm a w.r.t function f
$\Gamma(\nu, c_1, c_2)$	class of functions
H	a complexity of the bandit ν
H_2	a complexity of the bandit ν
$\mathbb{P}_\nu$	probability distribution under the bandit ν
$\mathbb{E}_\nu$	expectation under $\mathbb{P}_\nu$
$\mathcal{F}^\nu$	the filtration generated by the bandit ν