Adaptive Multiple-Arm Identification

Jiecao Chen [*1] **Xi Chen** [*2] **Qin Zhang** [*1] **Yuan Zhou** [*1]

Abstract

We study the problem of selecting K arms with the highest expected rewards in a stochastic n-armed bandit game. This problem has a wide range of applications, e.g., A/B testing, crowdsourcing, simulation optimization. Our goal is to develop a PAC algorithm, which, with probability at least $1-\delta$, identifies a set of K arms with the aggregate regret at most ϵ. The notion of aggregate regret for multiple-arm identification was first introduced in Zhou et al. (2014) , which is defined as the difference of the averaged expected rewards between the selected set of arms and the best K arms. In contrast to Zhou et al. (2014) that only provides instance-independent sample complexity, we introduce a new hardness parameter for characterizing the difficulty of any given instance. We further develop two algorithms and establish the corresponding sample complexity in terms of this hardness parameter. The derived sample complexity can be significantly smaller than state-of-the-art results for a large class of instances and matches the instance-independent lower bound upto a $\log(\epsilon^{-1})$ factor in the worst case. We also prove a lower bound result showing that the extra $\log(\epsilon^{-1})$ is necessary for instance-dependent algorithms using the introduced hardness parameter.

1. Introduction

Given a set of alternatives with different quality, identifying high quality alternatives via a sequential experiment is an important problem in multi-armed bandit (MAB) literature, which is also known as the "pure-exploration" problem. This problem has a wide range of applications. For example, consider the A/B/C testing problem with multiple website designs, where each candidate design corresponds to an alternative. In order to select high-quality designs, an agent could display different designs to website visitors and measure the attractiveness of an design. The question is: how should the agent adaptively select which design to be displayed next so that the high-quality designs can be quickly and accurately identified? For another example, in crowdsourcing, it is critical to identify high-quality workers from a pool of a large number of noisy workers. An effective strategy is testing workers by gold questions, i.e., questions with the known answers provided by domain experts. Since the agent has to pay a fixed monetary reward for each answer from a worker, it is important to implement a cost-effective strategy for to select the top workers with the minimum number of tests. Other applications include simulation optimization, clinical trials, etc.

More formally, we assume that there are n alternative arms, where the i-th arm is associated with an unknown reward distribution $\mathcal{D}_i$ with mean θ_i. For the ease of illustration, we assume each $\mathcal{D}_i$ is supported on $[0,1]$. In practice, it is easy to satisfy this assumption by a proper scaling. For example, the traffic of a website or the correctness of an answer for a crowd worker (which simply takes the value either 0 or 1), can be scaled to $[0,1]$. The mean reward θ_i characterizes the quality of the i-th alternative. The agent sequentially pulls an arm, and upon each pulling of the i-th arm, the *i.i.d.* reward from $\mathcal{D}_i$ is observed. The goal of "top-K arm identification" is to design an adaptive arm pulling strategy so that the top K arms with the largest mean rewards can be identified with the minimum number of trials. In practice, identifying the exact top-K arms usually requires a large number of arm pulls, which could be wasteful. In many applications (e.g., crowdsourcing), it is sufficient to find an "*approximate set*" of top-K arms. To measure the quality of the selected arms, we adopt the notion of *aggregate regret* (or regret for short) from Zhou et al. (2014). In particular, we assume that arms are ordered by their mean $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_n$ so that the set of the best K arms is $\{1,\ldots,K\}$. For the selected arm set T with

[*]Equal contribution [1]Computer Science Department, Indiana University, Bloomington, IN, USA [2]Stern School of Business, New York University, New York, NY, USA. Correspondence to: Jiecao Chen <jiecchen@umail.iu.edu; supported in part by NSF CCF-1525024 and IIS-1633215>, Xi Chen <xchen3@stern.nyu.edu; supported by Google Faculty Research Fellowship>, Qin Zhang <qzhangcs@indiana.edu; supported in part by NSF CCF-1525024 and IIS-1633215>, Yuan Zhou <yzhoucs@indiana.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

the size $|T| = K$, the aggregate regret $\mathcal{R}_T$ is defined as,

$$\mathcal{R}_T = \frac{1}{K}\left(\sum_{i=1}^{K}\theta_i - \sum_{i\in T}\theta_i\right). \tag{1}$$

The set of arms T with the aggregate regret less than a pre-determined tolerance level ϵ (i.e. $\mathcal{R}_T \leq \epsilon$) is called ϵ-top-K arms. In this paper, we consider the ϵ-top-K-arm problem in the "fixed-confidence" setting: given a target confidence level $\delta > 0$, the goal is to find a set of ϵ-top-K arms with the probability at least $1-\delta$. This is also known as the PAC (probably approximately correct) learning setting. We are interested in achieving this goal with as few arm pulls (sample complexity) as possible.

To solve this problem, Zhou et al. (2014) proposed the OptMAI algorithm and established its sample complexity $\Theta\left(\frac{n}{\epsilon^2}\left(1+\frac{\ln\delta^{-1}}{K}\right)\right)$, which is shown to be asymptotically optimal. However, the algorithm and the corresponding sample complexity in Zhou et al. (2014) are *non-adaptive* to the underlying instance. In other words, the algorithm does not utilize the information obtained in known samples to adjust its future sampling strategy; and as a result, the sample complexity only involves the parameters K, n, δ and ϵ but is independent of $\{\theta_i\}_{i=1}^n$. Chen et al. (2014) developed the CLUCB-PAC algorithm and established an instance-dependent sample complexity for a more general class of problems, including the ϵ-top-K arm identification problem as one of the key examples. When applying the CLUCB-PAC algorithm to identify ϵ-top-K arms, the sample complexity becomes $O((\log H^{(0,\epsilon)} + \log\delta^{-1})H^{(0,\epsilon)})$ where $H^{(0,\epsilon)} = \sum_{i=1}^n \min\{(\Delta_i)^{-2}, \epsilon^{-2}\}$, $\Delta_i = \theta_i - \theta_{K+1}$ for $i \leq K$, $\Delta_i = \theta_K - \theta_i$ for $i > K$. The reason why we adopt the notation $H^{(0,\epsilon)}$ will be clear from Section 1.1. However, this bound may be improved for the following two reasons. First, intuitively, the hardness parameter $H^{(0,\epsilon)}$ is the total number of necessary pulls needed for each arm to identify whether it is among the top-K arms or the rest so that the algorithm can decide whether to accept or reject the arm (when the arm's mean is ϵ-close to the boundary between the top-K arms and the rest arms, it can be either selected or rejected). However, in many cases, even if an arm's mean is ϵ-far from the boundary, we may still be vague about the comparison between its mean and the boundary, i.e. either selecting or rejecting the arm satisfies the aggregate regret bound. This may lead to fewer number of pulls and a smaller hardness parameter for the same instance. Second, the worst-case sample complexity for CLUCB-PAC becomes $O((\log n + \log\epsilon^{-1} + \log\delta^{-1})n\epsilon^{-2})$. When δ is a constant, this bound is $\log n$ times more than the best non-adaptive algorithm in Zhou et al. (2014).

In this paper, we explore towards the above two directions and introduce new instance-sensitive algorithms for the problem of identifying ϵ-top-K arms. These algorithms significantly improve the sample complexity by CLUCB-PAC for many common instances and almost match the best non-adaptive algorithm in the worst case.

Specifically, we first introduce a new parameter H to characterize the hardness of a given instance. This new hardness parameter H could be smaller than the hardness parameter $\widetilde{H}$ used in the literature, in many natural instances. For example, we show in Lemma 1 that when $\{\theta_i\}_{i=1}^n$ are sampled from a continuous distribution with bounded probability density function (which is a common assumption in Bayesian MAB and natural for many applications), for $K = \gamma n$ with $\gamma \leq 0.5$, our hardness parameter is $H = O(n/\sqrt{\epsilon})$ while $\widetilde{H} = \Omega(n/\epsilon)$.

Using this new hardness parameter H, we first propose an easy-to-implement algorithm– ADAPTIVETOPK and relate its sample complexity to H. In Theorem 1, we show that ADAPTIVETOPK uses $O\left(\left(\log\log(\epsilon^{-1}) + \log n + \log\delta^{-1}\right)H\right)$ to identify ϵ-top-K arms with probability at least $1-\delta$. Note that this bound has a similar form as the one in Chen et al. (2014), but as mentioned above, we have an $\sqrt{\epsilon}$ -factor improvement in the hardness parameter for those instances where Lemma 1 applies.

We then propose the second algorithm (IMPROVEDTOPK) with even less sample complexity, which removes the $\log n$ factor in the sample complexity. In Theroem 2, we show that the algorithm uses $O\left(\left(\log\epsilon^{-1} + \log\delta^{-1}\right)H\right)$ pulls to identify ϵ-top-K arms with probability $1-\delta$. Since H is always $\Omega(n/\epsilon^2)$ (which will be clear when the H is defined in Section 1.1), the worst-case sample complexity of IMPROVEDTOPK matches the best instance-independent bound shown in Zhou et al. (2014) up to an extra $\log(\epsilon^{-1})$ factor (for constant δ). We are also able to show that this extra $\log(\epsilon^{-1})$ factor is a necessary expense by being instance-adaptive (Theorem 3). It is also noteworthy that as a by-product of establishing IMPROVEDTOPK, we developed an algorithm that approximately identifies the k-th best arm, which may be of independent interest. Details are deferred to the full version of this paper. [1]

We are now ready to introduce our new hardness parameters and summarize the main results in technical details.

1.1. Summary of Main Results

Following the existing literature (see, e.g., Bubeck et al. (2013)), we first define the gap of the i-th arm

$$\Delta_i(K) = \begin{cases}\theta_i - \theta_{K+1} & \text{if } i \leq K\\ \theta_K - \theta_i & \text{if } i \geq K+1.\end{cases} \tag{2}$$

[1] Full version of this paper is available online at `https://arxiv.org/abs/1706.01026`.

Note that when $K = 1$, $\Delta_i(K)$ becomes $\theta_1 - \theta_i$ for all $i \geq 2$ and $\Delta_1(K) = \theta_1 - \theta_2$. When K is clear from the context, we simply use Δ_i for $\Delta_i(K)$. One commonly used hardness parameter for quantifying the sample complexity in the existing literature (see, e.g., Bubeck et al. (2013); Karnin et al. (2013)) is $\widetilde{H} \triangleq \sum_{i=1}^{n} \Delta_i^{-2}$. If there is an extremely small gap Δ_i, the value of $\widetilde{H}$ and thus the corresponding sample complexity can be super large. This hardness parameter is natural when the goal is to identify the exact top-K arms, where a sufficient gap between an arm and the boundary (i.e. θ_K and θ_{K+1}) is necessary. However, in many applications (e.g., finding high-quality workers in crowdsourcing), it is an overkill to select the exact top-K arms. For example, if all the top-M arms with $M > K$ have very close means, then any subset of them of size K forms an ϵ-top-K set in terms of the aggregate regret in (1). Therefore, to quantify the sample complexity when the metric is the aggregate regret, we need to construct a new hardness parameter.

Given K and an error bound ϵ, let us define $t = t(\epsilon, K)$ to be the largest $t \in \{0, 1, 2, \ldots, K-1\}$ such that

$$\Delta_{K-t} \cdot t \leq K\epsilon \qquad \text{and} \qquad \Delta_{K+t+1} \cdot t \leq K\epsilon. \tag{3}$$

Note that $\Delta_{K-t} \cdot t = (\theta_{K-t} - \theta_{K+1}) \cdot t$ upper-bounds the total gap of the t worst arms in the top K arms and $\Delta_{K+t+1} \cdot t = (\theta_K - \theta_{K+t+1}) \cdot t$ upper-bounds the total gap of the t best arms in the non-top-K arms. Intuitively, the definition in (3) means that we can tolerate exchanging at most t best arms in the non-top-K arms with the t worst arms in the top-K arms.

Given $t = t(\epsilon, K)$, we define

$$\Psi_t = \min(\Delta_{K-t}, \Delta_{K+t+1}), \tag{4}$$

and

$$\Psi_t^{\epsilon} = \max(\epsilon, \Psi_t). \tag{5}$$

We now introduce the following parameter to characterize the hardness of a given instance,

$$H = H^{(t,\epsilon)} = \sum_{i=1}^{n} \min\{(\Delta_i)^{-2}, (\Psi_t^{\epsilon})^{-2}\}. \tag{6}$$

It is worthwhile to note that no matter how small the gap Δ_i is, since $\Psi_t^{\epsilon} \geq \epsilon$, we always have $H^{(t,\epsilon)} \leq n\epsilon^{-2}$. We also note that since Ψ_t is non-decreasing in t, $H^{(t,\epsilon)}$ is also non-increasing in t.

Our first result is an easy-to-implement algorithm (see Algorithm 1) that identifies ϵ-top-K arms with sample complexity related to $H^{(t,\epsilon)}$.

Theorem 1 *There is an algorithm that computes ϵ-top-K arms with probability at least $(1-\delta)$, and pulls the arms at most $O\left(\left(\log\log\epsilon^{-1} + \log n + \log\delta^{-1}\right) H^{(t,\epsilon)}\right)$ times.*

We also develop a more sophisticated algorithm with an improved sample complexity, the details of which are deferred to the full version of this paper.

Theorem 2 *There is an algorithm that computes ϵ-top-K arms with probability at least $(1-\delta)$, and pulls the arms at most $O\left(\left(\log\epsilon^{-1} + \log\delta^{-1}\right) H^{(t,\epsilon)}\right)$ times.*

Since $\Psi_t^{\epsilon} \geq \epsilon$ and $H^{(t,\epsilon)} \leq n\epsilon^{-2}$, the worst-case sample complexity by Theorem 2 is $O\left(\frac{n}{\epsilon^2}\left(\log\epsilon^{-1} + \log\delta^{-1}\right)\right)$. While the asymptotically optimal instance-independent sample complexity is $\Theta\left(\frac{n}{\epsilon^2}\left(1 + \frac{\ln\delta^{-1}}{K}\right)\right)$ (by Zhou et al. (2014)), we show that the $\log\epsilon^{-1}$ factor in Theorem 2 is necessary for instance-dependent algorithms using $H^{(t,\epsilon)}$ as a hardness parameter. In particular, we establish the following lower-bound result, the detailed proof of which is deferred to the full version of this paper.

Theorem 3 *For any n, K such that $n = 2K$, and any $\epsilon = \Omega(n^{-1})$, there exists an instance on n arms so that $H^{(t,\epsilon)} = \Theta(n)$ and it requires $\Omega(n\log\epsilon^{-1})$ pulls to identify a set of ϵ-top-K arms with probability at least 0.9.*

Note that since $H^{(t,\epsilon)} = \Theta(n)$ in our lower bound instances, our Theorem 3 shows that the sample complexity has to be at least $\Omega(H^{(t,\epsilon)}\log\epsilon^{-1})$ in these instances. In other words, our lower bound result shows that for any instance-dependent algorithm, and any $\epsilon = \Omega(n^{-1})$, there exists an instance where sample complexity has to be $\Omega(H^{(t,\epsilon)}\log\epsilon^{-1})$. While Theorem 3 shows the necessity of the $\log\epsilon^{-1}$ factor in Theorem 2, it is not a lower bound for *every* instance of the problem.

1.2. Review of and Comparison with Related Works

The problem of identifying the single best arm (i.e. the top-K arms with $K = 1$), has been studied extensively (Even-Dar et al., 2002; Mannor & Tsitsiklis, 2004; Audibert et al., 2010; Gabillon et al., 2011; 2012; Karnin et al., 2013; Jamieson et al., 2014; Kaufmann et al., 2016; Garivier & Kaufmann, 2016; Russo, 2016; Chen et al., 2016b). More specifically, in the special case when $K = 1$, our problem reduces to identifying an ϵ-best arm, i.e. an arm whose expected reward is different from the best arm by an additive error of at most ϵ, with probability at least $(1-\delta)$. For this problem, Even-Dar et al. (2006) showed an algorithm with an instance-independent sample complexity $O\left(\frac{n}{\epsilon^2}\log\delta^{-1}\right)$ (and this was proved to be asymptotically optimal by Mannor & Tsitsiklis (2004)). An instance-dependent algorithm for this problem was given by Bubeck et al. (2013) and an improved algorithm was given by Karnin et al. (2013) with an instance-dependent sample complexity of $O\left(\sum_{i=2}^{n}\max\{\Delta_i,\epsilon\}^{-2}(\log\delta^{-1} + \log\log\max\{\Delta_i,\epsilon\}^{-1})\right)$.

In the worst case, this bound becomes $O\left(\frac{n}{\epsilon^2}(\log\delta^{-1}+\log\log\epsilon^{-1})\right)$, almost matching the instance-independent bound in Even-Dar et al. (2006). When $K=1$, we have $t(\epsilon,K)=0$ and thus $H^{(t,\epsilon)}=H^{(0,\epsilon)}=\Theta\left(\sum_{i=2}^{n}\max\{\Delta_i,\epsilon\}^{-2}\right)$. Therefore, the sample complexity in our Theorem 2 becomes $O((\log\epsilon^{-1}+\log\delta^{-1})H)=O\left(\frac{n}{\epsilon^2}(\log\epsilon^{-1}+\log\delta^{-1})\right)$ in the worst-case, almost matching the bound by Karnin et al. (2013).

For the problem of identifying top-K arms with $K>1$, different notions of ϵ-optimal solution have been proposed. One popular metric is the misidentification probability (MISPROB), i.e., $\mathbf{Pr}(T\neq\{1,\ldots,K\})$. In the PAC setting (i.e., controlling MISPROB less than ϵ with probability at least $1-\delta$), many algorithms have been developed recently, e.g., Bubeck et al. (2013) in the fixed budget setting and Chen et al. (2014) for both fixed confidence and fixed budget settings. Gabillon et al. (2016) further improved the sample complexity in Chen et al. (2014); however the current implementation of their algorithm has an exponential running time. As argued in Zhou et al. (2014), the MISPROB requires to identify the exact top-K arms, which might be too stringent for some applications (e.g., crowdsourcing). The MISPROB requires a certain gap between θ_K and θ_{K+1} to identify the top-K arms, and this requirement is not unnecessary when using the aggregate regret. As shown in Zhou et al. (2014), when the gap of any consecutive pair between θ_i and θ_{i+1} among the first $2K$ arms is $o(1/n)$, the sample complexity has to be huge ($\omega(n^2)$) to make the MISPROB less than ϵ, while any K arms among the first $2K$ form a desirably set of ϵ-top-K arms in terms of aggregate regret. Therefore, we follow Zhou et al. (2014) and adopt the aggregate regret to define the approximate solution in this paper.

Kalyanakrishnan et al. (2012) proposed the so-called EXPLORE-K metric, which requires for each arm i in the selected set T to satisfy $\theta_i\geq\theta_K-\epsilon$, where θ_K is the mean of the K-th best arm. Cao et al. (2015) proposed a more restrictive notion of optimality—ELEMENTWISE-ϵ-OPTIMAL, which requires the mean reward of the i-th best arm in the selected set T be at least $\theta_i-\epsilon$ for $1\leq i\leq K$. It is clear that the ELEMENT WISE-ϵ-OPTIMAL is a stronger guarantee than our ϵ-top-K in regret, while the latter is stronger than EXPLORE-K. Chen et al. (2016a) further extended Cao et al. (2015) to pure exploration problems under matroid constraints. Audibert et al. (2010) and Bubeck et al. (2013) considered expected aggregate regret (i.e., $\frac{1}{K}\left(\sum_{i=1}^{K}\theta_i-\mathbf{E}\left(\sum_{i\in T}\theta_i\right)\right)$, where the expectation is taken over the randomness of the algorithm. Note that this notion of expected aggregate regret is a weaker objective than the aggregate regret.

Moreover, there are some other recent works studying the problem of best-arm identification in different setups, e.g., linear contextual bandit (Soare et al., 2014), batch arm pulls (Jun et al., 2016).

For our ϵ-top-K arm problem, the state-of-the-art instance-dependent sample complexity was given by Chen et al. (2014) (see Section B.2 in Appendix of their paper). More specifically, Chen et al. (2014) proposed CLUCB-PAC algorithms that finds ϵ-top-K arms with probability at least $(1-\delta)$ using $O\left(\left(\log\delta^{-1}+\log H^{(0,\epsilon)}\right)H^{(0,\epsilon)}\right)$ pulls. Since $H^{(0,\epsilon)}\geq H^{(t,\epsilon)}\geq\Omega(n)$ and $H^{(0,\epsilon)}\geq(\Psi_t^{\epsilon})^{-2}$, our Theorem 1 is not worse than the bound in Chen et al. (2014). Indeed, in many common settings, $H^{(t,\epsilon)}$ can be much smaller than $H^{(0,\epsilon)}$ so that Theorem 1 (and therefore Theorem 2) requires much less sample complexity. We explain this argument in more details as follows.

In many real-world applications, it is common to assume the arms θ_i are sampled from a prior distribution $\mathcal{D}$ over $[0,1]$ with cumulative distribution function $F_{\mathcal{D}}(\theta)$. In fact, this is the most fundamental assumption in Bayesian multi-armed bandit literature (e.g., best-arm identification in Bayesian setup (Russo, 2016)). In crowdsourcing applications, Chen et al. (2015) and Abbasi-Yadkori et al. (2015) also made this assumption for modeling workers' accuracy, which correspond to the expected rewards. Under this assumption, it is natural to let θ_i be the $(1-\frac{i}{n})$ quantile of the distribution $\mathcal{D}$, i.e. $F_{\mathcal{D}}^{-1}(1-\frac{i}{n})$. If the prior distribution $\mathcal{D}$'s probability density function $f_{\mathcal{D}}=\frac{\mathrm{d}F_{\mathcal{D}}}{\mathrm{d}\theta}$ has bounded value (a few common examples include uniform distribution over $[0,1]$, Beta distribution, or the truncated Gaussian distribution), the arms' mean rewards $\{\theta_i\}_{i=1}^{n}$ can be characterized by the following property with $c=O(1)$.

Definition 1 *We call a set of n arms $\theta_1\geq\theta_2\geq\cdots\geq\theta_n$ c-spread (for some $c\geq 1$) if for all $i,j\in[n]$ we have $|\theta_i-\theta_j|\in\left[\frac{|i-j|}{cn},\frac{c|i-j|}{n}\right]$.*

The following lemma upper-bounds $H^{(t,\epsilon)}$ for $O(1)$-spread arms, and shows the improvement of our algorithms compared to Chen et al. (2014) on $O(1)$-spread arms. The proof of Lemma 1 is based on simple calculations and is deferred to the full version of this paper.

Lemma 1 *Given a set of n c-spread arms, let $K=\gamma n\leq\frac{n}{2}$. When $c=O(1)$ and $\gamma=\Omega(1)$, we have $H^{(t,\epsilon)}=O(n/\sqrt{\epsilon})$. In contrast, $H^{(0,\epsilon)}=\Omega(n/\epsilon)$ for $O(1)$-spread arms and every $K\in[n]$.*

2. An Instance Dependent Algorithm for ϵ-top-K Arms

In this section, we show Theorem 1 by providing Algorithm 1 and proving the following theorem.

Algorithm 1 ADAPTIVETOPK(n, ϵ, K, δ)

Input: n: number of arms; K and ϵ: parameters in ϵ-top-K arms; δ: error probability
Output: ϵ-top-K arms

Let r denote the current round, initialized to be 0. Let $S_r \subseteq [n]$ denote the set of candidate arms at round r. S_1 is initialized to be $[n]$. Set $A, B \leftarrow \emptyset$
$\Delta \leftarrow 2^{-r}$
while $2 \cdot \Delta \cdot (K - |A|) > \epsilon K$ **do**
  $r \leftarrow r + 1$
  Pull each arm in S_r by $\Delta^{-2} \ln \frac{2nr^2}{\delta}$ times, and let $\widetilde{\theta}_i^r$ be the empirical-mean
  Define $\widetilde{\theta}_a(S_r)$ and $\widetilde{\theta}_b(S_r)$ be the $(K - |A| + 1)$th and $(K - |A|)$th largest empirical-means in S_r, and define

$$\widetilde{\Delta}_i(S_r) = \max\left(\widetilde{\theta}_i^r - \widetilde{\theta}_a(S_r), \widetilde{\theta}_b(S_r) - \widetilde{\theta}_i^r\right) \quad (7)$$

  while $\max_{i \in S_r} \widetilde{\Delta}_i(S_r) > 2 \cdot \Delta$ **do**
    $x \leftarrow \arg\max_{i \in S_r} \widetilde{\Delta}_i(S_r)$
    if $\widetilde{\theta}_x^r > \widetilde{\theta}_a(S_r)$ **then**
      $A \leftarrow A \cup \{x\}$
    else
      $B \leftarrow B \cup \{x\}$
    $S_r \leftarrow S_r \backslash \{x\}$
  $S_{r+1} \leftarrow S_r$
  $\Delta \leftarrow 2^{-r}$
Set A' as the $(K - |A|)$ arms with the largest empirical-means in S_{r+1}
return $A \cup A'$

Theorem 4 *Algorithm 1 computes ϵ-top-K arms with probability at least $1 - \delta$, and pulls the arms at most*

$$O\left(\left(\log\log(\Delta_t^\epsilon)^{-1} + \log\frac{n}{\delta}\right)\sum_{i=1}^{n}\min\{(\Delta_i)^{-2}, (\Delta_t^\epsilon)^{-2}\}\right)$$

times, where $t \in \{0, 1, 2, \ldots, K-1\}$ is the largest integer satisfying $\Delta_{K-t} \cdot t \leq K\epsilon$, and $\Delta_t^\epsilon = \max(\epsilon, \Delta_{K-t})$.

Note that Theorem 4 implies Theorem 1 because of the following reasons: 1) t defined in Theorem 4 is always at least $t(\epsilon, K)$ defined in (3); and 2) $\Delta_t^\epsilon \geq \Psi_t^\epsilon \geq \epsilon$.

Algorithm 1 is similar to the accept-reject types of algorithms in for example Bubeck et al. (2013). The algorithm goes by rounds for $r = 1, 2, 3, \ldots$, and keeps at set of undecided arms $S_r \subseteq [n]$ at Round r. All other arms (in $[n] \setminus S_r$) are either accepted (in A) or rejected (in B). At each round, all undecided arms are pulled by equal number of times. This number is chosen to ensure that the event $\mathcal{E}$, which is defined as the empirical means of all arms are within a small neighborhood of their true means, happens with probability at least $1 - \delta$ (See Definition 2 and Claim 1). Note that $\mathcal{E}$ is defined for all rounds and the length of the neighborhood becomes smaller as the algorithm proceeds. We are able to prove that when $\mathcal{E}$ happens, the algorithm returns the desired set of ϵ-top-K arms and has small query complexity.

To prove the correctness of the algorithm, we first show that when conditioning on $\mathcal{E}$, the algorithm always accepts a top-K arm in A (Lemma 3) and rejects a non-top-K arm in B (Lemma 4). The key observation here is that our algorithm never introduces any regret due to arms in A and B. We then use the key Lemma 5 to upper bound the regret that may be introduced due to the remaining arms. Once this upper bound is not more than ϵK (i.e. the total budget for regret), we can choose the remaining $(K - |A|)$ arms without further samplings. Details about this analysis can be found in Section 2.1.

We analyze of the query complexity of our algorithm in Section 2.2. We establish data-dependent bound by relating the number of pulls to each arms to both their Δ_i's and Δ_{K-t} (Lemma 6 and Lemma 7).

2.1. Correctness of Algorithm 1

We first define an event $\mathcal{E}$ which we will condition on in the rest of the analysis.

Definition 2 *Let $\mathcal{E}$ be the event that $|\widetilde{\theta}_i^r - \theta_i| < 2^{-r}$ for all $r \geq 1$ and $i \in S_r$.*

Claim 1 $\mathbf{Pr}[\mathcal{E}] \geq 1 - \delta$.

Proof: By Hoeffding's inequality, we can show that for any fixed r and i, $\mathbf{Pr}\left[|\widetilde{\theta}_i^r - \theta_i| \geq 2^{-r}\right] \leq 2(\frac{\delta}{2nr^2})^2 \leq \frac{\delta}{2nr^2}$. By a union bound,

$$\mathbf{Pr}[\neg\mathcal{E}] \leq \sum_{r=1}^{\infty}\sum_{i \in S_r}\mathbf{Pr}\left[|\widetilde{\theta}_i^r - \theta_i| \geq 2^{-r}\right] \leq \sum_{r=1}^{\infty}\frac{\delta}{2r^2} \leq \delta.$$

□

The following lemma will be a very useful tool for our analysis, the proof of which is deferred to the full version of this paper.

Lemma 2 *Given $\mu_1 \geq \ldots \geq \mu_n$ and $\Delta > 0$, assuming that $|\widetilde{\mu}_i - \mu_i| \leq \Delta$ for all $i \in [n]$, and letting $y_1 \geq \ldots \geq y_n$ be the sorted version of $\widetilde{\mu}_1, \ldots, \widetilde{\mu}_n$, we have $|y_i - \mu_i| \leq \Delta$ for all $i \in [n]$.*

We now prove that conditioned on $\mathcal{E}$, the algorithm always accepts a desired arm in A.

Lemma 3 *Conditioned on $\mathcal{E}$, during the run of Algorithm 1, $A \subseteq \{1, 2, \ldots, K\}$, that is, all arms in A are among the top-K arms.*

Proof: We prove by induction on the round r. The lemma holds trivially when $r = 0$ ($A = \emptyset$). Now fix a round $r \geq 1$, and let x be the arm that is added to A at Line 10 of Algorithm 1. By the induction hypothesis, assuming that before round r all arms in A are in $[K]$, our goal is to show $x \in [K]$.

By the *inner while* condition we have

$$\widetilde{\theta}^r_x - \widetilde{\theta}_a(S_r) > 2 \cdot 2^{-r}. \tag{8}$$

For any $m \in [K - |A| + 1, |S_r|]$, let j be the arm of the m-th largest true-mean in S_r, and j' be the arm of the m-th largest empirical-mean in S_r. Since $m \geq K - |A| + 1$, we must have $j \notin [K]$ and $\widetilde{\theta}^r_{j'} \leq \widetilde{\theta}_a(S_r)$. By Lemma 2 we also have $|\widetilde{\theta}^r_{j'} - \theta_j| < 2^{-r}$. We thus have

$$\theta_x > \widetilde{\theta}^r_x - 2^{-r} \overset{\text{by (8)}}{>} \widetilde{\theta}_a(S_r) + 2^{-r} > \widetilde{\theta}^r_{j'} + 2^{-r} > \theta_j.$$

That is, *at least* $|S_r| - K + |A|$ arms in S_r have true-means smaller than arm x. On the other hand, $|S_r| - K + |A|$ arms in S_r are not in $[K]$. We therefore conclude that x must be in $[K]$. □

By symmetry, we also have the following lemma, stating that when $\mathcal{E}$ happens, the algorithm always rejects a non-top-K arm in B. We omit the proof because it is almost identical to the proof of Lemma 3.

Lemma 4 *Conditioning on $\mathcal{E}$, during the run of Algorithm 1, $B \subseteq \{K + 1, K + 2, \ldots, n\}$.*

Lemma 5 *Conditioned on $\mathcal{E}$, for all rounds r and $i \in S_r$, it holds that*

$$\widetilde{\theta}^r_i - \widetilde{\theta}_a(S_r) > \theta_i - \theta_{K+1} - 2 \cdot 2^{-r} \tag{9}$$

and

$$\widetilde{\theta}_b(S_r) - \widetilde{\theta}^r_i > \theta_K - \theta_i - 2 \cdot 2^{-r}. \tag{10}$$

Consequently, we have $\widetilde{\Delta}_i(S_r) \geq \Delta_i - 2 \cdot 2^{-r}$ for all rounds r and $i \in S_r$.

Proof: We look at a particular round r. Let j be the arm with $(K - |A| + 1)$-th largest true-mean in S_r. Since by Lemma 3 we have $A \subseteq [K]$, it holds that $j \geq K + 1$. By Lemma 2, we also have $|\widetilde{\theta}_a(S_r) - \theta_j| < 2^{-r}$. We therefore have for any $i \in S_r$

$$\widetilde{\theta}^r_i - \widetilde{\theta}_a(S_r) > \theta_i - \theta_j - 2 \cdot 2^{-r} \geq \theta_i - \theta_{K+1} - 2 \cdot 2^{-r}. \tag{11}$$

With a similar argument (by symmetry and using Lemma 4), we can show that

$$\widetilde{\theta}_b(S_r) - \widetilde{\theta}^r_i > \theta_K - \theta_i - 2 \cdot 2^{-r}. \tag{12}$$

Combining (11), (12) and the definitions of $\widetilde{\Delta}_i(S_r)$ and Δ_i, the lemma follows. □

Now we are ready to prove the correctness of Theorem 4. By Lemma 3, all the arms that we add into the set A at Line 10 are in $[K]$. The rest of our job is to look at the arms in the set A'.

When the algorithm exits the *outer while* loop (at round $r = r^*$) and arrives at Line 16, we have by the condition of the *outer while* loop that

$$2 \cdot 2^{-r^*} \cdot (K - |A|) \leq \epsilon K. \tag{13}$$

Let $m = K - |A|$, and $C = [K] \backslash A = \{i_1, i_2, \ldots, i_m\}$ where $i_1 < i_2 < \ldots < i_m$. Let $\widetilde{\theta}_{j_1} \geq \widetilde{\theta}_{j_2} \geq \ldots \geq \widetilde{\theta}_{j_m}$ be the $(K - |A|)$ empirical-means of the arms that we pick at Line 16. Note that it is *not* necessary that $j_1 < \ldots < j_m$. By Lemma 2 and $\mathcal{E}$, for any $s \in [K - |A|]$, we have $|\widetilde{\theta}_{j_s} - \theta_{i_s}| \leq 2^{-r^*}$ and $|\widetilde{\theta}_{j_s} - \theta_{j_s}| \leq 2^{-r^*}$. By the triangle inequality, it holds that

$$|\theta_{j_s} - \theta_{i_s}| \leq 2 \cdot 2^{-r^*}. \tag{14}$$

We thus can bound the error introduced by arms in A' by

$$\sum_{i \in [K]} \theta_i - \sum_{i \in A \cup A'} \theta_i = \sum_{i \in C} \theta_i - \sum_{i \in A'} \theta_i \overset{\text{by (14)}}{\leq} 2 \cdot 2^{-r^*} \cdot (K - |A|) \overset{\text{by (13)}}{\leq} \epsilon K.$$

2.2. Query Complexity of Algorithm 1

Recall (in the statement of Theorem 4) that $t \in \{0, 1, 2, \ldots, K - 1\}$ is the largest integer satisfying

$$\Delta_{K-t} \cdot t \leq \epsilon K. \tag{15}$$

Lemma 6 *If the algorithm exits the* outer while *loop at round $r = r^*$, then we must have*

$$8 \cdot 2^{-r^*} \geq \Delta_{K-t}. \tag{16}$$

The proof is deferred to the full version of this paper.

Lemma 7 *For any arm i, let r_i be the round where arm i is removed from the candidate set if this ever happens; otherwise set $r_i = r^*$. We must have*

$$8 \cdot 2^{-r_i} \geq \Delta_i. \tag{17}$$

The proof is deferred to the full version of this paper.

With Lemma 6 and Lemma 7, we are ready to analyze the query complexity of the algorithm in Theorem 4. We can bound the number of pulls on each arm i by at most

$$\sum_{j=1}^{r_i} 2^{2j} \cdot \log(2nj^2/\delta) \leq O\left(\log(r_i \cdot n\delta^{-1}) \cdot 2^{2r_i}\right). \tag{18}$$

Now let us upper-bound the RHS of (18). First, if $i \in A$, then by (17) we know that $r_i \leq \log_2 \Delta_i^{-1} + O(1)$. Second, by (16) we have $r_i \leq r^* \leq \log_2 \Delta_{K-t}^{-1} + O(1)$. Third, since $2^{-r^*} \geq \epsilon/2$ (otherwise the algorithm will exit the *outer while* loop), we have $r_i \leq r^* \leq \log_2 \epsilon^{-1} + O(1)$. To summarize, we have $r_i \leq \log_2 \min\{\Delta_i^{-1}, \Delta_{K-t}^{-1}, \epsilon^{-1}\} + O(1) = \log_2 \min\{\Delta_i^{-1}, (\Delta_t^{\epsilon})^{-1}\} + O(1)$ (recall that $\Delta_t^{\epsilon} = \max\{\epsilon, \Delta_{K-t}\}$). We thus can upper-bound the RHS of (18) by $O\left((\log\log(\Delta_t^{\epsilon})^{-1} + \log\frac{n}{\delta}) \cdot \min\{(\Delta_i)^{-2}, (\Delta_t^{\epsilon})^{-2}\}\right)$. The total cost is a summation over all n arms.

Remark 1 *As we stated in Theorem 2, the* $\log n$ *factor from Theorem 1 can be removed. On the other hand, our lower bound result (see Theorem 3) shows that an extra* $\log(\epsilon^{-1})$ *in Theorem 2 is necessary when we make the algorithm adaptive. The proofs of Theorem 2 and Theorem 3 are deferred to the full version of this paper.*

3. Experiments

In this section we present the experimental results. While our theorems are presented in the PAC form, it is in general difficult to verify them directly because the parameter ϵ is merely an upper bound and the actual aggregate regret may deviate from it. In our experiment, we convert our Algorithm 1 to the fixed-budget version (that is, fix the budget of the number of pulls and calculate the aggregate regret). We compare our Algorithm 1 (AdaptiveTopK) with two state-of-the-art methods – OptMAI in Zhou et al. (2014) and CLUCB-PAC in Chen et al. (2014). The comparison between OptMAI /CLUCB-PAC and previous methods (e.g., the methods in Bubeck et al. (2013) and Kalyanakrishnan et al. (2012)) have already been demonstrated in Zhou et al. (2014) and Chen et al. (2014), and thus are omitted due to space constraints. To convert our algorithm to the fixed-budget version, we remove the outer while loop of Algorithm 1. As a replacement, we keep track of the total number of pulls, and stop pulling the arms once the budget is exhausted.

We test our algorithm on both synthetic and real datasets (due to space constraints, our results on real datasets are deferred to the full version of this paper) as described as follows.

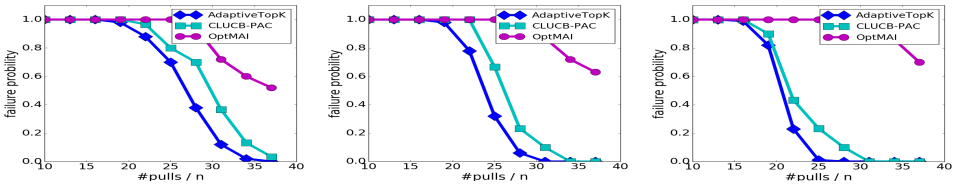
- TWOGROUP: the mean reward for the top K arms is set to 0.7 and that for the rest of the arms is set to 0.3.
- UNIFORM: we set $\theta_i = 1 - \frac{i}{n}$ for $1 \leq i \leq n$.
- SYNTHETIC-p: we set $\theta_i = (1 - \frac{K}{n}) + \frac{K}{n} \cdot (1 - \frac{i}{K})^p$ for each $i \leq K$ and $\theta_i = (1 - \frac{K}{n}) - \frac{n-K}{n} \cdot (\frac{i-K}{n-K})^p$ for each $i > K$. Note that SYNTHETIC-1 is identical to UNIFORM. When p is larger than 1, arms are made closer to the boundary that separates the top-K from the rest (i.e. $1 - \frac{K}{n}$). When p is smaller than 1, arms are made farther to the boundary. We normalize all the arms such that the mean values of the arms still span the whole interval $[0, 1]$. We consider $p = .5, 1, 6$.

We set the total number of arms $n = 1,000$, the tolerance parameter $\epsilon = 0.01$, and vary the parameter K. In AdaptiveTopK and CLUCB-PAC, another parameter δ (i.e., the failure probability) is required and we set $\delta = 0.01$.

For each dataset, we first fix the budget (total number of pulls allowed) and run each algorithm 200 times. For each algorithm, we calculate the empirical probability (over 200 runs) that the aggregate regret of the selected arms is above the tolerance threshold $\epsilon = 0.01$, which is called *failure probability*. A smaller failure probability means better performance. For each dataset and different K, we plot the curve of failure probability by varying the number of pulls. The results are shown in Figure 1-4.

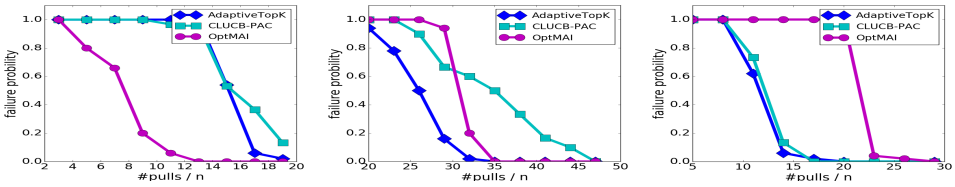
It can be observed from the experimental results that AdaptiveTopK (Algorithm 1) outperforms CLUCB-PAC in almost all the datasets. When K is relatively small, OptMAI has the best performance in most datasets. When K is large, AdaptiveTopK outperforms OptMAI . The details of the experimental results are elaborated as follows.

- For TWOGROUP dataset (see Figure 1), AdaptiveTopK outperforms other algorithms significantly for all values of K. The advantage comes from the adaptivity of our algorithm. In the TWOGROUP dataset, top-K arms are very well separated from the rest. Once our algorithm identifies this situation, it need only a few pulls to classify the arms. In details, the inner while loop (Line 7) of Algorithm 1 make it possible to accept/reject a large number of arms in one round as long as the algorithm is confident.
- As K increases, the advantage of AdaptiveTopK over other algorithms (OptMAI in particular) becomes more significant. This can be explained by the definition of $H^{(t,\epsilon)}$: $t = t(\epsilon, K)$ usually becomes bigger as K grows, leading to a smaller hardness parameter $H^{(t,\epsilon)}$.
- A comparison between SYNTHETIC-.5, UNIFORM, SYNTHETIC-6 reveals that the advantage of AdaptiveTopK over other algorithms (OptMAI in particular) becomes significant in both extreme scenarios, i.e., when arms are very well separated ($p \ll 1$) and when arms are very close to the separation boundary ($p \gg 1$).



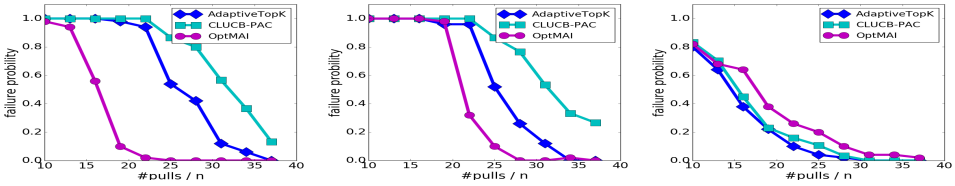
(a) $K = 100$ (b) $K = 250$ (c) $K = 500$

Figure 1: TWOGROUP dataset



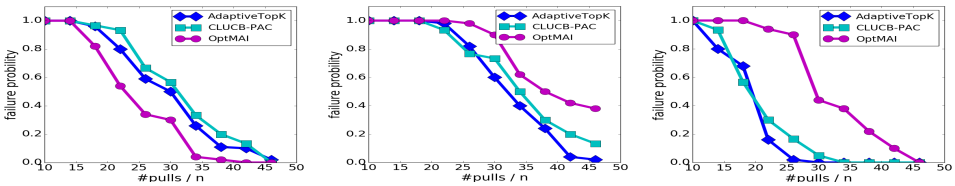
(a) $K = 100$ (b) $K = 250$ (c) $K = 500$

Figure 2: SYNTHETIC-.5 dataset



(a) $K = 100$ (b) $K = 250$ (c) $K = 500$

Figure 3: UNIFORM dataset



(a) $K = 100$ (b) $K = 250$ (c) $K = 500$

Figure 4: SYNTHETIC-6 dataset

4. Conclusion and Future Work

In this paper, we proposed two algorithms for a PAC version of the multiple-arm identification problem in a stochastic multi-armed bandit (MAB) game. We introduced a new hardness parameter for characterizing the difficulty of an instance when using the aggregate regret as the evaluation metric, and established the instance-dependent sample complexity based on this hardness parameter. We also established lower bound results to show the optimality of our algorithm in the worst case. Although we only consider the case when the reward distribution is supported on $[0, 1]$, it is straightforward to extend our results to sub-Gaussian reward distributions.

For future directions, it is worthwhile to consider more general problem of pure exploration of MAB under matroid constraints, which includes the multiple-arm identification as a special case, or other polynomial-time-computable combinatorial constraints such as matchings. It is also interesting to extend the current work to finding top-K arms in a linear contextual bandit framework.

References

Abbasi-Yadkori, Yasin, Bartlett, Peter, Chen, Xi, and Malek, Alan. Large-scale markov decision problems with KL control cost and its application to crowdsourcing. In *Proceedings of International Conference on Machine Learning (ICML)*, 2015.

Audibert, J.Y., Bubeck, S., and Munos, R. Best arm iden-

tification in multi-armed bandits. In *Proceedings of the Conference on Learning Theory (COLT)*, 2010.

Bubeck, Sebastian, Wang, Tengyao, and Viswanathan, Nitin. Multiple identifications in multi-armed bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2013.

Cao, Wei, Li, Jian, Tao, Yufei, and Li, Zhize. On top-k selection in multi-armed bandits and hidden bipartite graphs. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2015.

Chen, Lijie, Gupta, Anupam, and Li, Jian. Pure exploration of multi-armed bandit under matroid constraints. In *Proceedings of the Conference on Learning Theory (COLT)*, 2016a.

Chen, Lijie, Li, Jian, and Qiao, Mingda. Towards instance optimal bounds for best arm identification. arXiv preprint arXiv:1608.06031, 2016b.

Chen, Shouyuan, Lin, Tian, King, Irwin, Lyu, Michael R., and Chen, Wei. Combinatorial pure exploration of multi-armed bandits. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2014.

Chen, Xi, Lin, Qihang, and Zhou, Dengyong. Statistical decision making for optimal budget allocation in crowd labeling. *Journal of Machine Learning Research*, 16:1–46, 2015.

Even-Dar, Eyal, Mannor, Shie, and Mansour, Yishay. PAC bounds for multi-armed bandit and markov decision processes. In *Proceedings of the Annual Conference on Learning Theory (COLT)*, 2002.

Even-Dar, Eyal, Mannor, Shie, and Mansour, Yishay. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7:1079–1105, 2006.

Gabillon, Victor, Ghavamzadeh, Mohammad, Lazaric, Alessandro, and Bubeck, Sébastien. Multi-bandit best arm identification. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2011.

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

Gabillon, Victor, Lazaric, Alessandro, Ghavamzadeh, Mohammad, Ortner, Ronald, and Bartlett, Peter. Improved learning complexity in combinatorial pure exploration bandits. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2016.

Garivier, A. and Kaufmann, E. Optimal best-arm identification with fixed confidence. In *Proceedings of the Conference on Learning Theory (COLT)*, 2016.

Jamieson, Kevin, Malloy, Matthew, Nowak, Robert, and Bubeck, Sébastien. UCB : An optimal exploration algorithm for multi-armed bandits. In *Proceedings of the Conference on Learning Theory (COLT)*, 2014.

Jun, Kwang-Sung, Jamieson, Kevin, Nowak, Robert, and Zhu, Xiaojin. Top arm identification in multi-armed bandits with batch arm pulls. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016.

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

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

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

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

Russo, Daniel. Simple bayesian algorithms for best arm identification. In *Proceedings of the Conference on Learning Theory (COLT)*, 2016.

Soare, Marta, Lazaric, Alessandro, and Munos, Remi. Best-arm identification in linear bandits. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2014.

Zhou, Yuan, Chen, Xi, and Li, Jian. Optimal PAC multiple arm identification with applications to crowdsourcing. In *Proceedings of International Conference on Machine Learning (ICML)*, 2014.