On Identifying Good Options under Combinatorially Structured Feedback in Finite Noisy Environments

Yifan Wu YWU12@UALBERTA.CA
András György GYORGY@UALBERTA.CA
Csaba Szepesvári SZEPESVA@UALBERTA.CA
Department of Computing Science, University of Alberta, Edmonton, AB T6G 2E8 CANADA

Abstract

We consider the problem of identifying a good option out of finite set of options under combinatorially structured, noisy feedback about the quality of the options in a sequential process: In each round, a subset of the options, from an available set of subsets, can be selected to receive noisy information about the quality of the options in the chosen subset. The goal is to identify the highest quality option, or a group of options of the highest quality, with a small error probability, while using the smallest number of measurements. The problem generalizes best-arm identification problems. By extending previous work, we design new algorithms that are shown to be able to exploit the combinatorial structure of the problem in a nontrivial fashion, while being unimprovable in special cases. The algorithms call a set multi-covering oracle, hence their performance and efficiency is strongly tied to whether the associated set multi-covering problem can be efficiently solved.

1. Introduction

Consider the problem of identifying the most rewarding option(s) out of finitely many. At your disposal are a number of probing devices, or just *probes*, that give you noisy measurements of the quality of a select set of options. More precisely, each *probe* is associated with a *known subset* of options whose quality the probe will measure. In a sequential process, the goal is to select the probes so that one can stop early to return, with high probability, a sufficiently rewarding option (or a set of options). As a specific example, consider the problem of identifying the segment on a road network that is in the worst shape after a long winter. Measurements can be obtained by sending trucks checking the road for potholes along the paths they travel on. The trucks must return to their garage every day. Here, the options correspond to road segments, the probes correspond to a closed walk in the road network that starts from the garage. Somewhat ironically, a road segment is "rewarding" (from the point of view of how beneficial it is to sending there the repair team) if it has many potholes.[1] Measurements are noisy, as potholes are easy to miss.

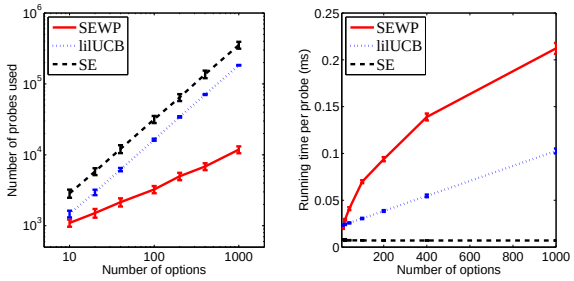

Figure 1. A specialized algorithm (SEWP) proposed in this paper can take nontrivial advantage of the probe structure as compared with simple adaptations of earlier algorithms, while being only marginally more expensive. All algorithms maintain the same error-rate. The plot on the left-hand-side uses a log-log-scale. Due to the special structure of the problem, the expected stopping time of the specialized algorithm scale linearly with $\sqrt{K}$, while the others scale linearly with K, the number of options.

Problems like the above one abound. Numerous quality assurance and surveying tasks are such that measurements give simultaneous information about multiple entities due to physical constraints on the measurement process. Application areas include technical computing (e.g., networking), biology (ecology, microbiology, etc.), physics, etc.

Of course, even though individual measurements might be

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

[1] In practice, one may want a whole "plan" at the end for the repair team. As often, we took the liberty of simplifying the problem to be able to focus on how the structure of probes should be used.

impossible, it is always possible to treat each probe as one that gives individual measurements for the options associated with it, though this could be wasteful (cf. Fig. 1). The main topic of the present paper is how to exploit, with efficient algorithms, when probes give information about multiple options.

The special case when each probe measures a single option, is known as the *best arm identification* problem, whose history goes back more than half a century (Bechhofer, 1958; Paulson, 1964), and with much activity in the last decade (e.g., Even-Dar et al. 2002, Mannor & Tsitsiklis 2004, Audibert et al. 2010, Kalyanakrishnan & Stone 2010, Bubeck et al. 2011, Kalyanakrishnan et al. 2012, Gabillon et al. 2012, Karnin et al. 2013, Kaufmann & Kalyanakrishnan 2013, Bubeck et al. 2013, Jamieson et al. 2014, Kaufmann et al. 2015, Zhou et al. 2014, Chen et al. 2014).

In this paper we consider two basic settings: identifying the best option with a prespecified error probability while using the smallest possible number of probes, and identifying a group of options of a fixed size, again with a prespecified error probability with the smallest possible number of probes. For the first setting, we propose two algorithms, SEWP and EGEWP described in Section 3, extending the works of Even-Dar et al. (2002) and Karnin et al. (2013). They work by constructing coverings with the probes of the sets of options not eliminated. The second algorithm removes a logarithmic term from the upper bound and it required a non-trivial extension of the median elimination method of Even-Dar et al. (2002). For the second setting, in Section 4, the quality of a group returned is assessed either by the quality of the worst option in the group (following Kalyanakrishnan & Stone (2010)), or by the average quality of options in the group (Zhou et al., 2014). We propose a single algorithm (SARWP) that essentially covers both cases. For the average quality, our distribution dependent upper bound is novel even in the bandit case and also near optimal in the worst case compared with the lower bound proposed by Zhou et al. (2014). For simple probe structures (singletons, or when a probe that covers all options is available), our algorithms are shown to be essentially unimprovable. We also give lower bounds for general probe structures. While both our lower and upper bounds express how the structure of the probes interferes with the structure of payoffs, they differ in subtle ways and it remains for future work to see whether there is a gap between them.

Due to space constraints, proofs and some experimental results are relegated to the appendix.

2. Preliminaries

In this section, we formulate the problem studied, as well as introducing the set covering problem, which will play an important role in our algorithms and analysis. We start by defining some notation.

2.1. Notation

The set of natural numbers will be denoted by $\mathbb{N}$, which includes zero. For a positive natural number n, $[n]$ denotes the set of integers between 1 and n: $[n] = \{1, \dots, n\}$. The power set, i.e., the set of all subsets of a set S, will be denoted by 2^S. As usual, functions, mapping set X to set Y will be viewed as elements of Y^X. For $v \in Y^X$, we will often write v_x instead of $v(x)$ to minimize clutter. This also helps with the next convention: When $U \subset X$, we will use v_U to denote the restriction of $v \in Y^X$ to U: $v_U(u) = v(u)$, $u \in U$. We identify $Y^{[n]}$ with Y^n (the set of n-tuples) in the natural way, which allows us to use notation v_U for $v \in Y^n \equiv Y^{[n]}$. The cardinality of a set S is denoted by $|S|$. Certain symbols will be reserved to denote elements of certain sets (i.e., p will always be an element of set $\mathcal{P}$). When using such reserved symbols, we will abbreviate (e.g.) $\sum_{p \in \mathcal{P}} f(p)$ to $\sum_p f(p)$. We will use $\log(\cdot)$ to denote the natural logarithm function.

2.2. Problem Formulation

A decision maker is given a pair $([K], \mathcal{P})$, where elements of $[K]$ are called arms, or, interchangeably, actions, and $\mathcal{P} \subset 2^{[K]}$ such that the sets in $\mathcal{P}$ cover $[K]$: $\cup \mathcal{P} = [K]$. Elements of $\mathcal{P}$ are called *probes*. A problem instance D, or *environment*, is specified by K distributions over the reals, $D = (D_1, \dots, D_K)$. The decision maker does not have direct access to these distributions. For $1 \le i \le K$, we think of distribution D_i as the distribution of "rewards" associated with arm i. We assume that the mean reward $\mu_i = \int x D_i(dx)$ of each arm is well defined. Further assumptions on D_i will be given later.

The goal of the decision maker is to find arms with the largest mean reward. For this, the decision maker can query the rewards of the arms by using the probes in a sequential manner. In particular, for each round $t = 1, 2, \dots$, first a random reward $X_{t,i} \sim D_i$ is generated for each arm i from its associated distribution. It is assumed that $X_{t,i}$ is independent of the other rewards $(X_{s,j})_{s \neq t \text{ or } j \neq i}$. We set $X_t = (X_{t,1}, \dots, X_{t,K}) \in \mathbb{R}^K$. In round $t = 1, 2, \dots$, the decision maker chooses a probe $p_t \in \mathcal{P}$ based on her past observations, to observe the values $X_{t,i}$ for each arm i in p_t; with our earlier introduced notation we can write that the decision maker observes $X_{t,p_t} \doteq (X_t)_{p_t} \in \mathbb{R}^{p_t}$. At the end of each round, the decision maker can decide between continuing or stopping to return a list of guesses (or a single guess) on the indices of the good arms. The goal is to stop as soon as possible, while avoiding poor guesses.

The following specific problem settings will be considered:

(i) *Fixed confidence, best-arm identification*. The *optimal arm* is unique: If $\mu^\star = \max_{i\in[K]} \mu_i$, $\max_{i:\mu_i\neq\mu^\star} \mu_i < \mu^\star$. The goal of the decision maker is to identify the index $i^\star = \operatorname{argmax}_{i\in[K]} \mu_i$ of the optimal arm. The decision maker is given a *confidence* parameter $0 \le \delta < 1$ and it is required that the guess returned after τ probes must be correct on an event $\mathcal{E}$ with probability at least $1-\delta$. Decision makers are compared based on their *probe complexity*, i.e., the number of probes they use when the "good event" $\mathcal{E}$ happens.

(ii) *PAC subset selection*. There are two subproblems that we consider. In both cases the decision maker is given a confidence, $0 \le \delta < 1$, a suboptimality threshold $\varepsilon > 0$ and a subset cardinality $1 \le m \le K$. The problems differ in how a quality $q(S,\mu)$ measure is assigned to a subset $S \subset [K]$ of arms. In both problems, the goal is to find a subset of arms of cardinality m such that $q(S,\mu) \ge \max_{P\subset[K]:|P|=m} q(P,\mu) - \varepsilon$ and with probability $1-\delta$, the decision maker must return a subset satisfying the above quality constraint. As before, decision makers are compared based on how many probes they use before stopping. The two quality measures considered are the reward of the worst arm in the set and the average reward: $q_{\min}(S,\mu) = \min_{i\in S} \mu_i$ and $q_{\text{avg}}(S,\mu) = \frac{1}{|S|}\sum_{i\in S} \mu_i$, $S \subset [K]$, $|S| = m$. We call the corresponding problems the *strong* and the *average* PAC subset selection problems.

An algorithm used by a decision maker to select probes, stop and return a guess will be said to be *admissible* with respect to a class of environments, if, for *any* environment within the class and any $0 \le \delta < 1$, the guess computed is correct (according to the previous requirements) with probability $1-\delta$.

The above problems have been considered in the past in the special case when $\mathcal{P}$ contains singletons only, by a number of authors (see Section 1 for some references). We shall call these the "bandit" problems. While one can readily apply the algorithms developed for the bandit case to our problem, the expectation is that the probe complexity of reasonable algorithms should improve considerably as $\mathcal{P}$ becomes "richer" (this was illustrated in Fig. 1). The question is how the structure of $\mathcal{P}$ together with the problem instance influences the problem complexity. For example, in the extreme case when $\mathcal{P}$ contains $[K]$, we expect the probe complexity of reasonable algorithms to scale sublinearly with K, whereas in the bandit case a linear scaling is unavoidable. The case when $\mathcal{P} = \{[K]\}$ will be called the *full information case*.

Note that since all probes "cost" the same amount (one unit of time), a reasonable algorithm will avoid any probe p that is entirely included in some other probe $p' \in \mathcal{P}$. Hence, we may as well assume that the set of probes does not have nontrivial chains in it.

We will present results for the class of environments $\mathcal{D}_{\text{sg}}$ with the following restrictions: For each $1 \le i \le K$, D_i is sub-Gaussian with common parameter $\sigma^2 = 1/4$:

$$\log \int_{\mathbb{R}} e^{-\lambda(x-\mu_i)} D_i(dx) \le \lambda^2\sigma^2/2 = \lambda^2/8$$

for all $\lambda \in \mathbb{R}$. To simplify the presentation of our results, without loss of generality, we *assume that* $\mu_1 \ge \mu_2 \ge \cdots \ge \mu_K$. (note that, obviously, the algorithms do not use this assumption). For further simplicity, we assume that $\Delta_i \in [0,1]$ for all $i \in [K]$ where $\Delta_i = \mu_1 - \mu_i$, $2 \le i \le K$. Our assumptions on the reward distributions D_i are satisfied if, for example, D_i has bounded support.

We will present algorithms, which will be shown to be admissible for $\mathcal{D}_{\text{sg}}$ and we will bound their probe complexities. The bounds on the probe complexities will be given in terms of the *(suboptimality) gaps* Δ_i, $2 \le i \le K$, i.e., they will be dependent on the distributions $D = (D_1, \ldots, D_K)$. Hence, we call them distribution dependent bounds. We will accompany our constructive results with lower bounds, putting a lower limit on the probe complexity of all admissible algorithms. Again, these will be given in terms of the gaps Δ_i.

2.3. Set Multi-Cover Problems

Probes allow one to "explore" multiple arms simultaneously. Clever algorithms should use the probes in a smart way to guarantee the necessary number of samples for each of the arms while using the smallest number of probes. If, for example, $n \in \mathbb{N}$ observations are enough from each of the arms to distinguish their mean payoff from that of the optimal arm, then an intelligent algorithm would try to create the smallest *covering* of $[K]$ using the subsets in $\mathcal{P}$ to meet this requirement. More generally, for $J \subset [K]$, we define

$$\mathcal{C}_{\text{IP}}(J,n) = \min\left\{ \textstyle\sum_p s_p : s \in \mathbb{N}^{\mathcal{P}}, \sum_{p:i\in p} s_p \ge n, i \in J \right\}$$

to be the *cost* of the smallest n-fold multi-covering of elements of J. Any $s \in \mathbb{N}^{\mathcal{P}}$ achieving the minimum is called an *optimal (integral) n-cover* of J, while a feasible vector s is called an n-cover. Given an n-cover $s \in \mathbb{N}^{\mathcal{P}}$, we will say that probe p belongs to s (writing $p \in s$) if $s_p > 0$. The optimization problem defining $\mathcal{C}_{\text{IP}}$ is a linear integer program (hence the IP in $\mathcal{C}_{\text{IP}}$). Relaxing the *integrality* constraint $s \in \mathbb{N}^{\mathcal{P}}$ to the nonnegativity constraint $s \in [0,\infty)^{\mathcal{P}}$, we get a so-called *fractional optimal n-cover* of J by solving the otherwise identical optimization problem. The resulting optimal value will be denoted by $\mathcal{C}_{\text{LP}}(J,n)$. Note that

the relaxed problem is a linear program, explaining "LP" in $\mathcal{C}_{\mathrm{LP}}$. While this linear program has potentially exponentially many variables in K, it can still be efficiently solved provided an efficiently computable membership oracle is available for its dual (Grötschel et al., 1993). Both $\mathcal{C}_{\mathrm{IP}}(J,n)$ and $\mathcal{C}_{\mathrm{LP}}(J,n)$ can be extended to non-integer values of n.

It follows immediately from the definitions that $\mathcal{C}_{\mathrm{LP}}(J,n) \le \mathcal{C}_{\mathrm{IP}}(J,n)$. Further, for any $a > 0$, $\mathcal{C}_{\mathrm{LP}}(J,a\,n) = a\,\mathcal{C}_{\mathrm{LP}}(J,n) = an\,\mathcal{C}_{\mathrm{LP}}(J,1)$. The *integrality gap* for a set multi-covering problem instance is given by $(\mathcal{P},J,n)$ is $\mathcal{C}_{\mathrm{IP}}(J,n)/\mathcal{C}_{\mathrm{LP}}(J,n)$ (Vazirani, 2001).

Our algorithms will need "small" n-covers for various subsets $J \subset [K]$. Depending on the structure of $\mathcal{P}$, calculating an optimal multi-cover of J may be easy or hard[2] (e.g., Slavik, 1998; Schrijver, 2003; Korte & Vygen, 2006). Thus, to keep the presentation general, our algorithms will rely on a set multi-covering *oracle* COrcl, which given $J,n,\mathcal{P}$, returns an n-fold multi-cover of J using the sets in $\mathcal{P}$. Denote by $\mathcal{C}_{\mathrm{O}}(J,n)$ the cost of the multi-cover returned by the oracle on J,n (as with $\mathcal{C}_{\mathrm{IP}}$ and $\mathcal{C}_{\mathrm{LP}}$ the dependence on $\mathcal{P}$ is suppressed). The oracle's integral (fractional) approximation gap, $\mathcal{G}_{IP}(O,\mathcal{P})$ ($\mathcal{G}_{LP}(O,\mathcal{P})$), is the worst-case multiplicative loss due to using COrcl in place of an optimal integral (fractional) cover. In particular, with $\star \in \{IP, LP\}$,

$$\mathcal{G}_\star(O,\mathcal{P}) = \sup_{n\in\mathbb{N}^+, J\subset[K]} \frac{\mathcal{C}_{\mathrm{O}}(J,n)}{\mathcal{C}_\star(J,n)}\,.$$

Let $d = \max_{p\in\mathcal{P}} |p|$ be the maximum number of actions that can be covered by a single probe. If the set-system $\mathcal{P}$ has no special structure, one possibility is to use the greedy algorithm G as the oracle. This algorithm works by sequentially setting $s_p = n$ for the probe $p \in \mathcal{P}$ that covers the maximum number of active arms in J and then deactivates the arms that are covered by p, until all arms are deactivated. Further, $\mathcal{G}_{LP}(O,\mathcal{P}) \le 1 + \log(d) \le 1 + \log(K)$. Lovász (1975) showed that $\mathcal{C}_{\mathrm{G}}(J,1) \le (1+\log d)\mathcal{C}_{\mathrm{LP}}(J,1)$. Then, $\mathcal{C}_{\mathrm{G}}(J,n) = n\,\mathcal{C}_{\mathrm{G}}(J,1) \le (1+\log d)n\,\mathcal{C}_{\mathrm{LP}}(J,1) = (1+\log d)\mathcal{C}_{\mathrm{LP}}(J,n)$, showing that the required inequality indeed holds. Raz & Safra (1997) proved that the exists some constant $c > 0$ such that, unless $P = NP$, no approximation ratio of $c\log(K)$ can be achieved, so in a worst-case the greedy algorithm is a near-optimal approximation algorithm.

3. Finding the Best Arm

In this section we present two algorithms and their analysis for the fixed confidence, best-arm identification problem. Recall that in this problem, given a set of probes $\mathcal{P}$ and a confidence $\delta \in (0,1]$, we need to design a sequential procedure that identifies the best arm $i^\star$ with probability at least $1-\delta$ using as few probes as possible.

3.1. Successive Elimination with Probes

The first algorithm modifies the successive elimination algorithm of Even-Dar et al. (2002) to take into account the richer observation structure of our problem. Recall that the algorithm of Even-Dar et al. (2002) works in phases, in each phase observing a certain number of rewards for each remaining candidate actions. At the end of the phase the provably suboptimal actions are eliminated. The number of observations in each phase depends only on the phase index. The process stops when the candidate set contains a single element. The main difference to the algorithm of Even-Dar et al. (2002) is that in each phase our algorithm, which we call Successive Elimination with Probes (SEWP), computes a set multi-covering for the remaining candidate actions given the probes, with a requirement adjusted to the phase index. The returned multi-cover is then used to get the observations for the remaining actions.

Algorithm 1 SuccessiveEliminationWithProbes (SEWP)

Inputs: K, δ, $\mathcal{P}$, observation scheduling function $f : \mathbb{N} \to \mathbb{N}$ and confidence function $g : \mathbb{N} \times (0,1] \to [0,\infty)$.
Initialize candidate set: $A_1 = [K]$.
for $t = 1, 2, \ldots$ **do**
  $C(t) \leftarrow \mathrm{COrcl}(A_t, f(t), \mathcal{P})$.
  Use each p in $C(t)$ for $C_p(t)$-times to get new observations.
  For each $i \in A_t$, let $\hat{\mu}_i(t)$ be the mean of all observations so far for arm i.
  $A_{t+1} \leftarrow \{i \in A_t : \hat{\mu}_i(t) + 2g(t,\delta) > \max_{j\in A_t} \hat{\mu}_j(t)\}$.
  if $|A_{t+1}| = 1$ **then**
    Return the arm in A_{t+1}.
  end if
end for

Our first result shows that Algorithm 1 is admissible and gives an upper bound on its probe complexity. To state it, define the scheduling and confidence functions

$$f(t) = 2^t, \qquad g(t,\delta) = \sqrt{\frac{\log(4Kt^2/\delta)}{2^{t+1}}}\,. \tag{1}$$

For simplicity, assume that the arms are ordered in decreasing order of their mean rewards and $\Delta_2 > 0$, i.e., the optimal arm is unique. For $2 \le i \le K$ define

$$\widehat{T}_i(\delta) = 1 + \max\left\{ s \,:\, g(s,\delta) \ge \frac{\Delta_i}{4} \right\}, \tag{2}$$

$$\hat{N}_i(\delta) = \frac{128}{\Delta_i^2} \log\left(\frac{54K}{\delta} \log \frac{4}{\Delta_i} \right) \tag{3}$$

[2] Computing the exact solution for the decision version of set covering (i.e., when $n = 1$), when $\mathcal{P}$ can be any covering system, is known to be NP-hard (Vazirani, 2001).

and let $\widehat{T}_{K+1}(\delta) = 0$ and $\hat{N}_{K+1}(\delta) = 0$. Note that $2^{\widehat{T}_i(\delta)+1} \leq \hat{N}_i(\delta)$, and both are decreasing with $i \geq 2$ increasing.

Theorem 1. *Pick any $0 \leq \delta < 1$ and let SEWP run with inputs $(K, \delta, \mathcal{P}, f, g)$ with f, g given by (1). Then, with probability at least $1-\delta$, SEWP returns the optimal arm $i^\star = 1$ within N probes, where N satisfies*

$$N \leq \mathcal{G}_{IP}(O, \mathcal{P}) \sum_{i=2}^{K} \sum_{t=\widehat{T}_{i+1}(\delta)+1}^{\widehat{T}_i(\delta)} \mathcal{C}_{\mathrm{IP}}([i], 2^t) . \quad (4)$$

Furthermore, with $\hat{M}_i(\delta) \doteq \hat{N}_i(\delta) - \hat{N}_{i+1}(\delta)$,

$$N \leq \mathcal{G}_{LP}(O, \mathcal{P}) \sum_{i=2}^{K} \hat{M}_i(\delta)\, \mathcal{C}_{\mathrm{LP}}([i], 1) . \quad (5)$$

The bound (4) may be tighter than that shown in (5), but perhaps the second is a bit easier to understand.[3] For simplicity, let us explain (5). Once (5) is explained, the meaning of (4) follows. The term $\mathcal{G}_{LP}(O, \mathcal{P})$ is the price of using an oracle combined with some upper bounding that allowed us to arrive at this simpler result by resorting to the linearity properties of $\mathcal{C}_{\mathrm{LP}}$. The rest is what we call a sequential fractional multi-cover with the requirements that arm i be covered $\hat{N}_i(\delta)$ times: In a sequential multi-cover, the covering is not done in a single-shot, but is done in phases. In the first phase, all the arms must be covered $\hat{M}_K(\delta)$ times. In the next phase, all the arms but the last must be covered $\hat{M}_{K-1}(\delta)$ times, etc., up to the last phase when arms one and two must be covered $\hat{M}_2(\delta)$ times. Note that the total requirements for an arm i are $\hat{M}_K(\delta) + \hat{M}_{K-1}(\delta) + \cdots + \hat{M}_i(\delta) = \hat{N}_K(\delta) - \hat{N}_{K+1}(\delta) + \hat{N}_{K-1}(\delta) - \hat{N}_K(\delta) + \cdots + \hat{N}_i(\delta) - \hat{N}_{i+1}(\delta) = \hat{N}_i(\delta)$. Roughly $\hat{N}_i(\delta) \approx O(1/\Delta_i^2)$ is the number of observations needed from arm i (and one) in order to be able to tell which of the two arms has a bigger mean reward. Now, compared to (5), (4) uses a more precise expression for the number of probes, by relying on the the phase structure of the algorithm.

An alternative choice of $f(t)$ and $g(t, \delta)$ is that $f(t) = 1$ and $g(t, \delta) = \sqrt{\frac{\log(4Kt^2/\delta)}{t}}$, which leads to $\hat{N}_i(\delta) = O\left(\frac{1}{\Delta_i^2} \log \frac{K}{\delta \Delta_i}\right)$ instead.

The proof, which borrows ideas from Even-Dar et al. (2002), is in Appendix A.1. To prove that SEWP is admissible, one only needs to show that when none of the confidence intervals based on g used in the elimination step fail, the optimal arm will not be eliminated. This essentially relied on Hoeffding's inequality, union bounds and calculations. To calculate the bound on the probe complexity bound, one shows that arm i will be eliminated after phase $\hat{T}_i(\delta)$. This happens because in each phase the confidence sets of all arms decrease at a uniform rate.

Now, we argue that this bound is tight up to a $\log K$ factor, at least in some cases. In particular, in the bandit case, the covering problem is trivial and we can use an optimal covering oracle. Then, $\mathcal{C}_{\mathrm{O}}([i], 2^t) = i2^t$, and hence the bound becomes $O\left(\sum_{i=1}^{K} \frac{1}{\Delta_i^2} \log\left(\frac{K}{\delta} \log \frac{1}{\Delta_i}\right)\right)$. Up to a log factor, this matches the lower bound of Kaufmann et al. (2015) which takes the form $\Omega\left(\sum_{i=1}^{K} \Delta_i^{-2} \log(1/\delta)\right)$. Furthermore, as noted by Jamieson et al. (2014) (based on a result of Farrell (1964)) the $\log\log \Delta^{-1}$ term is necessary.

To examine the tightness of the upper bound, we derive a distribution dependent lower bound on the probe complexity of algorithms admissible for $\mathcal{D}_{\mathrm{sg}}$. Call an environment D a Gaussian environment with common variance σ^2 if for any $1 \leq i \leq K$, D_i is a Gaussian with variance σ^2.

Theorem 2 (Distribution-dependent lower bound)**.** *For any algorithm admissible for $\mathcal{D}_{sg}$, any confidence $0 < \delta < 1/2$, any probe set $\mathcal{P}$, any sequence $0 = \Delta_1 < \Delta_2 \leq \ldots \Delta_K$, if D is a Gaussian environment with common variance $\sigma^2 = 1/4$ and means $\mu_1 = \mu_2 + \Delta_2 = \cdots = \mu_K + \Delta_K$, if N is the number of probes used by the algorithm on D then*

$$\mathbb{E}[N] \geq \min_{s \in [0,\infty)^{\mathcal{P}}} \sum_{p \in \mathcal{P}} s_p \quad \textit{s.t.} \sum_{p: 1 \in p} s_p \geq \frac{1}{4\Delta_2^2} \log \frac{1}{6\delta} ,$$
$$\textit{and} \sum_{p: i \in p} s_p \geq \frac{1}{4\Delta_i^2} \log \frac{1}{6\delta} , \quad 2 \leq i \leq K .$$

The proof can be found in Appendix A.2.

Note that the lower bound clearly reflects the structure of $\mathcal{P}$. However, even disregarding the constants and logarithmic factors, there is still a gap between our upper and lower bounds: In the upper bound, as explained before, the size of a *sequential* cover that appears, while in the lower bound, the size of a "one-shot" cover is seen. Note that in either the bandit or the full information case, there is no gap between these quantities. We were able to establish a gap of $\log(K)$ when considering sequential and one-shot *integral* covers. However, it remains a very interesting open question whether the gap can be closed in the fractional case.

3.2. An Alternative Algorithm to Find the Best Arm

The second algorithm is a generalization of the exponential gap elimination algorithm of Karnin et al. (2013), which improves the logarithmic term in the sample complexity from $\log(\frac{K}{\delta} \log \frac{1}{\Delta})$ to $\log(\frac{1}{\delta} \log \frac{1}{\Delta})$ for the bandit problem. So we expect that generalizing that algorithm to our setting will have a similar improvement regarding the $\log K$ term.

The exponential gap elimination algorithm of Karnin et al.

[3] In fact, if $\mathcal{C}_{\mathrm{O}}(\cdot, n)$ is monotone increasing, (4) will hold with $\mathcal{C}_{\mathrm{O}}$ replacing $\mathcal{G}_{IP} \cdot \mathcal{C}_{\mathrm{IP}}$, further tightening the bound.

(2013) calls the median elimination algorithm of Even-Dar et al. (2002) as a subroutine, which finds an ε-optimal arm using $O(K\varepsilon^{-2}\log(1/\delta))$ samples with probability at least $1-\delta$ (an arm is ε-optimal iff its expected reward is at least $\mu_1-\varepsilon$). So before generalizing the exponential gap elimination algorithm, we need to first design a counterpart for the median elimination algorithm.

3.2.1. Median Elimination With Probes

Simply replacing the uniform sampling in each phase in the median elimination algorithm of Even-Dar et al. (2002) with a set multi-cover does not work (shown in Appendix B.1), so a more careful design is needed. Our proposed algorithm, called Median Elimination With Probes (MEWP) is shown in Algorithm 2. It essentially runs the original median elimination algorithm for bandits over a one-cover of all arms (that is, each probe in the cover is treated as an arm in the bandit setting), and in each phase we eliminate half of the *probes* that do not seem to cover a good arm. We stop running median elimination when a single probe covers all the remaining arms. Then the algorithm enters its second stage where we use this probe until we identify an almost optimal arm from the remaining ones. In the next theorem we prove that the algorithm is admissible, and give an upper bound on the number of probes required to find an ε-optimal arm.

Algorithm 2 MedianEliminationWithProbes

Inputs: K, $\delta\in(0,1]$, $\varepsilon>0$, $\mathcal{P}$.
Set $\varepsilon_t=\frac{\varepsilon}{6}(\frac{3}{4})^t$, $\delta_t=\frac{\delta}{2^{t+1}}$.
$C\leftarrow$ COrcl$([K],1,\mathcal{P})$, and define a partition of the arms as $A_1=\{\pi_p\subset p\,:\,p\in C,\cup_{p\in C}\pi_p=[K]\}$.
for $t=1,2,\ldots$ **do**
  for all $\pi\in A_t$ **do**
    Use $\frac{4}{\varepsilon_t^2}\log\frac{3|\pi|}{\delta_t}$-times $p\in C$ that covers π to get observations for each arm in p.
    Let $\hat{\mu}_\pi(t)=\max_{i\in\pi}\hat{\mu}_i(t)$, where $\hat{\mu}_i(t)$ is the empirical mean reward of arm i based on the observations in the actual phase t.
  end for
  Find the median $m(t)$ of $\{\hat{\mu}_\pi(t):\pi\in A_t\}$.
  Let $A_{t+1}=\{\pi\in A_t\,:\,\hat{\mu}_\pi(t)\geq m(t)\}$.
  if $|A_{t+1}|=1$ **then**
    terminate the loop and let $\widehat{\pi}^*$ be the single element of A_{t+1}
  end if
end for
If $|\widehat{\pi}^*|>1$, use the probe that covers $\widehat{\pi}^*$ for $\frac{8}{\varepsilon^2}\log\frac{2|\widehat{\pi}^*|}{\delta}$-times.
Return the arm $\hat{i}^*\in\widehat{\pi}^*$ with the highest empirical mean based on these observations.

Theorem 3. *With probability at least $1-\delta$, MEWP returns an ε-optimal arm $\hat{i}^*$, and N, the total number of probes used by the algorithm is*

$$N=O\left(\frac{\mathcal{C}_{\mathrm{O}}([K],1)}{\varepsilon^2}\log\frac{|\pi_{\max}|}{\delta}\right). \tag{6}$$

where $|\pi_{\max}|=\max_{\pi\in A_1}|\pi|$.

Note that we have $|\pi_{\max}|$ inside the log term instead of the expected 1. It can be shown that the argument of the log term cannot be 1 in our problem setting, at least in the full information case (where it has to be K). Detailed discussion about this can be found in Appendix B.2.

3.2.2. Exponential Gap Elimination Algorithm

Algorithm 3 ExpGapEliminationWithProbes

Inputs: K, δ, $\mathcal{P}$.
Initialize candidate set: $A_1=[K]$. Set $\varepsilon_t=\frac{1}{4\cdot 2^t}$, $\delta_t=\frac{\delta}{50t^3}$.
for $t=1,2,\ldots$ **do**
  $C(t)\leftarrow$ COrcl$(A_t,1,\mathcal{P})$.
  Create a partition Π_t of A_t such that $\Pi_t=\{\pi_p\subset p\,:\,p\in C(t),\cup_{p\in C(t)}\pi_p=A_t\}$.
  for $\pi_p\in\Pi_t$ **do**
    Use probe p for $\frac{2}{\varepsilon_t^2}\log\frac{2|\pi_p|}{\delta_t}$-times to get observations for each arm in p.
  end for
  For each $i\in A_t$, let $\hat{\mu}_i(t)$ be the mean of all observations in phase t for arm i.
  $i_t\leftarrow$ MedianEliminationWithProbes$(A_t,\frac{\varepsilon_t}{2},\delta_t)$.
  Let $A_{t+1}=\{i\in A_t\,:\,\hat{\mu}_i(t)\geq\hat{\mu}_{i_t}(t)-\varepsilon_t\}$.
  if $|A_{t+1}|=1$ **then**
    Return the arm in A_{t+1}.
  end if
end for

Given the MEWP algorithm, we continue with generalizing the exponential gap elimination algorithm. The new algorithm, called Exponential Gap Elimination with Probes (EGEWP), is shown in Algorithm 3. The new idea here is to use the partition-based exploration technique (as in the MEWP algorithm) and replace the bandit-case median elimination subroutine with MEWP. The analysis follows a combination of the techniques of Karnin et al. (2013) and the proof of Theorem 3. However, due to the more complicated observation structure, we are only able to prove a Δ_2 dependent upper bound on the number of probes:

Theorem 4. *If the oracle COrcl always returns the optimal solution for integer programming, EGEWP finds the optimal arm with probability at least $1-\delta$ after using*

$$O\left(\frac{\mathcal{C}_{\mathrm{O}}([K],1)}{\Delta_2^2}\log\left(\frac{|p_{\max}|}{\delta}\log\frac{1}{\Delta_2}\right)\right) \tag{7}$$

probes where $|p_{\max}|=\max_{p\in\mathcal{P}}|p|$.

If COrcl is not guaranteed to return the optimal integer cover, the above theorem still holds by making the following modification to the algorithm to ensure that Π_{t+1} is not worse than Π_t for every t: if $|\{\pi \in \Pi_t : \pi \cap A_{t+1} \neq \varnothing\}| < \mathcal{C}_{\mathrm{O}}(A_{t+1}, 1)$, then use the same partition pattern from Π_t for Π_{t+1}.

Compared to the bound for SEWP, the $\log K$ term is replaced with $\log |p_{\max}|$. More specifically, in the full information case, the upper bound becomes $O\left(\frac{1}{\Delta_2^2} \log\left(\frac{K}{\delta} \log \frac{1}{\Delta_2}\right)\right)$, which is the same as the upper bound for SEWP. In the bandit case, the algorithm is exactly the same as the exponential gap elimination algorithm of Karnin et al. (2013), which enjoys an optimal $O\left(\sum_{i=1}^{K} \frac{1}{\Delta_i^2} \log\left(\frac{1}{\delta} \log \frac{1}{\Delta_i}\right)\right)$ upper bound on the number of probes, and is better than the upper bound for SEWP in bandit case. Therefore, although not formally proved, we expect that EGEWP enjoys an improved probe complexity compared with SEWP.

4. PAC Subset Selection

In this section, we consider the two PAC subset selection problems introduced in Section 2. The first, named *strong* PAC subset selection, is the same as the EXPLORE-m problem introduced by Kalyanakrishnan & Stone (2010) where the goal is to find m (ε, m)-optimal arms. The second problem, named *average* PAC subset selection, is to select a subset of m arms with ε-optimal average reward, introduced by Zhou et al. (2014).

The basic idea of our approach is to generalize our SEWP algorithm with two modifications: (i) First, besides rejecting the arms that cannot be in the best m arms after each phase, we also accept arms that have enough confidence to be one of the best m arms, which shares a similar idea with the Racing algorithm in Kaufmann & Kalyanakrishnan (2013). (ii) Specific stopping conditions are designed to meet the ε-relaxation in the problem definition.

To make it easier to express the probe complexity, we introduce a new symbol $\Delta_i^{(\varepsilon,m)}$ defined by $\Delta_i^{(\varepsilon,m)} = \max\{\mu_i - \mu_{m+1}, \varepsilon\}$ if $i \leq m$ and $\Delta_i^{(\varepsilon,m)} = \max\{\mu_m - \mu_i, \varepsilon\}$ if $i > m$. We then sort $\Delta_i^{(\varepsilon,m)}$ for all $i \in [K]$ in ascending order and let $S_{(i)}$ be the first i arms in the list, while $\Delta_{(i)}^{(\varepsilon,m)}$ denotes the i-th smallest entry.

Analogously to Theorem 1, let $f(t) = 2^t$, $g(t,\delta) = \sqrt{\frac{\log(4Kt^2/\delta)}{2^{t+1}}}$, and define

$$\hat{N}_{(i)}(\varepsilon,\delta) = \frac{128}{\left(\Delta_{(i)}^{(\varepsilon,m)}\right)^2} \log\left(\frac{54K}{\delta} \log \frac{4}{\Delta_{(i)}^{(\varepsilon,m)}}\right) \quad (8)$$

and let $\hat{N}_{(K+1)}(\varepsilon,\delta) = 0$.

Note that $\hat{N}_{(1)}(\varepsilon,\delta) = \hat{N}_{(2)}(\varepsilon,\delta)$ since $\Delta_{(1)}^{(\varepsilon,m)} = \Delta_{(2)}^{(\varepsilon,m)} = \max\{\mu_m - \mu_{m+1}, \varepsilon\}$. Also let $\hat{M}_{(i)}(\varepsilon,\delta) \doteq \hat{N}_{(i)}(\varepsilon,\delta) - \hat{N}_{(i+1)}(\varepsilon,\delta)$.

4.1. *Strong* PAC Subset Selection

First we propose an algorithm that returns a subset $\hat{S}^*$ containing m (ε, m)-optimal arms with high probability. An arm i is defined to be (ε, m)-optimal iff $\mu_i \geq \mu_m - \varepsilon$. This requirement is the same as $q_{\min}(\hat{S}^*, \mu) \geq q_{\min}([m], \mu) - \varepsilon$ where $q_{\min}(S, \mu) = \min_{i \in S} \mu_i$.

The algorithm, called Successive Accept Reject with Probes (SARWP) is shown in Algorithm 4. The following theorem shows that Algorithm 4 is admissible and the probe complexity is bounded.

Algorithm 4 SuccessiveAcceptRejectWithProbes

Inputs: K, m, ε, δ, $\mathcal{P}$, observation scheduling function $f : \mathbb{N} \to \mathbb{N}$ and confidence function $g : \mathbb{N} \times (0,1] \to [0, \infty)$.
Initialize candidate set $A_1 = [K]$, accepted arms $A_1^a = \varnothing$, rejected arms $A_1^r = \varnothing$.
for $t = 1, 2, \ldots$ **do**
  $C(t) \leftarrow \mathrm{COrcl}(A_t, f(t), \mathcal{P})$.
  Use each $p \in C(t)$ for $C_p(t)$-times to get new observations.
  For each $i \in A_t$, let $\hat{\mu}_i(t)$ be the mean of all observations so far for arm i. Sort the arms in A_t in descending order of $\hat{\mu}_i(t)$. Let H_t be the first $m - |A_t^a|$ arms and $L_t = A_t \setminus H_t$.
  if $\min_{i \in H_t} \hat{\mu}_i(t) \geq \max_{i \in L_t} \hat{\mu}_i(t) + 2g(t,\delta) - \varepsilon$ **then**
    Return $\hat{S}^* = A_t^a \cup H_t$ as selected subset.
  end if
  Let
  $A_{t+1}^a = A_t^a \cup \{i \in H_t : \hat{\mu}_i(t) > \max_{j \in L_t} \hat{\mu}_j(t) + 2g(t,\delta)\}$,
  $A_{t+1}^r = A_t^r \cup \{i \in L_t : \hat{\mu}_i(t) < \min_{j \in H_t} \hat{\mu}_j(t) - 2g(t,\delta)\}$,
  and $A_{t+1} = [K] - A_{t+1}^a - A_{t+1}^r$
end for

Theorem 5. *With probability at least $1 - \delta$, SARWP returns a subset $\hat{S}^*$ of size m within N probes, where $q_{\min}(\hat{S}^*, \mu) \geq q_{\min}([m], \mu) - \varepsilon$ and N satisfies $N \leq \mathcal{G}_{LP}(O, \mathcal{P}) \sum_{i=2}^{K} \hat{M}_{(i)}(\varepsilon,\delta) \mathcal{C}_{\mathrm{LP}}\left(S_{(i)}, 1\right)$.*

The upper bound on the probe complexity is in a similar form to the one for SEWP in Theorem 1, while here the number of samples required for arm i is determined by $\Delta_i^{(\varepsilon,m)}$ instead of Δ_i. This complexity measure matches existing work for the bandit case (Kalyanakrishnan et al.,

2012; Kaufmann & Kalyanakrishnan, 2013). In the bandit case, the upper bound matches the worst case lower bound in Kalyanakrishnan et al. (2012): $\Omega(K\varepsilon^{-2}\log(m/\delta))$, up to logarithmic factors. We do not have a distribution dependent lower bound like Theorem 2 and even in the bandit case a distribution dependent lower bound for $\varepsilon > 0$ is still an open question (Kaufmann & Kalyanakrishnan, 2013).

4.2. *Average* PAC Subset Selection

Next we consider the problem that aims to find a subset whose aggregate regret is ε-optimal. Given a subset $S \subset [K]$ and $|S| = m$, the *aggregate regret* of S is defined as $R_S = \frac{1}{m}\left(\sum_{i\in[m]}\mu_i - \sum_{i\in S}\mu_i\right) = q_{\text{avg}}([m],\mu) - q_{\text{avg}}(S,\mu)$ where $q_{\text{avg}}(S,\mu) = \frac{1}{|S|}\sum_{i\in S}\mu_i$. The aggregate regret of S is said to be ε-optimal iff $R_S \le \varepsilon$.

To address the problem of finding an average ε-optimal subset, Algorithm 4 can still be employed by only modifying the stopping condition according to the different objective. The new stopping condition is described as follows:

Stopping condition for average PAC subset selection: First for each $i \in A_t$, we construct an adversarial estimation $\hat{\mu}'_i(t)$ by setting $\hat{\mu}'_i(t) = \hat{\mu}_i(t) - g(t,\delta)$ if $i \in H_t$ and $\hat{\mu}'_i(t) = \hat{\mu}_i(t) + g(t,\delta)$ if $i \in L_t$. Then we sort the arms in descending order according to $\hat{\mu}'_i(t)$ and let H'_t be the first $m - |A^a_t|$ arms while L'_t be the remaining. The algorithm stops and returns $\hat{S}^* = A^a_t \cup H_t$ if

$$\sum_{i\in H_t\setminus H'_t}(\hat{\mu}_i(t) - g(t,\delta)) \ge \sum_{i\in H'_t\setminus H_t}(\hat{\mu}_i(t) + g(t,\delta)) - m\varepsilon\,.$$

This way of constructing "adversarial estimation" is similar to the one in the CLUCB algorithm of Chen et al. (2014), where the goal is to identify a subset with the highest reward sum without ε relaxation.

The next theorem shows that with the modified stopping condition, Algorithm 4 is admissible and bounds its probe complexity. Define

$$b(m,\varepsilon) = \max\left\{a \in \mathbb{N}^+ : \mu_{m-a+1} - \mu_{m+a} \le \frac{m\varepsilon}{a}\right\}, \tag{9}$$

or $b(m,\varepsilon) = 1$ if $\mu_m - \mu_{m+1} > m\varepsilon$. Then we have the following result:

Theorem 6. *With probability at least $1-\delta$, SARWP with modified stopping condition returns a subset $\hat{S}^*$ of size m within N probes, where $q_{avg}(\hat{S}^*,\mu) \ge q_{avg}([m],\mu) - \varepsilon$ and N satisfies $N \le \mathcal{G}_{LP}(O,\mathcal{P})\sum_{i=2}^{K}\hat{M}_{(i)}(m\varepsilon/b,\delta)\mathcal{C}_{\text{LP}}\left(S_{(i)},1\right)$, where $b = b(m,\varepsilon)$.*

Compared with Theorem 5, the complexity here is measured by $\Delta_i^{(m\varepsilon/b,m)}$ instead. This distribution dependent complexity measure is novel even in the bandit case since the algorithm in Zhou et al. (2014) comes with distribution independent guarantee only. Regarding the worst case performance, since $b(m,\varepsilon) \le \min\{m, K-m\}$, in bandit case our upper bound can be further relaxed to $O\left(\frac{K}{\varepsilon^2}\log\left(\frac{K}{\delta}\log\frac{1}{\varepsilon}\right)\right)$ if $m \le K/2$ and $O\left(\frac{K(K-m)^2}{m^2\varepsilon^2}\log\left(\frac{K}{\delta}\log\frac{K-m}{m\varepsilon}\right)\right)$ if $m > K/2$. Compared with the worst case lower bound in Zhou et al. (2014): $\Omega\left(\frac{K}{\varepsilon^2}\left(1+\frac{\log(1/\delta)}{m}\right)\right)$ for $m \le K/2$ and $\Omega\left(\frac{K-m}{m}\cdot\frac{K}{\varepsilon^2}\left(\frac{K-m}{m}+\frac{\log(1/\delta)}{m}\right)\right)$ for $m > K/2$, although our upper bound does not exactly match this worse case lower bound, our distribution dependent quantity $b(m,\varepsilon)$ shows how the different $\frac{K}{\varepsilon^2}$ and $\frac{K(K-m)^2}{m^2\varepsilon^2}$ terms appear for small m and large m compared with $K/2$.

5. Conclusions

We introduced a generalized version of the best arm identification problem, where a decision maker can query multiple arms at a time. This generalization describes several real world problems that are not adequately modeled by the standard best-arm identification problem. We generalized several existing algorithms and provided distribution dependent upper and lower bounds on the probe complexity, and showed that our algorithms achieve essentially the best possible performance in special cases. In the PAC subset selection problems our complexity measure either matches existing works for the bandit case or provides some new insights. One very interesting question that remains for future work is whether there is a real gap between our lower and upper bounds. However, much work remains to be done: We view our paper as opening a whole new practical and exciting research area of investigating richer feedback structures in "winner selection" problems. Interesting questions include how to change the algorithms when the subsets to be returned are restricted, or when probes are associated with costs.

Acknowledgement

This work was supported by the Alberta Innovates Technology Futures through the Alberta Ingenuity Centre for Machine Learning (AICML) and NSERC.

References

Audibert, J.-Y., Bubeck, S., and Munos, R. Best arm identification in multi-armed bandits. In *Proceedings of the Annual Conference on Learning Theory (COLT)*, 2010.

Bechhofer, R. E. A sequential multiple-decision procedure for selecting the best one of several normal populations with a common unknown variance, and its use with various experimental designs. *Biometrics*, 14(3):408–429, 1958.

Bubeck, S., Munos, R., and Stoltz, G. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852, 2011.

Bubeck, S., Wang, T., and Viswanathan, N. Multiple identifications in multi-armed bandits. In *Proceedings of International Conference on Machine Learning (ICML)*, 2013.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, 2006.

Chen, S., Lin, T., King, I., Lyu, M. R., and Chen, W. Combinatorial pure exploration of multi-armed bandits. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

Even-Dar, E., Mannor, S., and Mansour, Y. PAC bounds for multi-armed bandit and Markov decision processes. In *Proceedings of the Annual Conference on Learning Theory (COLT)*, pp. 255–270, 2002.

Farrell, R. H. Asymptotic behavior of expected sample size in certain one sided tests. *The Annals of Mathematical Statistics*, 35(1):36–72, 1964.

Gabillon, V., Ghavamzadeh, M., and Lazaric, A. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

Grötschel, M., Lovász, L., and Schrijver, A. *Geometric Algorithms and Combinatorial Optimization*. Springer, 2 edition, 1993.

Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. lil' ucb : An optimal exploration algorithm for multi-armed bandits. In *Proceedings of the Annual Conference on Learning Theory (COLT)*, 2014.

Kalyanakrishnan, S. and Stone, P. Efficient selection of multiple bandit arms: Theory and practice. In *Proceedings of International Conference on Machine Learning (ICML)*, 2010.

Kalyanakrishnan, S., Tewari, A., Auer, P., and Stone, P. PAC subset selection in stochastic multi-armed bandits. In *Proceedings of International Conference on Machine Learning (ICML)*, 2012.

Karnin, Z., Koren, T., and Somekh, O. Almost optimal exploration in multi-armed bandits. In *Proceedings of International Conference on Machine Learning (ICML)*, 2013.

Kaufmann, E. and Kalyanakrishnan, S. Information complexity in bandit subset selection. In *Proceedings of the Annual Conference on Learning Theory (COLT)*, 2013.

Kaufmann, E., Cappé, O., and Garivier, A. On the complexity of best arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 2015. (to appear).

Korte, B. H. and Vygen, J. *Combinatorial optimization: theory and algorithms*. Springer, 3 edition, 2006.

Lovász, L. On the ratio of optimal integral and fractional covers. *Discrete mathematics*, 13(4):383–390, 1975.

Mannor, S. and Tsitsiklis, J. N. The sample complexity of exploration in the multi-armed bandit problem. *The Journal of Machine Learning Research*, 5:623–648, 2004.

Paulson, E. A sequential procedure for selecting the population with the largest mean from k normal populations. *The Annals of Mathematical Statistics*, 35(1):174–180, 1964.

Raz, R. and Safra, S. A sub-constant error-probability low-degree test, and a sub-constant error-probability PCP characterization of NP. In *STOC*, pp. 475–484, 1997.

Schrijver, Alexander. *Combinatorial Optimization: Polyhedra and Efficiency*. Springer, 2003.

Slavik, Petr. *Approximation Algorithms for Set Cover and Related Problems*. PhD thesis, State University of New York at Buffalo, 1998. AAI9833643.

Vazirani, V.V. *Approximation algorithms*. Springer, 2001.

Zhou, Y., Chen, X., and Li, J. Optimal pac multiple arm identification with applications to crowdsourcing. In *Proceedings of International Conference on Machine Learning (ICML)*, 2014.