Abstract—This paper introduces the informational multi-armed bandit (IMAB) model, in which at each round, a player chooses an arm, observes a symbol, and receives an unobserved reward in the form of the symbol’s self-information. Thus, the expected reward of an arm is the Shannon entropy of the probability mass function of the source that generates its symbols. The player aims to maximize the expected total reward associated with the entropy values of the arms played. Under the assumption that the alphabet size is known, two UCB-based algorithms are proposed for the IMAB model which consider the biases of the plug-in entropy estimator. The first algorithm optimistically corrects the bias term in the entropy estimation. The second algorithm relies on data-dependent confidence intervals that adapt to sources with small entropy values. Performance guarantees are provided by upper bounding the expected regret of each of the algorithms. Furthermore, in the Bernoulli case, the asymptotic behavior of these algorithms is compared to the Lai-Robbins lower bound for the pseudo regret. Additionally, under the assumption that the exact alphabet size is unknown, and instead the player only knows a loose upper bound on it, a UCB-based algorithm is proposed, in which the player aims to reduce the regret caused by the unknown alphabet size in a finite time regime. Numerical results illustrating the expected regret of the algorithms presented in the paper are provided.

Index Terms—Multi-armed bandits, self-information rewards, entropy estimation, upper confidence bounds, support size estimation.

I. INTRODUCTION

MULTI-ARMED bandit (MAB) problems are sequential decision problems where a player makes iterative decisions in an unfamiliar environment to optimize a total outcome. More specifically, at every round the player is given a choice of $K$ arms, each affiliated with an unknown probability mass function (PMF) for its reward. The player chooses an arm to play based on its previous arm choices and received rewards, and then receives a random reward generated by the chosen arm. The player’s objective is to maximize the total expected reward it receives from all the rounds it has played. If the player knew the expected reward of each arm, it could maximize its total expected reward by repeatedly choosing the arm with the highest expected reward. However, since the player does not know in advance the expected reward of each arm, it must balance two conflicting acts, namely, exploration and exploitation. When making an arm choice, the player wants to exploit the knowledge it accumulated and choose the arm with the highest expected reward, however, naively choosing repeatedly the arm with the highest estimated reward can be sub-optimal since this estimate can be erroneous. To that end, the player periodically dedicates rounds to exploration, aiming to increase the estimation precision of the expected rewards. Balancing the wish to exploit current observations and maximize the immediate reward with the need to explore other arms to increase estimation precision and thus future rewards lies at the heart of MAB decision algorithms; it is known as the exploration-exploitation trade-off.

In the classical MAB problem [2], the reward of an arm is independently and identically distributed over different rounds, and so the expected reward of each arm is the mean of its reward distribution. Furthermore, it can be estimated by the sample mean of the observed rewards which is an unbiased estimator. The classical model has been extended in numerous ways to include, among other models, MAB with linear reward functions [3], [4], [5], [6], Markovian dynamics and rewards [7], [8], [9], [10], [11], [12], and combinatorial bandits with monotone reward functions [13]. In these models, the prevalent measure for the performance of the player’s arm choices is the total expected regret, which measures the cumulative difference between the expected reward of the optimal arm and that of the arms that were sequentially chosen by the player.

In this paper, we consider a different reward structure, which is based on the informativeness of the arm. If we consider each of the $K$ arms as an information source emitting independent and identically distributed (IID) symbols, a player may have the goal of sampling from the source which is most informative, to wit, has maximal entropy. We thus henceforth refer to a MAB problem with entropy rewards as informational MAB (IMAB). We remark in passing that the algorithms and analysis in this paper can be directly adapted to the problem of minimizing the entropy, but for simplicity of exposition we next only consider entropy maximization.

At each round, the player observes a random symbol generated from the PMF of its chosen arm. Letting the true probability of the generated symbol $x$ playing arm $i$ be $p_i(x)$,
the instantaneous reward associated with this symbol is its self-information \(-\log p_i(x)\), where the probability \(p_i(x)\) can capture, for example, the occurrence probability of an event of interest that the player aims to monitor. Using the symbol observations from the previously played arms, the player aims to choose the arm with maximal entropy. Evidently, this model is different from standard MAB since the expected reward of each arm, to wit, its entropy, is a non-linear functional of the PMF, rather than its mean (which is a linear functional). Moreover, at each round \(t\), the instantaneous reward function depends on the probability of a symbol and not its value \(x(t)\). Therefore, the player does not directly observe the instantaneous rewards \(-\log p_i(x(t))\), but can only estimate it based on its previous observations. As a result, and as we next discuss, the IMAB problem is intimately related to the problem of confidence bounds in entropy estimation.

Clearly, there could be different ways to measure this quantity, and in this work we focus on the natural choice of Shannon’s entropy. This is motivated both by the standard interpretation of entropy as a measure of uncertainty, the convenient analytic properties of the entropy functional, and its practical applicability in applications such as gambling [14], data acquisition [16], earthquake prediction [17], video surveillance [18], elderly assisting living [19], [20] and general anomaly detection [21]. Entropy is also used in finance [22], where some applications are portfolio selection with respect to risk, capital increment, diversification, and opinion pricing. Furthermore, analyzing the entropy reward function is related to other notable information-theoretic functionals such as conditional entropy, mutual information, and KL-divergence. These functionals are used to quantify fundamental communication measures such as channel capacity, and rate-distortion functions [23], and can be expressed by addition or subtraction of entropy terms. Thus the techniques we propose in this work can be extended to these functionals. Finally, examples of concrete applications for our model are autonomous exploration [24], [25], coverage [26], and biodiversity in biological systems [27], [28]; next, we elaborate on each example. Robotic networks have a promising role in exploring unknown terrains [24], [25] looking for distinguished points of interest and special features in the terrain. As an illustrative example, consider a planetary rover that can point its camera to one of \(K\) possible directions, and can change its direction at each time point. Then, without any other prior knowledge, a reasonable choice of the next direction that the rover should focus its attention can be based on the entropy of that direction - if the entropy of samples in a direction is very low, then clearly not much could be gained from examining this direction from a closer view. Here, the entropy serves as an unsupervised measure for the informativeness of this direction. The problem of coverage in robotics for monitoring and surveillance in drone networks [26] is also a highly relevant application that can be captured by our model. Here the drones aim to learn the best configuration of drone positioning on a grid, out of \(K\) possible ones, to discover events of interest without knowing in advance their occurrence probability in each part of the grid. Furthermore, entropy is also a measure of biodiversity which is desirable in ecological systems and conservation [27], [28]. Our model can also capture the exploration of learning the best method, out of \(K\) possible ones, to help an ecosystem to recover after devastating events such as fires, floods, and earthquakes where measuring all the members of each species is impossible.

A highly successful algorithmic approach to MAB problems is optimism in the face of uncertainty, where the uncertainty in the reward estimation of each arm is replaced by an optimistic estimate that is based on upper confidence bounds (UCB) [29], [30]. The performance guarantees of the expected regret of UCB algorithms are achieved by utilizing concentration inequalities (see primers in [31], [32]), which bound the probability that the unbiased sampled mean of the reward is outside a chosen distance, known as the confidence interval, of the expected reward of an arm. For the entropy functional, the plug-in estimator of the entropy is well-known to be biased, and in fact, there are no finite variance unbiased estimators of the entropy in general alphabet discrete settings [33]. This paper exploits multiple characteristics of the entropy function to overcome the biasedness of the entropy estimation and the non-linear dependence of the entropy estimation on the symbol samples which excludes the straightforward utilization of the Hoeffding inequality.

Concretely, it is known the bias of the plug-in estimator is upper bounded by \(\log(1 + \frac{|X|}{n})\) where \(|X|\) is the alphabet size and \(n\) is the number of samples used for estimation [33]. Moreover, the entropy functional satisfies a bounded-difference inequality with respect to (w.r.t.) to the samples, and so an application of McDiarmid’s inequality [34] resulted in a concentration inequality bound for the plug-in estimator w.r.t. its (biased) mean. Therefore, our first approach for confidence intervals in entropy estimation is to use a bias-corrected plug-in entropy estimator. Nonetheless, the drawback of this approach is that the additional bias term in the confidence interval leads to a large interval in case the alphabet of the arm is large, even if the entropy of this arm is very low. Therefore, we develop a second type of confidence interval bounds that is based on a total variation bound on the entropy difference of a pair of PMFs [35], [36]. This bound further hinges on a concentration inequality for the total variation, which depends on a functional of the arm’s PMF (denoted \(\zeta(p)\) in what follows), which essentially quantifies the effective alphabet size of the arm (given by \(\zeta(p)/|X|\)). We then show that \(\zeta(p)\) itself can be estimated from the samples at the previous rounds, and this estimate can be used in a UCB algorithm in lieu of the true value. Our upper confidence intervals directly depend on the alphabet size. In practice, the true alphabet size (i.e., support size, the number of letters with strictly positive probability) may not be known to the player in advance and only a loose upper bound on the support size is available (possibly arbitrarily large) Therefore, we additionally propose a UCB algorithm that incorporates support size estimation. For the sake of simplicity of presentation, we derive these bounds for the bias-corrected estimator approach, but it can nonetheless be similarly applied to confidence interval bounds that are based on a total variation bound.
A. Main Contributions and Paper Outline

In Sec. II we formulate the IMAB problem, and in Sec. III we state a generic UCB algorithm for the IMAB problem, which takes a choice of an entropy estimator and a choice of an upper confidence bound (on that estimator) as inputs. In later sections we specify this algorithm for particular choices, and derive regret bounds. First, in Sec. IV we bound the regret of a UCB algorithm that is based on a bias-corrected plug-in estimator, and obtain a regret upper bound which roughly scales as the regret of the standard UCB algorithm (for mean-based rewards), yet only after a large number of rounds $O(\exp(\sqrt{|X|}))$ where here $X$ is the maximal alphabet size of the arms. Then, in Sec. V we present a UCB algorithm that is based on concentration of total variation distance, with the goal of ameliorating this dependence on the alphabet size in case the PMFs of the arms are close to the vertices of the simplex (to wit, there exist a letter whose probability is close to 1). This regime is where elaborated UCB algorithms may lead to improved regret bounds. From a practical point of view, this fits anomaly detection scenarios, in which the arms are mostly “idle”, and thus most of the time emit the high probability symbol, and only occasionally a different symbol (“anomaly”). The player then needs to find the arm which is the “least idle”. The UCB algorithms in this section are based on data-dependent confidence intervals, similarly to UCB-V, which has the merit of adapting the bound to cases in which the entropy of the source is much smaller compared to its maximal value. While our motivation is the large alphabet case, in order to facilitate ideas in a clean way, we first consider Bernoulli arms, for which the probability of the symbol ‘1’ being close to zero indicates that the source is mostly idle. We compare this upper bound to the Lai-Robbins lower bound [2, Thm. 1], which reveals the asymptotic optimality of the proposed UCB algorithm. We then extend the analysis to alphabets of arbitrary size, and show that the regret bound is tight in most regime. We then discuss the gap between the bounds, occurring in the regime of low entropy. Afterward, in Sec. VI we relax the assumption that the alphabet sizes $|X|$ (support sizes) are known to the player, and assume a bound of $\kappa^{-1}$ on the minimal probability in the support. This leads to an upper bound of $\kappa$ on the support size, which may significantly overestimate the true support size. In turn, this leads to an overestimate of the bias of the entropy estimator, then to an overestimate of the confidence interval used by the player, and consequently, to an excessively large regret. To ameliorate this phenomenon, we propose a confidence interval bound that is based on online support size estimation, and analyze the corresponding regret. The proposed algorithm leads to improved regret bound that can be smaller by a factor of $\sqrt{\kappa}$ compared to using the naive bias-corrected plug-in estimator (with $\kappa$ used for the alphabet size). Finally, in Sec. VII we provide a few numerical examples that support the theoretical findings, and in Sec. VIII we summarize the paper and discuss future research directions.

B. Related Work

We conclude the introduction by mentioning related work in the entropy estimation and bandit problem literature. The general problem of entropy estimation is well studied [33], [34], [35], [37], [38], [39], [40], [41], [42], [43], [44]. These papers lead to tight (and even optimal) entropy estimators, and here we build upon their ideas to obtain a confidence interval bound, which is both tailored to the IMAB problem and can also be efficiently estimated from data. In the multi-armed bandit literature, information-theoretic functionals have been used in recent years to decrease the expected regret of several MAB models [45], [46], [47]. Using the mutual information of the probabilities for arm sampling at two consecutive rounds, information-directed sampling (IDS) outperforms both UCB-based algorithms and Thompson sampling [48], [49] in problems with special structures, such as dependent on prior models [45], and bandits with arm-dependent heteroscedastic noise [46]. Nonetheless, these works utilize informational measures to create exploration-exploitation trade-offs, and the reward structure is standard. Extending reward estimation for reward functions whose mean depends on the higher moments, or even the complete knowledge of distribution function is the focus on the works [13], [50]. The work [50] considers MAB problems with a known parametric family of distributions with unknown parameters, however, it is limited to the case where the parameters take discrete values and the support of the distributions is equal to its alphabet. Moreover, no complexity guarantees are provided for the algorithm presented in [50], as a function of the number of discretization points that the values of the parameters can take. A straightforward implementation is of complexity of $O(D + |X|^{-1}) = \Omega(1 + \frac{D}{|X| - 1})$, assuming a uniform discretization over $[0, 1]$ with $D + 1$ points, where for large alphabet size $|X|$ and large values of $D$ (i.e., high precision regime), $1 + \frac{D}{|X| - 1}$ can be approximated by $e^D$. The work [13] relies on stochastically dominant confidence bounds that require a bounded and monotonically increasing instantaneous reward function with respect to the observation $x$, however the self-information $-\log(p_i(x))$ is monotonically decreasing even in $p_i(x)$. Furthermore, it is not bounded in the interval $[0, 1]$. The IMAB problem is also related to bandits with heavy-tail distributions [51], [52], [53], where the reward function may not have a finite moment generating function. These works aim to develop more robust estimators for the mean of IID samples for heavy-tail regimes to achieve logarithmic regret, and do not consider general functions. The work [53] does not rely on such an analysis and instead exploits the KL-divergence as a measure of distance between two distributions. Nonetheless, its algorithm and analysis target the estimation of the mean and cannot be readily adopted for information rewards. We discuss this point further in the summary of this paper.

II. Problem Formulation

We first define a few notation conventions that will be used in the rest of the paper. For $a, b \in \mathbb{R}$, we denote $\max\{a, b\} := a \lor b$ and $\min\{a, b\} := a \land b$, as well as $(t)_+ := t \lor 0$. Furthermore, we denote by $X \cup Y$ the concatenation of the vector $X$ and the vector $Y$. To focus the reader on the first-order terms we denote the linear-times-polynomial function by

$$\Lambda_k(s) := s \log^k s, \quad (1)$$
We further denote by the tightest constants possible.

We have opted for the simplicity of our bounds over obtaining

We remark at this point that in what follows, as customary, we have opted for the simplicity of our bounds over obtaining the tightest constants possible.

Consider the following IMAB problem. Let \( \{X_i\}_{i=1}^K \) be a set of \( K \geq 2 \) memoryless sources, each defined on a possibly different discrete alphabet \( \mathcal{X}_i \), such that \( p_i(x) := \mathbb{P}[X_i = x] \). We further denote by \( p_t = \{p_t(x)\}_{x \in \mathcal{X}_i} \) the full PMF of the \( i \)-th source, and denote its support by \( \mathcal{S}(p_i) := \{x \in \mathcal{X}_i : p_i(x) > 0\} \), that is, \( \mathcal{S}(p_i) \) consists on all the symbols in \( \mathcal{X}_i \) that occur with positive probability. The IMAB problem is a game in which at each round \( t \), the player chooses one of the sources \( i \in [K] := \{1, 2, \ldots, K\} \) and observes the \( i \)-th symbol \( X_i(t) \) from that source. In the context of MAB, each of the sources is referred to as an arm. In this paper, we assume that the random reward associated with this arm choice and this observation is the self-information \( -\log p_t(X_i(t)) \), and so the expected reward of sampling only arm \( i \) is \( \mathbb{E}\log p_t(X_i(t)) = H(p_i) := H_i \), which is the entropy of the \( i \)-th source. The goal of the player is to choose the arm with the maximal expected reward, that is, the maximal entropy, \( i^* = \arg \max_{i \in [K]} H_i \). The player, which does not know in advance the PMFs \( p_t \) (and so also not the entropy values \( H_i \)) estimates the expected reward \( H_i \) of each arm from its previous actions and observations. The first part of the paper, i.e., Sections IV-V assumes that the player knows in advance the alphabet size of the sampled random variables. Later on, in Section VI, the scenario in which the player does not know in advance the exact alphabet size is considered, and instead it only has access to a loose upper bound on it. This models the scenario in which the potential alphabet size is considerably larger than the actual support size.

We denote the arm choice of the player at round \( t \) by \( I(t) \), and we let \( N_i(t) = \sum_{\tau=1}^t \mathbb{1}[I(\tau) = i] \) be the number of times in which arm \( i \) was sampled up to round \( t \). To measure the performance of the policies used by the player, we will adopt the standard expected pseudo-regret [30, Ch. 1]

\[
R(t) := t \cdot H_{i^*} - \sum_{i \in [K]} \mathbb{E}(N_i(t)) \cdot H_i. 
\]

Letting \( \Delta_i := H_{i^*} - H_i \) denote the gap of the \( i \)-th arm, we may equivalently represent the expected pseudo-regret as

\[
R(t) = \sum_{i \in [K]: \Delta_i > 0} \mathbb{E}(N_i(t)) \cdot \Delta_i. 
\]

III. THE UPPER CONFIDENCE BOUND ALGORITHM FOR ENTROPY REWARDS

In this section, we present a generic UCB algorithm for the IMAB problem. Similarly to the UCB algorithm with standard rewards [30, Sec. 2.2] [54, Ch. 1], the algorithm is based on an entropy estimator for which an upper confidence bound is known to hold with high probability. In general, let \( Y := \{Y_t\}_{t \in [n]} \) be \( n \) IID samples from a PMF \( p \) over a finite alphabet \( \mathcal{Y} \). Suppose that there exists an entropy estimator \( \hat{H}(Y, n) \) and an upper confidence deviation (UCD) function \( \text{UCD}(Y, n, \delta) \) (for \( \delta \in (0, 1) \)) for which the upper confidence bound

\[
\hat{H}(p) \leq \hat{H}(Y, n) + \text{UCD}(Y, n, \delta) 
\]

holds with probability larger than \( 1 - \delta \). Note that both the estimator \( \hat{H}(Y, n) \) and the confidence deviation \( \text{UCD}(Y, n, \delta) \) may depend on the observed source symbols \( Y \). The algorithm keeps a set of observed samples from each of the arms up to any round \( t \). Based on the samples of each arm, the algorithm computes the value of the estimator and the upper confidence deviation for each of the arms. The played arm is then the one maximizing the estimated entropy plus the confidence deviation, that is, the right-hand side (RHS) of (4) for the set of observations of each of the arms. The new observed sample is then added to the set of observations of that played arm.

The algorithm takes as input the following:

- The parameters of the information sources, namely, the number of arms \( K \) and the alphabet sizes \( \{\mathcal{X}_i\}_{i \in [K]} \).
- When these exact values are not known in advance, with a slight abuse of notations, we use the input parameters \( \{K_i\}_{i \in [K]} \) such that \( |\mathcal{X}_i| \leq K_i \) for all \( i \in [K] \). These inputs are used in Section VI which considers the unknown alphabet case;
- A sequence of entropy estimators \( \{\hat{H}(\cdot, n)\}_{n \in \mathbb{N}_+} \) and a sequence of upper confidence deviation functions \( \{\text{UCD}(\cdot, n, \delta)\}_{n \in \mathbb{N}_+} \);
- A real confidence parameter \( \alpha > 2 \) and a confidence function \( \delta(t) \equiv \delta_\alpha(t) \) which determines the required reliability of the confidence interval at any round \( t \).

At each round \( t \in \mathbb{N}_+ \), the algorithm plays a chosen arm \( i \in [K] \) and observes the sample \( X_i(t) \). The output of the algorithm is \( \{N_i(t)\}_{i \in [K], t \in \mathbb{N}_+} \), the number of times each of the arm have been played up to each of the rounds \( t \) (or, equivalently, the played arm at each round \( \{I(t)\}_{t \in \mathbb{N}_+} \). Given this output, the pseudo-regret at round \( t \) is given by \( \sum_{i \in [K]: \Delta_i > 0} N_i(t) \Delta_i \) whose expected value is \( R(t) \), as in (3).

The input, the actions and the policy of the player are summarized in Algorithm 1. Therein, \( X_i(t) \) is the set of samples available to the player at round \( t \) from the \( i \)-th arm.

Table I summarizes the entropy estimators \( \hat{H}(\cdot, n) \), upper confidence bounds UCB\((\cdot, n) \), and functions \( \delta_\alpha(t) \) that we develop for Algorithm 1. Additionally, Table I refers to the relevant theorems that provide performance guarantees for the chosen inputs.

IV. UPPER CONFIDENCE BOUNDS WITH BIAS-CORRECTED ENTROPY ESTIMATION

A straightforward idea for estimating entropy is the plug-in estimator, in which the PMF of the source is estimated via the empirical PMF of the samples, and then the entropy of the empirical PMF is used to estimate the entropy of the source. As discussed in the introduction, the plug-in estimator for the entropy concentrates around its expected value [34], yet suffers from a negative bias [33]. Thus, a natural method of
from some distribution specifically, let $Y$ obtaining an upper confidence bound is by correcting this bias.

**Algorithm 1** A General UCB-Entropy Algorithm

1: **procedure** **AN UPPER CONFIDENCE BOUND ALGORITHM** $(K, \{X_i\}_{i \in [K]}, \hat{H}(\cdot, n), UCD(\cdot, \cdot, n), \alpha, \delta_\alpha(t))$

2: set $X_i(0) = \phi$ and $N_i(0) = 0$ for all $i \in [K]$

3: for $t = 1, 2, \ldots$ do

4: play $I(t) \in \arg\max_{i \in [K]} \{\hat{H}(X_i(t-1), N_i(t-1)) + UCD(X_i(t-1), \delta_\alpha(t), N_i(t-1))\}$

5: set $X_{I(t)}(t) = X_{I(t)}(t-1) \cup X_{I(t)}(t)$ and $N_{I(t)}(t) = N_{I(t)}(t-1) + 1$

6: The observation of the chosen arm is concatenated to the sequence of observations

7: end for

8: return $\{N_i(t)\}_{i \in [K]}, t \in \mathbb{N}_+$

9: **end procedure**

obtaining an upper confidence bound is by correcting this bias. Specifically, let $Y = \{Y_t\}_{t \in [n]}$ be a sequence of IID samples from some distribution $p$ over the alphabet $\mathcal{Y}$, and let (with a slight abuse of notation) $\hat{p}(n) = \{\hat{p}(y, n)\}_{y \in \mathcal{Y}}$ be the empirical mean of the $n$ samples, where $\hat{p}(y, n) := \frac{1}{n} \sum_{t=1}^{n} 1 \{Y_t = y\}$ for all $y \in \mathcal{Y}$. Then, the plug-in estimator $H(\hat{p}(n))$ is biased, and, as was proved in [33],

$$H(p) - B(n) \leq \mathbb{E}[H(\hat{p}(n))] \leq H(p),$$

where

$$B(n) := \log \left(1 + \frac{|\mathcal{Y}| - 1}{n}\right)$$

for $n \geq 1$. Therefore, the bias-corrected estimator $H(\hat{p}(n)) + B(n)$ has a nonnegative bias. Let

$$UCD_{\text{bias}}(\delta, n) := B(n) + \sqrt{\frac{2 \log^2(n)}{n} \log \left(\frac{2}{\delta}\right)}.$$ (7)

The concentration result of the plug-in estimator from [34, p. 168] implies the following confidence interval bound. For the sake of clarity of exposition, we present all the relevant proofs for this section in Appendix II.

**Proposition 1:** Let $Y = \{Y_t\}_{t \in [n]}$ be IID from a discrete distribution $p$ over a finite alphabet $\mathcal{Y}$ such that $p(y) := P[Y = y]$. Then, assuming $n \geq 2$, it holds for any $\delta \in (0, 1)$ that

$$|H(\hat{p}(n)) - H(p)| \leq UCD_{\text{bias}}(\delta, n),$$

with probability larger than $1 - \delta$.

We may now specify the general Algorithm 1 to the upper confidence bound of Proposition 1, and obtain the following guarantee on the expected regret. To this end, let us denote

$$\Gamma_{\text{bias}}(\alpha, \beta, \mathcal{Y}, \Delta, t) := \max \left\{ \frac{|\mathcal{Y}|-1}{e^\beta \Delta^2/2 - 1}, 15 \cdot L_2 \left( \frac{8 \cdot \log(2t^n)}{(1 - \beta)^2 \Delta^2} \right) \right\}. \quad (9)$$

**Theorem 2:** Assume that Algorithm 1 is run with a plug-in entropy estimator $H(\hat{p}(n)) = H(\tilde{p}(n))$, and upper confidence deviation $UCD(Y, \delta, n) = UCD_{\text{bias}}(\delta, n)$ with $\delta = \delta_\alpha(t) = t^{-\alpha}$ and $\alpha > 2$. Let $\beta \in (0, 1)$ be given. Then, the pseudo-regret is bounded as

$$R(t) \leq \sum_{i \in [K]: \Delta_i > \Delta} \left( \Gamma_{\text{bias}}(\alpha, \beta, \mathcal{X}_i, \Delta_i, t) \cdot \Delta_i + \frac{2(\alpha - 1)}{\alpha - 2} \cdot \Delta_i \right). \quad (10)$$

---

**TABLE I**

**SUMMARY OF INPUTS FOR ALGORITHM 1 AND RESULTS**

| Thm. | Alphabet | $\hat{H}(Y, n)$ | $\delta_\alpha(t)$ | UCB($Y, \delta, n$) | PMF based UCB | Pseudo regret |
|------|----------|-----------------|---------------------|-------------------|-----------------|---------------|
| Thm. 2 | discrete, finite, known | $H(\hat{p}(n))$ | $t^{-\alpha}$ | (7) | no | (10) |
| Thm. 6 | binary, known | $H(\hat{p}(n))$ | $6t^{-\alpha}$ | (12) | yes | (18) |
| Thm. 9 | binary, known | $H(\hat{p}(n))$ | $4t^{-\alpha}$ | (20) | yes | (24) |
| Thm. 11 | discrete, finite, known | $H(\hat{p}(n))$ | $t^{-\alpha}$ | (28) | yes | (30) |
| Thm. 14 | finite but unknown support | $H(\hat{p}(n))$ | $t^{-\alpha}$ | (51) | no | (54) |
We remark that $\beta$ is a parameter that is used in our analysis of the pseudo regret for the sake of serving tighter upper bounds and is not part of Algorithm 1. The bound on the regret of Algorithm 1 with a bias-corrected entropy estimator in (10) is comprised of a few terms. However, for any $i \in [K]$, there is only a single term that blows-up as $\Delta_i \downarrow 0$, given by
\begin{equation}
\frac{c_1 \cdot \log(t)}{\Delta_i} \cdot \log^2 \left( \frac{c_2 \cdot \log(t)}{\Delta_i^2} \right),
\end{equation}
for some constants $c_1, c_2$. Thus, the regret scales as $O(\log(t)/\Delta)$, where the only difference from the standard UCB regret [30, Thm. 2.1] is the additional poly-logarithmic term. Then, if we consider for simplicity the two-arm case ($K = 2$) with $\Delta_1 = 0$ and $\Delta_2 \equiv \Delta$, since $\Delta t$ is always an upper bound on the pseudo-regret, we may obtain the $O(\frac{\log(t)}{\Delta}) = O(\sqrt{t})$, which roughly matches the gap-independent bound in the standard MAB problem (e.g., [54, Thm. 2.10]).

Nonetheless, from a different perspective, assuming that the gaps are all constants, then if $\log^2(t) = \tilde{O}(|X_i|)$ the regret will be determined by the first term in (9), and so the regret bound is large as long as $t = O(\max_{i\in[K]} \exp(\sqrt{|X_i|}))$. In the next section we develop a UCB algorithm that ameliorates this unfavorable behavior.

V. Upper Confidence Bounds With a Total Variation Bound

As we have seen in Theorem 2, the upper bound on the regret of the UCB algorithm with a bias-corrected entropy estimator is severely affected by the size of the alphabets $X_i$. Therefore, a natural question is whether improved bounds can be obtained whenever the entropy of sources is much less than the alphabet size. In this section, we propose algorithms that adapt to arms with very low entropy. The idea is similar to UCB-V [55], [56] that replaces the distribution-independent confidence interval of the standard UCB algorithm (which hinges, e.g., on Hoeffding’s inequality, assuming bounded rewards), with a distribution-dependent confidence interval (which hinges, e.g., on Bernstein’s inequality). Our next proposed algorithms will similarly use a data-dependent UCD. For such algorithms, the confidence interval, which in principle depends on the unknown distribution, is also required to be estimated from the given observations. For the sake of illustration, let us first consider the simpler case of Bernoulli arms, for which $p_i(1) = \mathbb{P}[X_i = 1]$ is close to 0 for some arm $i \in [K]$. The entropy of this arm is much smaller than the maximal possible value of $\log |X_i| = \log 2$. A multiplicative Chernoff’s inequality (or Bernstein’s inequality, see Lemma 17) results in a confidence interval of $O\left(\frac{p_i(1) \log(1/\delta)}{n}\right)$ in the estimation of $p_i(1)$ using $n$ samples from the source. Since $p_i(1) \ll 1$, this is much smaller than the $O\left(\frac{\log(1/\delta)}{n}\right)$ which stems from standard Chernoff’s bound (or Hoeffding’s inequality). This confidence interval on $p_i(1)$ then leads to an improved confidence interval bound on the error of the plug-in estimator of the entropy. Since this confidence interval bound depends on the unknown $p_i(1)$, it should also be estimated by the player, using its estimation of $p_i(1)$. The estimation error of the confidence interval is then another source of error that is addressed by our analysis.

Thus, in what follows, we begin with the Bernoulli case in Sec. V-A, which leads to a more transparent bound than the general case, and compare it in detail to the Lai-Robbins impossibility result [2, Thm. 1]. We later on generalize this type of algorithm to arbitrary arm alphabets in Sec. V-B. As in the Bernoulli case, the confidence interval of arms with low entropy is smaller than for arms with large entropy. The PMF of these arms is close to the vertices of the probability simplex, which in the Bernoulli case implies a low value of $p_i(1)$. In the general alphabet case, the value of $p_i(1)$ is replaced by a functional $\zeta(p) \in [0,1]$, which satisfies that low values of $\zeta(p)$ are indicator of being close to the vertices of the simplex, and that can also be efficiently be estimated from the data. As a result, we additionally show that the effective alphabet size for the IMAB problem is $\zeta_i |X_i|$, which demonstrates the utility of this functional.

A. The Bernoulli Case

In this section, we consider the Bernoulli case, in which $X_i = \{0, 1\}$ for all the $K$ arms, $i \in [K]$. For brevity we use $h_0(p) := -p \log p - (1-p) \log(1-p)$ to denote the binary entropy function. Furthermore, we assume for simplicity of exposition that $p_i(1) = \mathbb{P}[X_i = 1] \leq 1/2$ for all $i \in [K]$. The results can be extended in a straightforward manner to remove this assumption. The proofs of the theoretical results presented in Section V-A are included in Appendix III. The proposed UCB algorithm and its regret analysis are based on the following confidence deviation function
\begin{equation}
\text{UCD}_{\text{ber}}(q, \delta, n) := \sqrt{\frac{12 q \log\left(\frac{q}{\delta}\right)}{n} \log\left(\frac{n}{q \log\left(\frac{q}{\delta}\right)}\right)} + \frac{18 \log\left(\frac{q}{\delta}\right) \log(n)}{n},
\end{equation}
and the corresponding confidence interval bound for the plug-in entropy estimator:

Proposition 3: Let $Y_i = \{Y_{i,e}\}_{e \in [n]}$ be IID from a Bernoulli distribution with parameter $p = \mathbb{P}[Y_i = 1]$, and let $\hat{p}(n) = \frac{1}{n} \sum_{e=1}^{n} \mathbb{1}\{Y_{i,e} = 1\}$ be the empirical probability of ‘1’. Let $\delta \in [0, \frac{1}{2}]$ be given. If $n \geq 200 \cdot \log\left(\frac{1}{\delta}\right)$ then
\begin{equation}
|h_0(\hat{p}(n)) - h_0(p)| \leq \text{UCD}_{\text{ber}}(\hat{p}(n), \delta, n),
\end{equation}
with probability larger than $1 - \delta$.

Remark 4: The confidence interval bound of Proposition 3 follows from the relation
\begin{equation}
|h(p) - H(q)| \leq d_{TV}(p, q) \log\left(\frac{|\mathcal{Y}|}{d_{TV}(p, q)}\right),
\end{equation}
where $d_{TV}(p, q)$ is the total variation distance between PMFs $p$ and $q$ defined on a common alphabet $\mathcal{Y}$, and that holds as long as $d_{TV}(p, q) \leq \frac{1}{2}$ [57, Lemma 2.7]. Specifically, this bound is used with $p$ being the true PMF of the arm and $q$ being the empirical PMF of the arm after its sampling by the player. Since the total variation between the true PMF

\footnote{We define this functional explicitly in (26).}
and the empirical PMF decreases when an arm is sampled, the relation (14) implies that the entropy of the empirical arm PMF approaches to the true entropy of the arm PMF. Nonetheless, the bound (14) is not the sharpest known bound, and, e.g., it also holds that [35, Thm. 6]

\[ |H(p) - H(q)| \leq d_{TV}(p, q) \log(|Y| - 1) + h_b(d_{TV}(p, q)), \]

(15)

(see also an additional refinement in [36]). In our proofs this type of bounds is utilized in the regime \( d_{TV}(p, q) = O(1) \), for which both (14) and (15) are of the same order of \( \Theta(\frac{|Y|}{d_{TV}(p, q)}) \). Thus, we exclusively use the simpler bound (14).

Remark 5: In the special case of a binary alphabet, the binary entropy \( h_b(p) \) is maximized when \( p = 1/2 \). Additionally, it is symmetric around \( p = 1/2 \) and strictly increases in the interval \([0, 1/2] \). Therefore, instead of using UCB for entropy estimation we can consider the following strategy where we optimistically look for the arm with the minimal \( p \)-value of \( X \), which we can extend to the general alphabet case. These regret bounds are achieved by UCB strategies for entropy estimation.

Next, we state the regret bound on Algorithm 1, based on the confidence interval of Proposition 3. To this end, let us denote

\[ \Gamma_{\text{ber}}(\alpha, \beta, q, \Delta, t) := \max \left\{ 6 \cdot \Delta \left( \frac{36\alpha \log(t)}{(T - \beta)\Delta} \right), \frac{5120\alpha \log(t)}{\beta^2\Delta^2} \cdot \log^2 \left( \frac{48}{\beta^2\Delta^2} \right), \frac{88\sqrt{\alpha \log(t)}}{\beta\Delta} \cdot \log \left( \frac{48}{\beta^2\Delta^2} \right) \right\}, \]

(16)

and use the notation \( \hat{p}(Y, n) \) for the empirical probability of \( '1' \) in \( Y = \{Y_t\}_{t \in [n]} \).

Theorem 6: Assume that \( X_i = \{0, 1\} \) for all \( i \in [K] \). Further assume that Algorithm 1 is run with the plug-in entropy estimator \( \hat{H}(Y, n) \equiv H(\hat{p}(Y, n)) \) and upper confidence deviation \( UCD(Y, \delta, n) \equiv UCD_{\text{ber}}(\hat{p}(Y, n), \delta, n) \),

(17)

(as defined in (12)) with \( \delta \equiv \delta_0(t) = 6t^{-\alpha} \) with \( \alpha > 2 \). Then,

\[ R(t) \leq \sum_{i \in [K]} \inf_{\hat{p}(Y, n)} \left[ \Gamma_{\text{ber}}(\alpha, \beta, p_i(1), \Delta_i, t) \cdot \Delta_i + \frac{16(\alpha - 1)}{\alpha - 2} \cdot \Delta_i \right]. \]

(18)

Let us inspect the regret bound of (18) of Theorem 6 in the regime of small gaps. By inserting the definition of the \( \Gamma_{\text{ber}}(\cdot) \), the dominating term as \( \Delta_i \rightarrow 0 \) is on the order of \( O(\frac{p_i \log(t)}{\Delta_i}) \) and all other terms are \( O(\log^2(\frac{1}{\Delta_i})) \). Thus, e.g., in case \( \Delta_i = \Theta(p_i) \) (as \( t \rightarrow \infty \)), then the regret is only \( O(\log(t) \cdot \log^2(\frac{1}{\Delta})) \). This is a similar behavior to the UCB-V algorithm [55], [56] with standard bounded rewards. Nonetheless, in general, the bound of Theorem 6 is sub-optimal. Indeed, recall that in the standard Bernoulli MAB problem, say with two arms \( (K = 2) \), the regret depends on the difference \( p_1(1) - p_2(1) \) between the \( '1' \)-probability of each arm, which is exactly the gap between their rewards. However, in the IMAB problem, the gap is \( h_b(p_1(1)) - h_b(p_2(1)) \), and due to the non-linearity of the entropy functional this gap depends on both the difference \( p_1(1) - p_2(1) \) as well as the location of \( p_1(1) \).

To further elucidate this phenomenon, next we state the Lai-Robbins lower bound [2] on the pseudo-regret. We follow the clear statement made in [54, Thms. 2.14 and 2.16]:

Theorem 7 (Lai-Robbins lower bound): Consider the IMAB problem with \( K \) arms. A problem instance \( I \) is the collection \( \{p_i\}_{i \in [K]} \) with \( p_i \). Suppose that an IMAB algorithm is such that \( R(t) = O(C_{I,a} t^a) \) for each problem instance \( I \) and \( a > 0 \). Fix an arbitrary problem instance \( I \). Then,

\[ \liminf_{t \to \infty} \frac{R(t)}{\log(t)} \geq \sum_{i \in [K]} \Delta_i \cdot \frac{D_{KL}(p_i || p_i^*) \Delta_i}{\Lambda}, \]

(19)

where \( D_{KL}(|\cdot| |\cdot|) := \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \) is the Kullback-Leibler divergence between the probability measures \( p \ll q \), and where \( \Delta_i = \max_{x \in [K]} h_b(p_j) - h_b(p_i) \).

The proof of Theorem 7 for the IMAB problem is omitted since it is essentially identical to the proof of the standard Lai-Robbins lower bound for Bernoulli bandits, which can be found in [2, Thm. 1] [30, Thm. 2.2] [54, Ch. 2].

Strictly speaking, Theorem 7 is asymptotic and does not specify the minimal \( t \) required for its validity, and is also valid for a fixed set of gaps. Nonetheless, we next informally compare the order of convergence it implies with the one attained in Theorem 6 while considering the effect of varying the gap. To simplify the next discussion, we will assume \( K = 2 \) with \( p_2 < p_1 \). Let \( \Delta \equiv \Delta_2 = h_b(p_1) - h_b(p_2) \). In the standard reward case, if \( p_2 \) is bounded away from zero, then \( D_{KL}(p_2 || p_1) = \Theta(\Delta^2) \) (e.g., using Pinsker’s inequality and its reversed version), and then the lower bound is \( \Omega(\frac{1}{\log(t)}) \). This lower bound is roughly achieved by the basic UCB algorithm [30, Thm. 2.1], and the UCB-V algorithm leads to data-dependent improvement in the constant. If \( p_1 = o(1) \) then \( D_{KL}(p_2 || p_1) = \Omega(\Delta^2) \) and then the lower bound is \( \Omega(\frac{1}{\log(t)}) \). This is achieved by the KL-UCB algorithm [58]. We next move on to compare the regret bound of Theorem 6.

Before making this comparison, we note that the upper bound has an extra multiplicative logarithmic factor, which can be as large as \( \Theta((\log(t))^2) \). To focus on the first-order terms in the regret bound, we next ignore these additional factors in the discussion. We next consider a few different regimes.

To begin, let us assume that \( p_1 = p \) and \( p_2 = p - \Delta \) with \( p \) fixed and \( \Delta \downarrow 0 \), then \( \Delta = h_0(p_1) - h_0(p_2) = \Theta(\Delta^2) \) and the ratio in the lower bound is \( \Theta(\frac{\Delta^2}{\Delta}) \) as in standard bandits. This roughly matches the upper bound of Theorem 6 on the pseudo-regret achieved by the algorithm, and no significant improvements are anticipated.

\(^2\)Note that for binary alphabets, if \( q(1) = 0 \) or \( q(1) = 1 \) and \( p(1) \neq q(1) \) then \( D_{KL}(p || q) = \infty \).
Next, we consider the regime in which the probabilities of
the arms are close to 1/2. The binary entropy function
“flattens” in this region, and is markedly different from the
standard linear reward function. Thus, on an intuitive level, this
is not a difficult instance of the problem. More explicitly, let
us assume that both \( p_1 = \frac{1}{2} - \alpha \) and \( p_2 = \frac{1}{2} - \Delta \). Then, both
\( \Delta = h_b(p_1) - h_b(p_2) = \Theta(\alpha^2) \) and \( \Delta = h_b(p_1) - h_b(p_2) = \Theta(\alpha^2) \),
and the ratio in the lower bound is asymptotically \( \Theta(\log t) \)
even if \( \alpha \downarrow 0 \) and so also \( \Delta \downarrow 0 \). In this regime, the
regret of Algorithm 1 is upper bounded as
\[
\Theta(\log(\frac{\delta}{\alpha}))
\]
where the regime of current interest
is not a difficult instance of the problem. More explicitly, let
we now have the following confidence interval bound.
Specifically, consider the following upper confidence deviation
\[
\text{UCD}^{(1/2)}(q, \delta, n) := \sqrt{\frac{1}{2} - q} \cdot \sqrt{\log(\frac{\delta}{n})} + \frac{9 \log(\frac{\delta}{n})}{n}. \tag{20}
\]
We now have the following confidence interval bound.

**Proposition 8:** Let \( Y = \{Y_i\}_{i \in [n]} \) be IID from a Bernoulli with parameter \( p \) and \( p \) be the empirical probability of \( '1' \).
Assume that \( p > \frac{1}{2} \), that \( n \geq 60 \log(\frac{\delta}{\alpha}) \), and let \( \delta \in [0, \frac{1}{2}] \) be given. Then
\[
|h_b(p(n)) - h_b(p)| \leq \text{UCD}^{(1/2)}(p, \delta, n), \tag{21}
\]
with probability larger than \( 1 - \delta \), and
\[
|h_b(p(n)) - h_b(p)| \leq \text{UCD}^{(1/2)}(p, \delta, n), \tag{22}
\]
with probability larger than \( 1 - \delta \).

We restricted \( p_i \) to \( \left[ \frac{1}{2}, \frac{3}{4} \right] \) since the regime of current interest
is such that the arm probabilities are close to 1/2, and this restriction simplifies the exposition. With this result, the following regret bound can be easily derived using the same methods used in the proof of Theorem 6, and so its proof is omitted. For simplicity, we only state it for \( K = 2 \) arms.

**Theorem 9:** Assume that \( X_i = \{0, 1\} \) and that \( p_i \in \left[ \frac{1}{2}, \frac{3}{4} \right] \)
for \( i \in \{1, 2\} \) where \( \Delta = h_b(p_1) - h_b(p_2) \) with \( p_2 < p_1 < \frac{1}{2} \). Further assume that Algorithm 1 is run with the plug-in entropy estimator \( \hat{H}(Y, n) = H(p(Y, n)) \) and upper
confidence deviation
\[
\text{UCD}(Y, \delta, n) = \text{UCD}^{(1/2)}(p(Y, n), \delta, n), \tag{23}
\]
(as defined in (20)) with \( \delta \equiv \delta_\alpha(t) = 4t^{-\alpha} \) with \( \alpha > 2 \). Then,
\[
R(t) \leq \frac{784 \left( \frac{1}{2} - p_2 \right)^2 \alpha \log(t)}{\Delta} + 60 \alpha \log(t) + \frac{8(\alpha - 1)}{\alpha - 2} \cdot \Delta. \tag{24}
\]

In the regime above, \( \left( \frac{1}{2} - p_2 \right)^2 = \alpha^2 = \Theta(\Delta) \) and so the
regret of the algorithm is \( \Theta(\log(t)) \), just as the Lai-Robbins lower bound. This result can be easily extended to multiple arms, and can also be combined with the result of Theorem 6 by taking the minimal confidence bound out of those used in Theorem 6 and that of Theorem 9. This will result in the minimum of the regret upper bounds of both theorems.

Finally, we consider the other extremal regime for the
arms’ Bernoulli probabilities, to wit, the small value regime,
in which the derivative of the binary entropy function is
ubounded. For concreteness, assume that \( p_1 = \gamma > 0 \) and \( p_2 = 0 \) for a small \( \gamma > 0 \). It then can be easily derived that the gap is \( \Delta = h_b(p_1) - h_b(p_2) = \Theta(\gamma \log \frac{1}{\gamma}) \) and that \( D_{KL}(p_2||p_1) = \Theta(\gamma) \) (adopting the convention that
\( 10 \log(0/\gamma) = 0 \) in the definition of the KL divergence, which agrees with continuity assumptions). So, the lower bound is
\( \Theta(\log \frac{1}{\gamma} \cdot \log t) \). According to Theorem 6, the pseudo-regret
bound achieved by the algorithm is also \( O(\log \frac{1}{\gamma} \cdot \log t) \), and this agrees with the lower bound.

We may also compare the gap-independent lower bound with the gap-independent regret bound that Theorem 6 implies.
This bound is given by
\[
R(t) = O \left( \min \left\{ \log t \cdot \log \left( \frac{1}{\gamma} \right), \left( \gamma \log \frac{1}{\gamma} \right) \cdot t \right\} \right)
\leq O \left( \log^2(t) \right), \tag{25}
\]
where the maximum regret is achieved for the gap \( \gamma = \Theta(\log(\frac{1}{\gamma})) \) (as can be shown by equating both terms and eval-
uating the order of the solution). This result roughly matches the gap-independent lower bound of \( R(t) = O(\log(t)) \), which can be established by modifying, e.g., the argument in [54, Sec. 2.3 and 2.4]. We next briefly describe this argument.

Therein, the standard MAB problem is reduced to a best-arm identification problem (essentially, a binary hypothesis testing problem), between a uniform Bernoulli source \( p_1(1) = \frac{1}{2} \) and an \( \epsilon \)-biased source, that is, an arm for which \( p_1(1) = \frac{1}{2} - \epsilon \) for some \( \epsilon > 0 \). Since the KL divergence is \( \Theta(\epsilon^2) \), then as long as \( t = \Theta(\epsilon^{-2}) \), the arm identifier resulting from any MAB algorithm will have a constant fraction of errors. Consequently, since the gap in this standard MAB problem is \( \Theta(\epsilon) \), a lower bound on the pseudo-regret is given by \( \Theta(\epsilon t) \) which can be chosen as large as \( R(t) = \Omega(\sqrt{t}) \), to obtain the gap-independent bound. In the IMAB problem and the regime considered here, the KL divergence is on the order of \( \Theta(\gamma) \) (instead of \( \Theta(\gamma^2) \)), and the gap is \( \Theta(\gamma \log(\frac{1}{\gamma})) \). Repeating the same argument then leads to a lower bound of \( R(t) = O(\log(t)) \).

Intuitively, this is also not a difficult instance of the problem (just as in the regime in which the probabilities are close to half) since here it is easy to statistically distinguish between the arms with the larger entropy (the KL divergence between the distributions is linear \( D_{KL}(p_1||p_2(1)) = \Theta(\gamma) \) rather than quadratic, while the gap is only logarithmically above linear \( \Theta(\gamma \log(1/\gamma)) \)).

**B. The General Alphabet Case**

In this section, we extend the data-dependent UCD bound of the previous section to general alphabets, of cardinality larger than 2. To this end, let \( p \) and \( q \) be two probability mass functions over an alphabet \( Y \). We consider the distribution-dependent functional
\[
\zeta(p) := 1 - \sum_{y \in Y} p^2(y), \tag{26}
\]
which can be easily seen to equal \( \zeta(p) = 1 - e^{-H^2(p)} \), where \( H^2(p) \) is the second-order Rényi entropy. Note that if \( H^2(p) \ll 1 \) then \( \zeta(p) \approx H^2(p) \ll 1 \). As we shall
see, $\zeta(p_i)|X_i|$ is a measure of the effective alphabet size of the $i$th arm. In addition, $\zeta(p_i)$ can also be accurately estimated from the data, and thus can be used by the player in determining its confidence interval. The proofs of the theoretical results of this Section V-B are relegated to Appendix IV.

Let the plug-in estimator of $\zeta(p)$ be given by

$$\hat{\zeta}(n) = \hat{\zeta}(Y, n) := 1 - \sum_{y \in \mathcal{Y}} \hat{p}^2(n, y).$$  (27)

The proposed UCB algorithm and its regret analysis are based on the following confidence interval function

$$\text{UCD}_n(\hat{\zeta}, \mathcal{Y}, \delta, n) := 3\sqrt{\frac{\zeta(\mathcal{Y})}{n} \log \left( \frac{n|\mathcal{Y}|}{36\zeta} \right)} + \frac{3}{2} \sqrt{\frac{\log \left( \frac{3}{\delta} \right)}{n}} \left( \frac{n|\mathcal{Y}|^2}{9} \right) + 2|\mathcal{Y}|^{1/2} \log^{1/4}(\frac{3}{\delta}) \log (n|\mathcal{Y}|^{2/3}) \cdot n^{3/4},$$  (28)

and the following confidence interval bound for the plug-in entropy estimator.

**Proposition 10:** Let $Y = \{Y_\ell\}_{\ell \in [n]}$ be IID from a PMF $p$ over an alphabet $\mathcal{Y}$, and let $\hat{p}(n) = \{\hat{p}(n, y)\}_{y \in \mathcal{Y}}$ with $\hat{p}(n, y) = \frac{1}{n} \sum_{\ell=1}^{n} 1\{Y_\ell = y\}$ be the empirical PMF of $Y$. Let $\delta \in [0, 0.2]$ be given. Then, if $n \geq 112 \cdot \log \left( \frac{3}{\delta} \right)$ it holds that

$$|H(\hat{p}(n)) - H(p)| \leq \text{UCD}_n(\hat{\zeta}(n), \mathcal{Y}, \delta, n),$$  (29)

with probability larger than $1 - \delta$.

To state the upper bound on the regret, we define, as before

$$\Gamma_n(\alpha, \zeta(p_i), \Delta_i, t) = \max \left\{ 288 \frac{\zeta(\mathcal{Y})}{|\mathcal{Y}|^2} \Lambda_1^2 \left( \frac{2|\mathcal{Y}|}{3\Lambda_i} \right), \right.$$

$$36230 \frac{\alpha^{1/3} \log^{1/3}(t)}{|\mathcal{Y}|^{2/3}} \Lambda_i^{1/3} \left( \frac{2|\mathcal{Y}|}{3\Lambda_i} \right), \right.$$

$$\frac{135}{|\mathcal{Y}|^2} \Lambda_2 \left( \frac{9|\mathcal{Y}|^2 \alpha \log(t)}{\Delta_i^2} \right), \right.$$

$$\frac{3}{|\mathcal{Y}|^{2/3}} \Lambda_i^{1/3} \left( \frac{27|\mathcal{Y}|^{4/3} \alpha^{1/3} \log^{1/3}(t)}{\Delta_i^{4/3}} \right), \right.$$

$$30 \cdot \alpha \log \left( \frac{2^{1/n} t}{n} \right), 119 \zeta(p_i)|\mathcal{Y}|. \}$$  (30)

**Theorem 11:** Assume that Algorithm 1 is run with the plug-in entropy estimator

$$\hat{H}(Y, n) = H(\hat{p}(Y, n)),$$  (31)

and upper confidence deviation

$$\text{UCD}(\hat{\zeta}, \mathcal{Y}, \delta, n) = \text{UCD}_n(\hat{\zeta}(Y, n), \mathcal{Y}, \delta, n),$$  (32)

with $\delta = \delta_0(t) = t^{-\alpha}$ with $\alpha > 2$. Then,

$$R(t) \leq \sum_{i \in [K]: \Delta_i > 0} \left[ \Gamma_n(\alpha, \zeta(p_i), \Delta_i, t) \cdot \Delta_i + \frac{4(\alpha - 1)}{\alpha - 2} \cdot \Delta_i \right].$$  (33)

To show the improvement of using $\text{UCD}_n(\zeta, \mathcal{Y}, \delta, n)$ of (28) over $\text{UCD}_{\text{bias}}(\delta, n)$ of (7), we observe that for a non-asymptotic time $t$ that is upper bounded by a polynomial function of $|X_i|$, the contribution of a sub-optimal arm $i$ to the regret bound of Theorem 11 scales as

$$\hat{O} \left( \frac{\zeta(p_i)|X_i| + 1}{\Delta_i} + \frac{|X_i|^{2/3}}{\Delta_i^{1/3} + \zeta(p_i)|X_i| + \Delta_i} \right),$$  (34)

where the $\hat{O}()$ hides logarithmic terms in the gap, the alphabet size and the number of rounds. For a fixed gap, the dependence on the alphabet size is $|X_i|^{2/3} \vee \zeta(p_i)|X_i|$ which can be much smaller than the $|X_i|$ dependence obtained in Theorem 2 for the biased-based UCB.

To investigate the tightness of the regret bound of Theorem 11 for asymptotically large $t$, we may again compare it to the Lai-Robbins lower bound of Theorem 7 in the regime of small gaps (as in the Bernoulli case). For a sub-optimal arm $i$, the dominant term in the bound (with respect to $t$) then scales with $t$ as

$$\hat{O} \left( \frac{\log |X_i|}{\Delta_i} \cdot \log t \right),$$  (35)

where we may note the mild logarithmic dependence on the alphabet size. For simplicity of exposition, we next focus on the case of $K = 2$ arms, whose PMFs are $p_1$ and $p_2$ from a common probability simplex of dimension $|X| - 1$, and where without loss of generality (w.l.o.g.) it is assumed that $H(p_1) > H(p_2)$. As in the Bernoulli case, we will consider three regimes. In the first one, both $p_1$ and $p_2$ are at the interior of the simplex, in the second one $p_1$ is a uniform distribution, which is at the center of the simplex and has maximal entropy, and in the third one, both $p_1$ and $p_2$ approach a deterministic distribution, which is at one of the vertices of the simplex, and has zero entropy. We next consider each case in detail. First, suppose that $p_1$ is in the interior of the simplex, and that $p_2 \to p_1$. Using a Taylor approximation, it holds that

$$H(p_2) = H(p_1) + \nabla H(p)|_{p=p_1} \cdot (p_1 - p_2) + O \left( \|p_1 - p_2\|^2 \right),$$  (36)

where $\| \cdot \|$ is the Euclidean norm. Since $p_1$ is assumed to be in the interior of the simplex, and where $\nabla H(p)^T = -1^T - \log p(1), \log p(2), \ldots, \log p(|X|))$ has uniformly bounded entries, it holds that

$$\Delta = H(p_1) - H(p_2) = \Theta(\|p_1 - p_2\|).$$  (37)

At the same time, the KL divergence between $p_1$ and $p_2$ whenever $p_2 \to p_1$ approaches the chi-square divergence between them. Concretely, recall that the chi-square divergence for a pair of PMFs $p$ and $q$ with an alphabet $X$ such that $q(x) > 0$ for all $x \in X$ is defined as $\chi^2(p||q) := \sum_{x \in X} p(x)q(x)$. Then, [59, Thm. 4.1] states that

$$D_{\text{KL}}(p_2||p_1) = \frac{[1 + o(1)]}{2} \cdot \chi^2(p_2||p_1)$$  (38)

where $o(1) \to 0$ as $p_2 \to p_1$. Since $p_1$ is assumed to be in the interior of the probability simplex, it thus holds that

$$D_{\text{KL}}(p_2||p_1) = \Theta \left( \sum_{x \in X} (p_2(x) - p_1(x))^2 \right) = \Theta \left( \|p_1 - p_2\|^2 \right).$$  (39)
Combing both (37) and (39) results that the Lai-Robbins lower bound for the regret contribution of the sub-optimal arm \( i \) is \( \Omega\left(\frac{1}{\|p_1 - p_2\|} \cdot \log t\right) \), which agrees with the asymptotic scaling in (35) (after setting \( \Delta_i = \Theta(\|p_1 - p_2\|) \), as follows from (37)). Hence, as in the Bernoulli case, the asymptotic scaling of the upper bound on the regret in Theorem 11 roughly matches the scaling of the Lai-Robbins lower bound (as said, ignoring \( \log \log (t) \) terms). Second, consider the case in which \( p_1 \) is at the center of the simplex and thus has maximal entropy \( H(p_1) = \log |\mathcal{X}_1| \). Hence, the gradient of the entropy at this point vanishes in any direction within the probability simplex. Furthermore, the Hessian matrix \( H(p) \) of the entropy function \( H(p) \) is readily seen to be a diagonal matrix, whose value on the \( j \)th diagonal entry is given by \(-1/p(x) \) where \( x \) is the \( j \)th letter in \( \mathcal{X} \) (in an arbitrary order). A second-order Taylor approximation then implies that
\[
\Delta = H(p_1) - H(p_2) = \frac{1}{2} (p_1 - p_2)^T H(p_1) (p_1 - p_2) + O(\|p_1 - p_2\|^3)
\]
\[= \Theta(\|p_1 - p_2\|^2). \tag{40}
\]
Similarly to the previous case, the KL divergence is as in (39), that is \( D_{KL}(p_2||p_1) = \Theta(\|p_1 - p_2\|^2) \), and the Lai-Robbins lower bound for the regret contribution of the sub-optimal arm \( i \) is \( \Omega(\log t) \). By contrast, the regret term of the UCB algorithm is \( \tilde{O}(\log|\mathcal{X}| \cdot \log t) \), which is sub-optimal. Nonetheless, as in the Bernoulli case (see Proposition 8), we may derive an ameliorated confidence interval, which utilizes the vanishing gradient of the entropy function at the center of the probability simplex. In turn, this confidence interval would lead to an improved regret, which matches the Lai-Robbins lower bound. The details follow from a simple generalization of Proposition 8 and Theorem 9 and thus omitted. Consider, third, the case in which \( p_2 \) is at a vertex of the simplex, say, w.l.o.g., that \( p_2 = (1, 0, \ldots, 0) \), and hence, \( H(p_2) = 0 \). Now, assume that \( p_1 \rightarrow p_2 \), and specifically, that
\[
p_1 = \left(1 - \sum_{j=2}^{|\mathcal{X}|} \gamma_j, \gamma_2, \gamma_3, \ldots, \gamma_{|\mathcal{X}|}\right) \tag{41}
\]
with \( \max_{j \in |\mathcal{X}| \setminus \{1\}} \gamma_j \rightarrow 0 \). Then
\[
D_{KL}(p_2||p_1) = \log \left(\frac{1}{1 - \sum_{j=2}^{|\mathcal{X}|} \gamma_j}\right) = \Theta \left(\sum_{j=2}^{|\mathcal{X}|} \gamma_j\right), \tag{42}
\]
whereas
\[
\Delta = H(p_1) = \left[1 - \sum_{j=2}^{|\mathcal{X}|} \gamma_j\right] \log \left[1 - \frac{1}{1 - \sum_{j=2}^{|\mathcal{X}|} \gamma_j}\right] - \sum_{j=2}^{|\mathcal{X}|} \gamma_j \log \gamma_j
\]
\[= \Theta \left(\sum_{j=2}^{|\mathcal{X}|} \gamma_j \log \frac{1}{\gamma_j}\right). \tag{43}
\]
The regret upper bound of the algorithm (Theorem 9) is
\[
\tilde{O} \left(\frac{\log|\mathcal{X}|}{\sum_{j=2}^{\gamma_j} \gamma_j \log \frac{1}{\gamma_j}} \cdot \log t\right), \tag{44}
\]
whereas the lower bound (Theorem 7) is
\[
\tilde{O} \left(\frac{\sum_{j=2}^{\gamma_j} \gamma_j \log \frac{1}{\gamma_j}}{\sum_{j=2}^{\gamma_j} \gamma_j} \cdot \log t\right). \tag{45}
\]
As a more concrete example, if \( \gamma_j \equiv \gamma \) for all \( j \in |\mathcal{X}| \setminus \{1\} \) then the upper bound is \( \tilde{O}(\frac{\log |\mathcal{X}|}{\gamma} \cdot \log t) \) whereas the lower bound is \( \tilde{O}(\log \frac{1}{\gamma} \cdot \log t) \). Evidently, in this case, the upper and lower bounds on the scaling of the coefficient of the asymptotic log t term as a function of the gap do not match for small gaps, by a factor of \( 1/\gamma \). An interesting open problem is to close this gap, which most likely will require an improved algorithm (see the discussion in Section VIII).

For a gap-independent bound, we only need to consider the terms which blow-up as \( \Delta_i \downarrow 0 \), and this leads to a bound of the order \( \tilde{O}(\sqrt{\zeta(p_i) |\mathcal{X}|} + 1)^{3/4} + \sqrt{|\mathcal{X}|} \cdot t^{1/4} \), for which the leading term \( t \) is multiplied by roughly \( \sqrt{\zeta(p_i) |\mathcal{X}|} \), which again improves the dependence on the alphabet size.

VI. MULTI-ARMED BANDITS WITH ENTROPY REWARDS AND SUPPORT ESTIMATION

The previous sections assume that the player knows in advance the alphabet sizes of the sampled random variables. In practice, the exact alphabet size may not be known by the player in advance. Additionally, even when the player does know in advance the alphabet size, the actual support size can be drastically smaller than the alphabet size, and this leads to unnecessarily large confidence bounds which increase the player’s regret. To address this problem, we consider the case where the alphabet size is unknown, however, it is loosely upper bounded by a parameter \( \kappa \) via the assumption
\[
\min_{a \in \mathbb{N}_+} \{p(a) : p(a) > 0\} \geq \frac{1}{\kappa} \quad \text{for every arm } i \in [K].
\]
We next derive upper confidence bounds for unknown support and present a UCD with support size estimation that can be input to Algorithm 1. The proof of the theoretical results derived in this section are relegated to Appendix V.

A. A Concentration Inequality for Arm’s Support Size

Let \( Y \sim p \) be a discrete random variable over an \( \mathbb{N}_+ \), such that \( p(a) = \mathbb{P}[Y = a] \) for \( a \in \mathbb{N}_+ \). Assume that
\[
\min_{a \in \mathbb{N}_+} \{p(a) : p(a) > 0\} \geq \frac{1}{\kappa} \tag{46}
\]
for some given \( \kappa \geq 1 \), and so it holds that the support size of \( Y \) is at most \( \kappa \). Since the exact alphabet symbols of \( Y \) which take values in \( \mathbb{N}_+ \) are unknown, we denote the support of \( p \) w.r.t. the set \( \mathbb{N}_+ \), i.e., \( S(p) := \{a \in \mathbb{N}_+ : p(a) > 0\} \). Given \( n \) IID samples \( Y = (Y_1)_{i \in [n]} \) from \( p \), our goal is to find confidence interval for the support size \( S(p) := |S(p)| \).

For a given dataset \( (Y_i)_{i \in [n]} \), let
\[
N_a(n) := \sum_{i=1}^{n} \mathbb{1}[Y_i = a] \tag{47}
\]
be the number of appearances of \( a \) in \( (Y_i)_{i \in [n]} \). The simplest intuitive estimator is given by the plug-in estimator
\[
\hat{S}(Y, n) = \sum_{a \in \mathbb{N}_+} \mathbb{1}[N_a(n) > 0], \tag{48}
\]
which is the number of different symbols that appeared in $(Y_i)_{i \in [n]}$.

**Proposition 12:** Let $Y = \{Y_t\}_{t \in [n]}$ be IID from a discrete distribution $\rho$ over a finite alphabet $\mathcal{Y}$ such that $\rho(y) := P[Y = y]$. Then, assuming $n \geq 1$, it holds for any $\delta \in (0, 1)$ that

$$
\hat{S}(Y, n) \leq S(p) \leq \left( \hat{S}(Y, n) + \sqrt{\frac{1}{2} \log \left( \frac{1}{\delta} \right)} \cdot (1 - e^{-\frac{n}{\delta}})^{-1},
$$

with probability larger than $1 - \delta$.

### B. A UCB Regret Bound With Support Size Estimation

Next, we present a UCB policy with support size estimation and an upper bound on the resulting regret. We apply our analysis to the finite time regime where $\sqrt{\alpha \log(t)/2} < S(p) < \kappa$. In this regime the number of available samples for an arm is $o(\kappa \log \kappa)$ for which it is known that the plug-in estimator for the support size is known to grossly underestimate the true support size (see [60] for a detailed discussion). Furthermore, the regime $\sqrt{\alpha \log(t)/2} < S(p)$ considers the case in which the bound of Proposition 12 is non-trivial, and allows for $\delta$, chosen as $t^{-\alpha}$, to be positive.

In the results of the previous sections, both the biased-corrected entropy estimator and the confidence bounds depend on the alphabet size. Additionally, the pseudo-regret terms depend (monotonically) on the alphabet size. So, if the true support size is much smaller than the alphabet size, and if the player is aware of this, it can significantly reduce its pseudo-regret (bound). Nonetheless, when the player does not necessarily know the support size in advance, it can estimate it to reduce its pseudo-regret. Next, for simplicity of presentation, we focus on the biased-corrected entropy estimation presented in Section IV.

Define

$$B_{SE}(Y, \delta, n) := \log \left( 1 + \frac{1}{n} \left( \hat{S}(Y, n) + \sqrt{\frac{1}{2} \log \left( \frac{1}{\delta} \right)} \cdot (1 - e^{-\frac{n}{\delta}})^{-1} \right) \right),$$

and

$$\text{UCD}_{SE}(Y, \delta, n) := B_{SE}(Y, \delta, n) + \sqrt{\frac{2 \log^2(n)}{n} \log \left( \frac{2}{\delta} \right)}.$$

Now, we present the following upper confidence interval bound for the bias-corrected entropy estimator with an unknown support size.

**Proposition 13:** Let $Y = \{Y_t\}_{t \in [n]}$ be IID from a discrete distribution $\rho$ over a finite alphabet $\mathcal{Y}$ such that $\rho(y) := P[Y = y]$. Then, assuming $n \geq 2$, it holds for any $\delta \in (0, 1)$ that

$$|H(\hat{p}(n)) - H(p)| \leq \text{UCD}_{SE}(Y, \delta, n),$$

with probability larger than $1 - 2\delta$.

Using this confidence interval, we may adapt Algorithm 1 to include upper confidence bounds on the support size estimation, for each arm, and then derive an upper bound on the pseudo-regret of Algorithm 1 when the support size is estimated. To this end, denote

$$\Gamma_{SE}(\alpha, \beta, S, \kappa, \Delta, t) := \max \left\{ \frac{2\sqrt{\kappa \left( \sqrt{S} + (\alpha \log(t)/2)^{\frac{3}{2}} \right)} \cdot \left( 1 - e^{\beta \Delta/2 - 1} \right)^{\frac{1}{2}}}{\left( 1 - \beta \right)^{\frac{3}{2}} \Delta^2}, \frac{15}{8} \cdot \log \left( \frac{2t^{\alpha}}{(1) - \beta \Delta^2} \right) \right\}.$$

**Theorem 14:** Let $\beta \in (0, 1)$ be given, $\delta \equiv \delta_{\alpha}(t) = t^{-\alpha}$ and $\alpha > 2$. Additionally, denote by $S(p_i)$ the support size of arm $i$ and assume that $\{\kappa_i\}_{i \in [K]}$ are given and are such that $S(p_i) \leq \kappa_i$ for all $i \in [K]$. Assume that Algorithm 1 is run with a plug-in entropy estimator $H(Y, n) \equiv H(\hat{p}(n))$, and upper confidence deviation $\text{UCD}(Y, \delta, n) \equiv \text{UCD}_{SE}(Y, \delta, n)$. Then, the pseudo-regret is bounded as

$$R(t) \leq \sum_{i \in [K]; \Delta_i > 0} \left[ \Gamma_{SE}(\alpha, \beta, \kappa_i, \Delta_i, t) \cdot \Delta_i + \frac{4(\alpha - 1)}{\alpha - 2} \cdot \Delta_i \right].$$

Note that $\Gamma_{SE}(\alpha, \beta, S, \kappa, \Delta, t) = O(\kappa^{2/3})$, for given $\beta$ and $\Delta$, whenever $\sqrt{S} \cdot \alpha \log(t) = o(\kappa^{2/3})$. In this case, estimating the support significantly decreases the pseudo-regret by an order of $\kappa^{1/3}$ with respect to that of the bias-correction entropy estimator we present in Section IV where $|\mathcal{Y}| = \kappa$, cf. (9).

### VII. Numerical Experiments

In this section we present numerical experiments that illustrate the average total regret achieved by Algorithm 1 for the various upper confidence bounds we develop. We examined the setups summarized in Table II, each includes two arms, the subscript 1 denotes quantities of the first arm, similarly, the subscript 2 denotes quantities of the second arm.

We set $\alpha = 2.1$ and ran each setup for $1.5 \times 10^6$ rounds. Additionally, for each setup, we averaged the total regret across 100 Monte Carlo realizations.

Figure 1 presents numerical results for the binary alphabet, i.e., Setups 1-3. Additionally, Figure 2 presents the numerical results for the ternary alphabet, i.e., Setups 4-6. The lines ‘Bias’ depict the average total regret of the bias-corrected confidence interval used in (7) with the plug-in entropy estimator. To evaluate the sensitivity of (7) to the alphabet size compared with the support size, we examine multiple alphabet sizes $\kappa$ as is described in Table II. The lines ‘TV’ denote the PMF-based confidence intervals that are used with the plug-in entropy estimator. Here too $\kappa$ denotes the alphabet size known to the player. In the case of a binary alphabet $\kappa = 2$, we take the minimum between the confidence interval (12) and the confidence interval (20). In the case of a larger alphabet, i.e., $\kappa \geq 3$, we use the general alphabet confidence interval (28). Finally, The lines ‘Bias SE’ depict the
average total regret of the bias-corrected confidence interval with support estimation used in (51) along with the plug-in entropy estimator.

For the case of binary support size that is known to the player, i.e., \( \kappa = 2 \), it is evident from Figure 1 that the Bernoulli PMF-based confidence intervals (12) and (20)
provide a significant reduction in the average total regret in comparison to the combination of the bias-corrected estimator and confidence interval used in (7). Furthermore, we can see that, as expected, the bias correction approach suffers from significantly increased regret values as the probabilities of drawing the symbol ‘1’, i.e., \( p(1) \), of the arms get closer to the boundary points of the interval \([0, 1/2]\). The Bernoulli PMF-based confidence intervals exhibit robustness in these regimes and have not suffered from such stark degradation in performance. It is also evident from Figure 1 that the resulted regrets of both the bias-corrected and the PMF-based approaches increase as the known alphabet size, i.e., \( \kappa \) increases and the support size remains fixed. We note that since the alphabet size affects the biased corrected approach through a logarithmic term in (7) it is robust to small variations in the alphabet size. In fact, even when \( \kappa \) is set to \( 10^3 \) the resulting increase in regret is small. Nonetheless, as we continue to increase \( \kappa \) to \( 10^5 \) the regret increases significantly. Interestingly, Figure 1 shows that the PMF approach is very sensitive to the choice of the alphabet size \( \kappa \) where the increase in regret is noticeable even for \( \kappa = 10 \). Finally, we can see from Figure 1 that regret of the bias-corrected approach with support estimation is robust to the changes in \( \kappa \) even for very large values of \( \kappa \) such as \( 10^5 \).

As we increase the alphabet size from two (binary) to three (ternary), Figure 2 shows that the general PMF-based confidence interval (28) does not perform as well as the binary ones, i.e., (12) and (20). This occurs since the general PMF-based confidence interval (28) targets scenarios where \( \zeta(p_i) \cdot |X_i| \) is sufficiently small. Furthermore, we can see that \( \kappa \) impacts the experience regret of sources with ternary support size, i.e., Figure 2 in a very similar way to its impact on the regret in the case of a binary support size, see Figure 1.

In addition to Setups 1-6 that capture scenarios with small alphabet sizes, we consider a scenario with a large alphabet size, namely, one with \( 10^4 \) symbols. For the first arm, the total probability of the first \( 10^4 - 1 \) symbols is \( 5 \times 10^{-3} \), these probabilities are chosen randomly by generating \( 10^4 - 1 \)
Weinberger and Yemini: Multi-Armed Bandits with Self-Information Rewards

In this paper, we have introduced the IMAB problem, in which a player aims to maximize the information it observes from a set of possible sources, and concretely focused on the entropy functional. We have proposed a basic bias-corrected UCB algorithm, and showed its inefficiency whenever the maximal-arm entropy is very low compared to the log-alphabet size. We thus proposed refined UCB algorithms that are based on data-dependent UCDs, and which significantly improve the bias-corrected UCB algorithm. Specifically, we have first considered the Bernoulli case (binary alphabets) and showed that its pseudo-regret bound has almost optimal gap-independent regret bound, and agrees order-wise with Lai-Robbins imposibility lower bound for the gap-dependent regret bound, unless the maximal-entropy has a PMF close to a uniform one (and thus the maximal possible entropy). In this high-entropy regime, we have proposed a dedicated confidence interval for the entropy, which when combined with the previous data-dependent lower bound closes the gap to the Lai-Robbins lower bound in all possible regimes of the maximal-entropy arm.

Fig. 3. Average total regret as a function of the number of rounds for a two-armed bandit model for arms with an alphabet size of $10^5$ and a support size of $10^5$. (Setup 7).

Acknowledgment

This work was supported by the Israel Science Foundation (Grant No. 1417/16) and the Israel National Institute for Information Technology (INIT) grant. We would like to thank the anonymous reviewers for their valuable comments, which helped us improve the quality of this paper.
Furthermore, the UCB-V algorithm [56] and the KL-UCB algorithm [58], [62], [63], [64] refine the confidence interval bounds that are used for optimistically estimating the reward of each arm, either using Bernstein-type concentration inequalities (UCB-V) or using the tight large-deviations bounds of the Chernoff bound with the KL divergence as the rate function (KL-UCB). The confidence intervals we have derived in this paper for the IMAB problem are based on the total variation, and similarly attempt to achieve tight confidence bounds for the entropy. In the Bernoulli case, we have used a relative (multiplicative) Chernoff bound on a confidence interval, which is similar in this context to a Bernstein-type bound, and this has led to almost optimal regret bounds. For general-size alphabets, it is conceivable that a KL-UCB algorithm with refined analysis will be able to close the gap to the Lai-Robbins lower bound. Technically, this amounts to defining a modified KL-inf type index, which amounts to the entropy of the PMF with the largest possible entropy given a KL distance to the empirical PMF of that arm (c.f. [58, Algorithm 1]). Future work calls for a refined analysis of this type.

In the last part of the paper, we have relaxed the assumption that the player knows in advance the alphabet of each arm. In practice, the alphabet of the arm may be very large compared to the support of the PMF, thus it is of interest to develop UCB algorithms for this case. To that end, we additionally developed a bias-corrected UCB algorithm with support estimation which implements a UCB approach for estimating both the support size and the resulted entropy of each arm.

Beyond the above, as surveyed in [30, Sec. 2.4 and references therein], the bounds can be refined to distribution-free bounds, high probability bounds, algorithms such as ε-greedy or Thompson sampling can be considered, and the probabilistic arm model may be generalized to heavy-tailed distributions. The IMAB problem proposed in this paper may be addressed according to these avenues. Finally, in the considered IMAB model the arms are statistically independent. In practical setups, such as in surveillance and coverage problems, the underlying probability functions of different arms can be correlated. Correlated arm counterparts of the classical MAB problem have been investigated in works such as [65]. Exploring related models for IMAB setups can be another interesting avenue of future research.

APPENDIX I
INVERTING POLYLOGARITHMIC FUNCTIONS OVER LINEAR FUNCTIONS

Lemma 15: Let \( r \in [1, 2] \) be given. There exists a constant \( c_r > 0 \) so that if \( x \geq c_r \log^r (1/y) \) then \( \log^r x \leq y \). This bound is orderwise tight as \( y \downarrow 0 \). Specifically, this holds for the constants \( c_1 = 2, c_{4/3} = 3 \) and \( c_2 = 15 \).

Proof: On \( \mathbb{R}_+ \), the function \( x \rightarrow \log^r x \) has a unique maximum at \( x = e^r \), and its maximal value is \((\frac{1}{e^r})^r\) (which is less than 1 for any \( r \in [1, 2] \)). So, \( \log^r \frac{x}{y} \) is monotonically decreasing for \( x \geq e^r \). If \( y \geq (\frac{1}{e^r})^r \) then \( \log^r y \leq y \) for all \( x \in \mathbb{R}_+ \) and the claim of the lemma trivially holds. Otherwise, if \( y \in [0, (\frac{1}{e^r})^r] \) then setting \( x = c_r \log^r (1/y) / y \) results

\[
\frac{x}{y} = y \log^r \left( \frac{c_r \log^r (1/y)}{y} \right) \\
\leq y \cdot \left[ \frac{\log(c_r) + r \log(1/y) + \log(1/y)}{c_r^{1/r} \cdot \log(1/y)} \right]^r \\
\leq y \cdot \left[ \frac{\log(c_r) + (r + 1) \log(1/y)}{c_r^{1/r} \cdot \log(1/y)} \right]^r \\
\leq y \cdot \sup_{y' \in [0,(\frac{1}{e^r})^r]} \left[ \frac{\log(c_r) + (r + 1) \log(1/y')}{{c_r}^{1/r} \cdot \log(1/y')} \right]^r \\
= y \cdot \left[ \frac{\log(c_r)}{c_r^{1/r} \cdot r \log(\frac{1}{e^r})} + \frac{(r + 1)}{c_r^{1/r}} \right]^r. \tag{55}
\]

For any given power \( r \), the term inside the square brackets can be made arbitrarily small by taking \( c_r \sim \infty \), and specifically, can be made less than 1, which results \( \log^r \frac{x}{y} \leq y \) for the aforementioned choice of \( x \), with some numerical constant \( c_r \). The minimal constant can be found by checking (55) numerically, and this leads to the constants in the claim of the lemma. Finally, this value of \( x \) is orderwise tight since if \( x = o(\log^r (1/y))/y \) (where the asymptotic-\( o \) notation is as \( y \downarrow 0 \)), then \( \log^r \frac{x}{y} = \omega(1) \). □

APPENDIX II
PROOFS FOR SECTION IV

Proof of Proposition 1: For the upper confidence bound it holds that

\[
\mathbb{P} \left( H(\hat{p}(n)) - H(p) > B(n) + \epsilon \right) = \\
\mathbb{P} \left( H(\hat{p}(n)) - E(H(\hat{p}(n))) > \epsilon + (H(p) - E(H(\hat{p}(n)))) + B(n) \right) \\
\leq (a) \mathbb{P} \left( H(\hat{p}(n)) - E(H(\hat{p}(n))) > \epsilon \right) \\
\leq (b) \exp \left( -\frac{n}{2} \left( \frac{\epsilon}{\log(n)} \right)^2 \right), \tag{56}
\]

where (a) follows from the bound on the bias in (5), and (b) follows from [34, p. 168]. Similarly, for the lower confidence bound it holds that

\[
\mathbb{P} \left( H(\hat{p}(n)) - H(p) < -\epsilon - B(n) \right) = \\
\mathbb{P} \left( H(\hat{p}(n)) - E(H(\hat{p}(n))) < -\epsilon - H(p) - E(H(\hat{p}(n))) - B(n) \right) \\
\leq (a) \mathbb{P} \left( H(\hat{p}(n)) - E(H(\hat{p}(n))) < -\epsilon \right) 
\]
(b) $\exp \left[ -\frac{n}{2} \left( \frac{\epsilon}{\log(n)} \right)^2 \right].$ (57)

Combining (56) and (57) shows that

$$\mathbb{P} \left( |H(\hat{p}(n)) - H(p)| > B(n) + \epsilon \right) \leq 2 \exp \left[ -\frac{n}{2} \cdot \frac{\epsilon^2}{\log^2(n)} \right]$$

(58)

for every $n \geq 2$ and $\epsilon > 0$. Setting the RHS of (58) to $\delta$ and simplifying leads to the claimed result. \qed

The proof of Theorem 2 requires the following lemma, which lower bounds the number of samples required for a sufficiently low upper confidence interval.

**Lemma 16:** Let an alphabet $\mathcal{Y}$ be given, let a gap $\Delta \in (0, \log|\mathcal{Y}|]$ be given, and let $\delta = t^{-\alpha}$. Then, for any $\beta \in (0, 1)$, if $n \geq \Gamma_{\text{bias}}(\alpha, \beta, \mathcal{Y}, \Delta, t)$ then $\text{UCD}_{\text{bias}}(t^{-\alpha}, n) \leq \Delta / 2$.

**Proof:** We may assume that $n > 1$. Let $\beta \in [0, 1]$ be given. Then, $\text{UCD}_{\text{bias}}(t^{-\alpha}, n) \leq \Delta / 2$ if both

$$B(n) \leq \beta \cdot \Delta / 2,$$

(59)

and

$$\sqrt{\frac{2 \log^2(n)}{n}} \log \left( \frac{2}{\beta} \right) \leq (1 - \beta) \cdot \Delta / 2,$$

(60)

holds. The first condition (59) is equivalent to

$$n \geq \frac{|\mathcal{Y}| - 1}{e^{\beta \cdot \Delta^2 / 4} - 1},$$

(61)

and the second condition (60) is equivalent to

$$\log^2(n) \leq \frac{(1 - \beta)^2 \Delta^2}{8 \log(2^{\alpha})}.$$

(62)

According to Lemma 15 this holds if

$$n \geq \frac{120 \cdot \log(2^{\alpha}) \cdot \log \left( \frac{8 \log(2^{\alpha})}{(1 - \beta)^2 \Delta^2} \right)}{(1 - \beta)^2 \Delta^2}$$

$$= 15 \cdot \Lambda_2 \cdot \frac{8 \log(2^{\alpha})}{(1 - \beta)^2 \Delta^2}$$

(63)

(recall the notation in (1)). Simplifying both expressions and optimizing over $\beta \in [0, 1]$ concludes the proof. \qed

With this result at hand, we may prove Theorem 2.

**Proof of Thm. 2:** The proof follows the analysis of [30, Proof of Thm. 2.1], with required modifications to entropy rewards structure. At round $t$, the player chooses a suboptimal $i$ arm with $\Delta_i > 0$ if

$$\hat{H}(X_i, (t-1), N_i, (t-1)) + \text{UCD}_{\text{bias}}(\delta_\alpha(t), N_i, (t-1))$$

$$\leq \hat{H}(X_i, (t-1), N_i, (t-1)) + \text{UCD}_{\text{bias}}(\delta_\alpha(t), N_i, (t-1)).$$

(64)

For this to occur at least one of the following events must occur too (sufficient conditions):

I. The entropy of the best arm is significantly underestimated:

$$H_i \geq \hat{H}(X_i, (t-1), N_i, (t-1)) + \text{UCD}_{\text{bias}}(\delta_\alpha(t), N_i, (t-1)).$$

(65)

II. The entropy of arm $i$ is significantly overestimated:

$$\hat{H}(X_i, (t-1), N_i, (t-1)) > H_i + \text{UCD}_{\text{bias}}(\delta_\alpha(t), N_i, (t-1)).$$

(66)

III. The upper confidence interval is significantly larger than the gap

$$\text{UCD}_{\text{bias}}(\delta_\alpha(t), N_i, (t-1)) > \Delta_i / 2.$$ (67)

If all three events I-III are false, then

$$\hat{H}(X_i, (t-1), N_i, (t-1)) + \text{UCD}_{\text{bias}}(\delta_\alpha(t), N_i, (t-1)) > H_i, \text{ or } H_i + \Delta_i,$$

$$\geq H_i + 2 \cdot \text{UCD}_{\text{bias}}(\delta_\alpha(t), N_i, (t-1))$$

$$\geq \hat{H}(X_i, (t-1), N_i, (t-1)) + \text{UCD}_{\text{bias}}(\delta_\alpha(t), N_i, (t-1)).$$

(68)

which contradicts the assumption that Algorithm 1 chooses $I_i = i$ at the $t$th round.

Next, we upper bound the expected pseudo-regret (3) of Algorithm 1 with the entropy estimator and confidence bound stated in the theorem. To that end, we upper bound the expected number of times a sub-optimal arm $i$ is played, i.e., $\mathbb{E}(N_i(t))$ as follows. Note that if $N_i(t) \geq \Gamma_{\text{bias}}(\alpha, \beta, X_i, \Delta_i, t)$ then event III does not occur. so,

$$\mathbb{E}(N_i(t)) = \mathbb{E} \left( \sum_{\tau=1}^{t} [I(\tau) = i] \right)$$

$$\leq \Gamma_{\text{bias}}(\alpha, \beta, X_i, \Delta_i, t) + \sum_{\tau = \Gamma_{\text{bias}}(\alpha, \beta, X_i, \Delta_i, t)+1}^{t} \mathbb{P}(I \text{ is true at round } \tau)$$

$$+ \mathbb{P}(\text{II is true at round } \tau).$$

(69)

For any $\tau \leq t$, the first probability in (69) is upper bounded as

$$\mathbb{P}(I \text{ is true at round } \tau) \leq \sum_{n=1}^{\tau} \mathbb{P} \left( \hat{H}(X_i(t), t \in [n], n) \right) \leq H_i$$

(70)

where (a) follows from the union bound, and (b) from the definition of the upper confidence deviation $\text{UCD}(\delta, n)$. The second probability in (69) is similarly upper bounded. Inserting these bounds back to the sum in (69) it then follows that

$$\sum_{\tau=1}^{t} \mathbb{P}(I \text{ is true at round } \tau) \leq 2 \sum_{\tau=1}^{\infty} \frac{1}{\tau^{\alpha-1}} \leq 2 \sum_{\tau=1}^{\infty} \frac{1}{\tau^{\alpha-1}}.$$
\[
\leq 2 \left[ 1 + \int_1^{\infty} \frac{1}{\tau^{\alpha-1}} d\tau \right] = \frac{2(\alpha - 1)}{\alpha - 2}.
\]
Substituting the upper bounds in the last two displays back to (69), and using the resulting bound in \( R(t) = \sum_{i \in [\mathcal{K}]: \Delta_i > 0} \mathbb{P}(N_i(t)) \Delta_i \) then concludes the proof. \( \square \)

**APPENDIX III**

**Proofs for Section V-A**

The proof of Proposition 3 is based on a standard concentration result on the empirical mean of a Bernoulli source. 

**Lemma 17:** In the setting of Proposition 3, each of the following events holds with probability larger than \( 1 - \delta \):

\[
|p - \hat{p}(n)| \leq \sqrt{\frac{3p \log \left( \frac{2}{\delta} \right)}{n}},
\]

\[
p \leq 2\hat{p}(n) + \frac{12\log \left( \frac{1}{\delta} \right)}{n},
\]

and

\[
\hat{p}(n) \leq 2p + \frac{3\log \left( \frac{1}{\delta} \right)}{n}.
\]

**Proof:** We will use the relative (multiplicative) Chernoff bound multiple times. This bound states that \([66, \text{Thm. 4.4}]

\[
P \left[ |\hat{p}(n) - p| \geq \xi p \right] \leq P \left[ |\hat{p}(n) - p| \geq \xi p \right] + P \left[ |\hat{p}(n) - p| \leq -\xi p \right] \leq 2e^{-\frac{\xi^2pn}{2}},
\]

for any \( \xi \in [0, 1] \) (and it holds for the pair of one-sided deviations each without the 2 pre-factor). Setting \( \xi = \sqrt{\frac{3\log \left( \frac{2}{\delta} \right)}{pn}} \) in (75) immediately leads to (72). Next, if \( p > \frac{12\log \left( \frac{1}{\delta} \right)}{n} \) then

\[
P [p \geq 2\hat{p}(n)] = P \left[ p - \frac{1}{2}p \leq \hat{p}(n) \right] \leq e^{-\frac{\xi^2pn}{2}}; \leq \delta,
\]

where \((a)\) is by setting \( \xi = 1/2 \) in the one-sided version of (75), and \((b)\) utilizes the assumption on \( p \). Thus, with probability larger than \( 1 - \delta \) it holds that

\[
p \leq 2\hat{p}(n) + \frac{12\log \left( \frac{1}{\delta} \right)}{n},
\]

which can be loosened to (73). Finally, If \( p > \frac{3\log \left( \frac{1}{\delta} \right)}{n} \) then

\[
P \left[ \hat{p}(n) > 2p \right] = P \left[ \hat{p}(n) - p \geq \xi p \right] \leq e^{-\frac{\xi^2pn}{2}} ; \leq \delta,
\]

where \((a)\) is by setting \( \xi = 1 \) in the one-sided version of (75), and \((b)\) utilizes the assumption on \( p \). Thus, with probability larger than \( 1 - \delta \) it holds that

\[
\hat{p}(n) \leq 2p + \frac{3\log \left( \frac{1}{\delta} \right)}{n},
\]

which can be loosened to (74). \( \square \)

The concentration of the empirical probability of the source then leads to a confidence bound on the entropy, as next shown in the proof of Proposition 3.

**Proof of Proposition 3:** If \( d_{TV}(p, \hat{p}(n)) \leq \frac{1}{2} \) then \([57, \text{Lemma 2.7}]\) implies that

\[
|h_b(\hat{p}(n)) - h_b(p)| \leq \sqrt{\frac{12p \log \left( \frac{2}{\delta} \right)}{n}} \log \left( \frac{4n}{12p \log \left( \frac{2}{\delta} \right)} \right)
\]

and we note that \(-\Lambda_1(s)\) is monotonic increasing for \( s \in [0, e^{-1}] \). For a pair of Bernoulli distributions \( p \) and \( q \) it holds that

\[
d_{TV}(p, q) = 2|p(1) - q(1)|,
\]

and so by (72) and (73) from Lemma 17 it holds that

\[
d_{TV}(p, q) \leq \sqrt{\frac{12p \log \left( \frac{2}{\delta} \right)}{n}},
\]

and

\[
p \leq 2\hat{p}(n) + \frac{12\log \left( \frac{1}{\delta} \right)}{n},
\]

simultaneously hold with probability larger than \( 1 - 2\delta \). To be in the monotonic increasing regime of \(-\Lambda_1(s)\) for any \( \hat{p}(n) \), we require that the upper bound on the total variation distance in (82), when substituted with the upper bound on \( p \) in (83), is less than \( e^{-1} \), to wit

\[
\sqrt{\frac{12 \left( 2\hat{p}(n) + \frac{12\log \left( \frac{1}{\delta} \right)}{n} \right)}{n}} \log \left( \frac{n}{\hat{p}(n) \log \left( \frac{2}{\delta} \right)} \right) \leq e^{-1}.
\]

This can be easily seen to be satisfied by the assumption \( n \geq 200 \cdot \log \left( \frac{2}{\delta} \right) \). Now, if \( 2\hat{p}(n) \geq \frac{12\log \left( \frac{1}{\delta} \right)}{n} \) then \( p \leq 4\hat{p}(n) \) and so by the assumption of \( n \) and the resulting monotonicity,

\[
|h_b(\hat{p}(n)) - h_b(p)| \leq \sqrt{\frac{12\hat{p}(n) \log \left( \frac{2}{\delta} \right)}{n} \frac{24 \log \left( \frac{2}{\delta} \right)}{n}} \log \left( \frac{n}{\hat{p}(n) \log \left( \frac{2}{\delta} \right)} \right)
\]

(after slightly deteriorating the constants to obtain a succinct expression). Otherwise, if \( \frac{12\log \left( \frac{1}{\delta} \right)}{n} \geq 2\hat{p}(n) \) then \( p \leq \frac{24 \log \left( \frac{1}{\delta} \right)}{n} \) and so by the assumption of \( n \) and the resulting monotonicity,

\[
|h_b(\hat{p}(n)) - h_b(p)| \leq \frac{18 \log \left( \frac{2}{\delta} \right)}{n} \log \left( \frac{n}{\hat{p}(n) \log \left( \frac{2}{\delta} \right)} \right)
\]

(after, again, slightly deteriorating the constants). To account for both cases, we sum the two deviation terms. Finally, to obtain (12), we replace \( \delta \) with \( 2\delta \). \( \square \)

Next, we turn to the proof of Theorem 6, which is based on a lemma analogous to Lemma 16. To this end, we further denote a simplified version of \( \Gamma_{ber}(\cdot) \) from (16), defined as

\[
\tilde{\Gamma}_{ber}(\alpha, \beta, q, \Delta, t) := \max \left\{ 2 \cdot \Lambda_1 \left( \frac{36q t \log(t)}{(1 - \beta) \Delta} \right), \frac{48 \cdot 965q t \log(t)}{\beta^2 \cdot \Delta^2} \cdot \log^2 \left( \frac{48}{\beta^2 \cdot \Delta^2} \right) \right\}.
\]

**Lemma 18:** With \( \delta \equiv \delta_\alpha(t) = 4t^{-\alpha} \), \( \beta \in (0, 1) \) and \( \alpha > 2 \), if \( n \geq \tilde{\Gamma}_{ber}(\alpha, \beta, q, \Delta, t) \) then \( \text{UCD}_{ber}(q, \delta_\alpha(t), n) \leq \Delta/2 \) where \( \text{UCD}_{ber}(\cdot) \) is as defined in (12).

**Proof:** We may assume that \( n \geq e \); this can easily be achieved by playing each arm for three rounds at the
beginning of Algorithm 1. Let $\beta \in [0, 1]$ be given. Then, UCDber$(q, 4t^{-\alpha}, n) \leq \Delta/2$ if both
\[
\sqrt{12 q \alpha \log(t) \log \left( \frac{n}{q \alpha \log(t)} \right)} \leq \beta \cdot \Delta/2,
\]
and
\[
\frac{18 \alpha \log(t) \log(n)}{n} \leq (1 - \beta) \cdot \Delta/2,
\]
hold. The first condition is satisfied if
\[
\frac{\log^2 \left( \frac{n}{q \log(\frac{\Delta}{\beta \Delta})} \right)}{\log \left( \frac{\Delta}{\beta \Delta} \right)} \leq \frac{\beta^2 \cdot \Delta^2}{48},
\]
for which Lemma 15 implies that this condition is satisfied if
\[
n \geq \frac{960 \alpha q \log(t) \beta^2 \cdot \Delta^2 \cdot \log^2 \left( \frac{48}{\beta^2 \cdot \Delta^2} \right)}{n}.
\]
(91)

The second condition is satisfied if
\[
\frac{\log(n)}{n} \leq \frac{(1 - \beta) \Delta}{36 \alpha \log(t)},
\]
for which Lemma 15 implies that this condition is satisfied if
\[
n \geq \frac{(1 - \beta) \Delta \log(36 \alpha \log(t))}{(1 - \beta) \Delta} = 2 \cdot \Lambda_1 \left( \frac{36 \alpha \log(t)}{(1 - \beta) \Delta} \right)
\]
(93)

(92)

(recall the notation (1)). The claim of the lemma then follows from the definition of $\hat{\Gamma}_{ber}(\cdot)$ in (87). 

We may now prove Theorem 6.

Proof of Theorem 6: The proof is similar to the proof of Theorem 2, and so we only highlight the main differences. In what follows it will be convenient to interchangeably use both UCD$(Y, \delta, n)$ and UCDber$(\hat{p}(Y, n), \delta, n)$ to denote the (same) upper confidence bound used by the algorithm. At round $t$, the player chooses a sub-optimal $i$ arm if $\Delta_i > 0$ and
\[
\hat{H}(X_i, (t - 1), N_i, (t - 1)) + UCD(X_i, (t - 1), \delta_i(t), N_i, (t - 1)) \leq \hat{H}(X_i, (t - 1), N_i, (t - 1)) + UCD(X_i, (t - 1), \delta_i(t), N_i, (t - 1)).
\]
(94)

For this to occur at least one of the following events must occur too (sufficient conditions):

\textbf{I’}. Either the entropy of the best arm is significantly underestimated
\[
\hat{H}(X_i, (t - 1), N_i, (t - 1)) + UCD(X_i, (t - 1), \delta_i(t), N_i, (t - 1)) \leq H_i^*.
\]
(95)

or
\[
\hat{p}(X_i, (t - 1), N_i, (t - 1)) - \frac{1}{2} p_i^* \leq -\frac{6 \log(1/\delta_i(t))}{N_i, (t - 1)}.
\]
(96)

\textbf{II’}. Either the entropy of arm $i$ is significantly overestimated
\[
\hat{H}(X_i, (t - 1), N_i, (t - 1)) > H_i + UCD(X_i, (t - 1), \delta_i(t), N_i, (t - 1)),
\]

or
\[
\hat{p}(X_i, (t - 1), N_i, (t - 1)) - 2 p_i \geq \frac{3 \log(1/\delta_i(t))}{N_i, (t - 1)}.
\]
(98)

\textbf{III’}. The upper confidence interval, which is based on an overestimation of $\hat{p}(X_i, (t - 1), N_i, (t - 1))$ is significantly larger than the gap
\[
UCDber \left( 2 p_i - 3 \log(1/\delta_i(t)) \right) \geq \frac{\Delta_i}{2},
\]
(99)

or
\[
N_i(t - 1) \leq 200 \alpha \log(t).
\]
(100)

As in the proof of Theorem 2, if all three events I’-III’ are false, then
\[
\hat{H}(X_i, (t - 1), N_i, (t - 1)) + UCD(X_i, (t - 1), \delta_i(t), N_i, (t - 1)) \geq H_i^* = H_i + \Delta_i \\
\geq H_i \leq 2 UCDber \left( 2 p_i - 3 \log(1/\delta_i(t)) \right) \geq \frac{\Delta_i}{2},
\]
(101)

\[\text{where in }\ast\text{ we have used the current assumption that } N_i(t - 1) \geq 200 \alpha \log(t), \text{ which assures that } UCDber(q, \delta_i(t), N_i, (t - 1)) \text{ is monotonically non-decreasing function of } q. \text{ Thus, in this case Algorithm 1 will not choose } I_i = i \text{ at the } t^{th} \text{ round; a contradiction.}
\]

By Lemma 18, if
\[
N_i(t - 1) \geq \hat{\Gamma}_{ber} \left( \alpha, \beta, 2 p_i + \frac{3 \log(1/\delta_i(t))}{N_i, (t - 1)}, \Delta_i, t \right),
\]
(102)

then the first part of the event III’ does not occur. By the definition of $\hat{\Gamma}_{ber}(\cdot)$ in (87), and by setting $\delta_i(t) = 4t^{-\alpha}$, the RHS in the last equation is upper bounded as

\[
\max \left\{ 2 \cdot \Lambda_1 \left( \frac{36 \alpha \log(t)}{(1 - \beta) \Delta_i} \right), \right. \\
\left. \frac{2560 \alpha q \log(t) \log^2 \left( \frac{48}{\beta^2 \cdot \Delta_i^2} \right)}{\beta^2 \cdot \Delta_i^2 N_i, (t - 1)} \cdot \log^2 \left( \frac{48}{\beta^2 \cdot \Delta_i^2} \right) \right\}.
\]
(103)

This can be guaranteed by requiring that $N_i(t - 1)$ is larger than each of the first two terms, as well as larger than twice of
each of the additive components of the third term. To conclude, a sufficient condition for the event III” not to occur is that
\[ N_1(t-1) \geq \max \left\{ 2 \cdot \Lambda_1 \left( \frac{36 \alpha \log(t)}{(1-\delta) \Delta_i^3} \right), \frac{1520 \alpha \log(t)}{\beta_1^2 \Delta_i^2} \cdot \log^2 \left( \frac{48}{\beta_1^2 \Delta_i^2} \right), \frac{88 \sqrt{\alpha \log(t)}}{\beta_1 \Delta_i} \cdot \log \left( \frac{48}{\beta_1^2 \Delta_i^2} \right) \right\} \]
\[ = \Gamma_{\text{ber}}(\alpha, \beta, p_1, \Delta_i, t). \]  
(104)

The second part of event III” does not occur if \( N_1(t-1) \geq 200 \alpha \log(t) \), which is already covered by the condition in (104) if we increase the pre-constant of the second term to 6, which is the definition of \( \Gamma_{\text{ber}}(\cdot) \) used in (16).

The analysis then follows as in the proof of Theorem 2, by using Lemma 17 and Proposition 3 to bound the probabilities of the events in I” and II. Note that the condition \( N_1(t-1) \geq 200 \cdot \log(t) \) required for the confidence bound to hold with high probability is already satisfied by (104). □

**Proof of Proposition 8:** By Taylor approximation at the point \( p \), for any \( q \in [0, \frac{1}{2}] \)
\[ h_b(q) = h_b(p) + h_b'(p)(q-p) + \frac{h_b''(\xi)}{2}(q-p)^2, \]  
(105)
where \( \xi \in [p, q] \cup [q, p] \). From Lemma 17, it holds with probability larger than \( 1 - 2\delta \) that both \( p \leq \hat{p}(n) + \frac{12 \log(\frac{2}{\delta})}{n} \) and \( |p - \hat{p}(n)| \leq \frac{3 \sqrt{2 \log(\frac{2}{\delta})}}{n} \). Under this event, since \( n \geq 60 \log(\frac{2}{\delta}) \) was assumed, it holds that \( \hat{p}(n) \geq \frac{1}{10} \). For \( q \in [\frac{3}{5}, \frac{1}{2}] \), it can be easily verified that
\[ |h_b'(q)| = \left| \log \frac{1 - q}{q} \right| \leq 5 \left( \frac{1}{2} - q \right), \]  
(106)
and for any \( q \in [\frac{1}{10}, \frac{1}{2}] \), it holds that \( |h_b''(q)| \leq 12 \). Hence, by (105), and under the high probability event
\[ |h_b(\hat{p}(n)) - h_b(p)| \leq \frac{5}{2} |\hat{p}(n) - p| + 6 (\hat{p}(n) - p)^2 \\
\leq \frac{7}{2} |\hat{p}(n) - p| \sqrt{\frac{\log(\frac{2}{\delta})}{n}} + \frac{9 \log(\frac{2}{\delta})}{n}. \]  
(107)
The proof of (22) is completed by replacing \( \delta \) with \( 2\delta \). The proof of (21) is similar, with a Taylor approximation for \( p \) around \( \hat{p}(n) \). □

**APPENDIX IV**

**Proofs for Section V-B**

The proof of Proposition 10 relies on a confidence interval bound for the entropy which is based on an empirical version of \( \zeta(p) \). We begin with the following bound.

**Lemma 19:** Consider the setting of Proposition 10. Then, for any \( \delta \in (0, 1) \)
\[ d_{\text{TV}}(p, \hat{p}(n)) \leq \sqrt{\frac{4 \zeta(p)}{|Y|} + \log (\frac{1}{\delta})}, \]  
(108)
with probability larger than \( 1 - \delta \).

**Proof:** The total variation \( d_{\text{TV}}(p, \hat{p}(n)) \) satisfies a bounded difference inequality with constant \( 1/n \) as a function of \( (Y_1, \ldots, Y_n) \), and so by McDiarmid’s inequality [67, Thm. 3.11]
\[ P \left[ |d_{\text{TV}}(p, \hat{p}(n)) - E[d_{\text{TV}}(p, \hat{p}(n))]| \geq \epsilon \right] \leq e^{-\frac{\epsilon^2}{2n}}. \]  
(109)
Recall that \( \hat{p}(n, y) = \frac{1}{n} \sum_{i=1}^n 1 \{ Y_i = y \} \). We next upper bound the expected value \( E[d_{\text{TV}}(p, \hat{p}(n))] \) as follows:
\[ E[d_{\text{TV}}(p, \hat{p}(n))] = \sum_{y \in Y} E |p(y) - \hat{p}(y, n)| \]
\[ \leq \sum_{y \in Y} \sqrt{E \left[ (p(y) - \hat{p}(y, n))^2 \right]} \]
\[ = \sum_{y \in Y} \sqrt{\frac{2}{n} p(y)(1 - p(y))} \]
\[ \leq \|Y\| \left[ \frac{1}{|Y|} \sum_{y \in Y} \frac{2}{n} p(y)(1 - p(y)) \right] \]
\[ = \frac{2\|Y\|}{n} \left[ \sum_{y \in Y} p(y)(1 - p(y)) \right] \]
\[ = \frac{2\zeta(p)|Y|}{n}, \]  
(110)
where the two inequalities follow from Jensen’s inequality. Setting \( e^{-2n\epsilon^2} = \delta \) directly leads to
\[ d_{\text{TV}}(p, \hat{p}(n)) \leq \sqrt{\frac{2\zeta(p)|Y|}{n} + \frac{1}{2n} \log \left( \frac{1}{\delta} \right)}, \]  
(111)
which is further slightly loosened to the claim of the lemma using \( \sqrt{a + \sqrt{b}} \leq \sqrt{2(a + b)} \) for \( a, b \in \mathbb{R}_+ \). □

Clearly, while \( \zeta(p) \) controls the size confidence interval of \( d_{\text{TV}}(p, \hat{p}(n)) \), it is a distribution-dependent quantity which is unknown to the player, and thus required to be estimated from the data. In this respect, the concentration of \( \zeta(p) \) to its estimated plug-in value is roughly on the same order of that of the total variation (in fact, it can be proved to be faster). Specifically, the following holds:

**Lemma 20:** Let the plug-in estimator of \( \zeta(p) \) be given by \( \hat{\zeta}(n) \equiv \hat{\zeta}(Y, n) \equiv 1 - \sum_{y \in Y} \hat{p}(n, y) \). Then, under the setting of Lemma 19, for any \( \delta \in (0, 1) \)
\[ \hat{\zeta}(n) - \sqrt{\frac{18 \log(\frac{1}{\delta})}{n}} - 1 \leq \zeta(p) \leq \hat{\zeta}(n) + \sqrt{\frac{18 \log(\frac{1}{\delta})}{n}}, \]  
(112)
with probability larger than \( 1 - \delta \).

**Proof:** Since \( \|p(Y, n) + \hat{\zeta}(n) - 1\|^2 \leq \frac{\epsilon}{n} \) for any \( p(Y, n) \in [0, 1] \), the plug-in estimator \( \hat{\zeta}(n) \equiv \hat{\zeta}(Y, n) \) satisfies a bounded difference inequality with constant \( 6/n \) as a function of \( (Y_1, \ldots, Y_n) \), and so by McDiarmid’s inequality [67, Thm. 3.11]
\[ P \left[ |\hat{\zeta}(Y, n) - E[\hat{\zeta}(Y, n)]| \geq \epsilon \right] \leq e^{-\frac{\epsilon^2}{2n}}. \]  
(113)
The plug-in estimator \( \hat{\zeta}(n) \) is biased, and easily seen to satisfy \( E[\hat{\zeta}(n)] = \zeta(p) + \frac{\zeta(p)}{n} \). The result follows since \( \zeta(p) \in [0, 1] \). □
We combine Lemma 19 and Lemma 20 to obtain a confidence interval bound which can be computed by the player according to its empirical data.

**Lemma 21**: Under the setting of Lemma 19, any $\delta \in (0, 1/e)$ it holds that

$$d_{TV}(p, \hat{p}(n)) \leq \sqrt{\frac{4\hat{\zeta}(n)|\mathcal{Y}|}{n} + \frac{\log\left(\frac{2}{\delta}\right)}{n}} + \frac{5|\mathcal{Y}|^{3/4}}{n^{3/4}} \log^{1/4}\left(\frac{2}{\delta}\right),$$

(114)

with probability larger than $1 - \delta$.

**Proof**: By combining Lemma 19 and Lemma 20, and a union bound, it holds with probability larger than $1 - 2\delta$ that

$$d_{TV}(p, \hat{p}(n)) \leq \frac{4\hat{\zeta}(p)|\mathcal{Y}| + \log\left(\frac{2}{\delta}\right)}{n} + \frac{4\hat{\zeta}(n)|\mathcal{Y}| + \sqrt{\frac{18\log\left(\frac{2}{\delta}\right)}{n}}}{n} |\mathcal{Y}| + \log\left(\frac{1}{\delta}\right) + \frac{288|\mathcal{Y}|^2 \log\left(\frac{2}{\delta}\right)}{n^3} + \frac{\log\left(\frac{1}{2\delta}\right)}{n},$$

(115)

where the last inequality follows from $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \in \mathbb{R}_+$. The proof is completed by substituting $\delta$ with $2\delta$.

With these results at hand we may prove Proposition 10.

**Proof of Proposition 10**: As in the proof of Proposition 3 if $d_{TV}(p, \hat{p}(n)) \leq \frac{1}{2}$ then

$$|H(\hat{p}(n)) - H(p)| \leq |\mathcal{Y}| \cdot A_1 \left(\frac{d_{TV}(p, \hat{p}(n))}{|\mathcal{Y}|}\right).$$

(116)

From Lemma 21

$$d_{TV}(p, \hat{p}(n)) \leq \sqrt{\frac{4\hat{\zeta}(n)|\mathcal{Y}|}{n} + \frac{\log\left(\frac{2}{\delta}\right)}{n} + \frac{5|\mathcal{Y}|^{1/2}}{n^{3/4}} \log^{1/4}\left(\frac{2}{\delta}\right)}$$

$: = a_1 + a_2 + a_3,$

(117)

with probability larger than $1 - \delta$, where $\{a_i\}_{i=3}^3$ were implicitly defined. To be in the monotonic increasing regime of $-\Lambda_1(s)$ of $s \in [0, e^{-1}]$, we require that this upper bound is less than $|\mathcal{Y}| e^{-1}$. This can be satisfied if

$$d_{TV}(p, \hat{p}(n)) \leq \frac{4\hat{\zeta}(n)|\mathcal{Y}|}{n\mathcal{Y}} + \frac{\log\left(\frac{2}{\delta}\right)}{n|\mathcal{Y}|^2} + \frac{5}{n^{3/4}} |\mathcal{Y}|^{1/2} \log^{1/4}\left(\frac{2}{\delta}\right)$$

$$\leq \frac{1}{e},$$

(118)

A simple sufficient condition for this can be obtained by bounding $\hat{\zeta}(n) \leq 1$ and $|\mathcal{Y}| \geq 2$, and requiring that each of the three terms is less than a third of $1/e$. This holds if $n \geq 112 \cdot \log\left(\frac{2}{\delta}\right)$ and $\delta \leq 0.2$.

Now, since monotonicity is satisfied, we may replace $d_{TV}(p, \hat{p}(n))$ with the high probability upper bound (114) of Lemma 21. We may consider three cases, according to which of the terms, which we denoted by $\{a_i\}_{i=3}^3$, is the largest.

- If $\max_{i \in [3]} a_i = a_1$ then the upper bound (114) is less than $3a_1 = \sqrt{\frac{36\hat{\zeta}(n)|\mathcal{Y}|}{n}}$. By the monotonicity property, (116) results

$$|H(\hat{p}(n)) - H(p)| \leq \sqrt{\frac{36\hat{\zeta}(n)|\mathcal{Y}|}{n} \log\left(\frac{n|\mathcal{Y}|}{36\hat{\zeta}(n)}\right)}$$

$$\leq 3 \sqrt{\frac{\hat{\zeta}(n)|\mathcal{Y}|}{n} \log\left(\frac{n|\mathcal{Y}|}{36\hat{\zeta}(n)}\right)}.$$  

(119)

- If $\max_{i \in [3]} a_i = a_2$ then the upper bound (114) is less than $3a_2 = \sqrt{\frac{9\log\left(\frac{2}{\delta}\right)}{n}}$. By the monotonicity property, (116) results

$$|H(\hat{p}(n)) - H(p)| \leq \sqrt{9 \log\left(\frac{2}{\delta}\right) \log\left(\frac{n|\mathcal{Y}|^2}{9 \log\left(\frac{2}{\delta}\right)}\right)}$$

$$\leq \frac{3}{2} \sqrt{\log\left(\frac{2}{\delta}\right) \log\left(n|\mathcal{Y}|^2\right)}.$$  

(120)

- If $\max_{i \in [3]} a_i = a_3$ then the upper bound (114) is less than $3a_3 = \frac{15|\mathcal{Y}|^{3/4}}{n^{3/4}} \log^{1/4}\left(\frac{2}{\delta}\right)$. By the monotonic property, (116) results

$$|H(\hat{p}(n)) - H(p)| \leq \frac{15|\mathcal{Y}|^{1/2} \log^{1/4}\left(\frac{2}{\delta}\right)}{n^{3/4}} \log\left(\frac{n|\mathcal{Y}|^{1/2}}{15 \log^{1/4}\left(\frac{2}{\delta}\right)}\right)$$

$$\leq \frac{2|\mathcal{Y}|^{1/2} \log^{1/4}\left(\frac{2}{\delta}\right)}{n^{3/4}} \log\left(n|\mathcal{Y}|^{2/3}\right).$$  

(121)

To agree with all three cases, we sum the three deviation terms, and this completes the proof.

We next turn to the proof of Theorem 11, which is based on a lemma analogous to Lemma 18. To this end, we further denote a simplified version of $\Gamma_n(\cdot)$ from (30) defined as

$$\tilde{\Gamma}_n(\alpha, \zeta, \Delta, t)$$

$$:= \max\left\{\frac{144\zeta}{|\mathcal{Y}|^2} A_1^2 \left(\frac{2|\mathcal{Y}|}{3\Delta}\right), \frac{135}{|\mathcal{Y}|^2} A_2 \left(\frac{9|\mathcal{Y}|^{2\alpha} \log(t)}{\Delta^2}\right), \frac{3}{|\mathcal{Y}|^{2/3}} A_4^{1/3} \left(\frac{27|\mathcal{Y}|^{1/3} \alpha^{1/3} \log^{1/3}(t)}{\Delta^{4/3}}\right)\right\}.$$  

(122)

**Lemma 22**: For $\delta \equiv \delta_n(t) = 2t^{-\alpha}$ and $\alpha > 2$ if $n \geq \tilde{\Gamma}_n(\alpha, \zeta, \Delta, t)$ then $\text{UCD}_n(\zeta, \delta, \mathcal{Y}, n) \leq \Delta/2$, where $\text{UCD}_n(\zeta, \delta, \mathcal{Y}, n)$ is as defined in (28).
Proof: We may assume\(^3\) that \(n \geq e\). Then, UCD\(_n\)(\(\zeta, |\mathcal{Y}|, \delta, n\)) \(\leq \Delta/2\) if all three conditions hold\(^4\)

\[
3 \sqrt{\frac{|\mathcal{Y}|}{n} \log \left( \frac{n|\mathcal{Y}|}{36\zeta} \right)} \leq \Delta/6, \tag{123}
\]

and

\[
3 \sqrt{\frac{\log \left( \frac{2}{\delta} \right)}{n} \log \left( \frac{n|\mathcal{Y}|^2}{9} \right)} \leq \Delta/6, \tag{124}
\]

as well as

\[
2 |\mathcal{Y}|^{1/2} \log^{1/4} \left( \frac{2}{\delta} \right) \log \left( \frac{n|\mathcal{Y}|^{2/3}}{n^{3/4}} \right) \leq \Delta/6. \tag{125}
\]

For the first condition, we write it equivalently as

\[
\log \left( \frac{n|\mathcal{Y}|^{2/3}}{36\zeta} \right) \leq \frac{3\Delta}{2|\mathcal{Y}|}. \tag{126}
\]

Lemma 15 then implies that this condition is satisfied if

\[
n \geq 144 \cdot \frac{\sqrt{\zeta}}{|\mathcal{Y}|} \Lambda_1 \left( \frac{2|\mathcal{Y}|}{3\Delta} \right). \tag{127}
\]

For the second condition, we write it equivalently as

\[
\log^2 \left( \frac{n|\mathcal{Y}|^2}{9} \right) \leq \frac{\Delta^2}{9|\mathcal{Y}|^2 \log \left( \frac{2}{\delta} \right)}. \tag{128}
\]

Lemma 15 implies that this condition is satisfied if

\[
n \geq 135 \cdot \frac{\sqrt{\zeta}}{|\mathcal{Y}|^2} \Lambda_2 \left( \sqrt{\frac{9|\mathcal{Y}|^2 \log \left( \frac{2}{\delta} \right)}{\Delta^2}} \right). \tag{129}
\]

For the last condition, we first require a slightly stronger condition (in terms of the numerical constant)

\[
\log^{4/3} \left( \frac{n|\mathcal{Y}|^{2/3}}{27|\mathcal{Y}|^{4/3} \log^{1/3} \left( \frac{2}{\delta} \right)} \right) \leq \frac{\Delta^{4/3}}{27|\mathcal{Y}|^{4/3} \log^{1/3} \left( \frac{2}{\delta} \right)} \tag{130}
\]

Lemma 15 implies that this condition is satisfied if

\[
n \geq \frac{3}{|\mathcal{Y}|^{2/3}} \Lambda_4 \left( \frac{27|\mathcal{Y}|^{4/3} \log^{1/3} \left( \frac{2}{\delta} \right)}{\Delta^{4/3}} \right). \tag{131}
\]

The claim of the lemma then follows from the definition of \(\hat{\Gamma}_n(\cdot)\) in (122).

We may now prove Theorem 11.

Proof of Theorem 11:

The proof begins as the proof of Theorem 6. We then define the events:

I'. Either the entropy of the best arm is significantly underestimated

\[
\hat{H}(X_i(t-1), N_i(t-1)) + \text{UCD}(X_i(t-1), \delta_{\alpha}(t), N_i(t-1)) \leq H_i, \tag{134}
\]

II'. Either the entropy of arm \(i\) is significantly overestimated

\[
\hat{H}(X_i(t-1), N_i(t-1)) \leq H_i + \text{UCD}(X_i(t-1), \delta_{\alpha}(t), N_i(t-1)), \tag{135}
\]

III'. The upper confidence interval, which is based on an overestimation of \(\hat{\zeta}(X_i(t-1), N_i(t-1))\) is significantly larger than the gap:

\[
\text{UCD}_n \left( \zeta(p_i) + \sqrt{\frac{18 \log \left( \frac{1}{\delta} \right)}{N_i(t-1)}}, \delta_{\alpha}(t), N_i(t-1) \right) \geq \frac{\Delta_i}{2}, \tag{136}
\]

or

\[
\hat{\zeta}(X_i(t-1), N_i(t-1)) - \zeta(p_i) \geq \frac{18 \log \left( \frac{1}{\delta} \right)}{N_i(t-1)} + \frac{1}{N_i(t-1)}. \tag{137}
\]

As in the proof of Theorem 2, if all three events I'-.III' are false, then

\[
\hat{H}(X_i(t-1), N_i(t-1)) + \text{UCD}(X_i(t-1), \delta_{\alpha}(t), N_i(t-1)) \geq H_i, \tag{138}
\]

or

\[
N_i(t-1) \leq \max \left( 30 \cdot \log \left( \frac{2}{\delta} \right), 119\zeta(p_i)|\mathcal{Y}| \right). \tag{139}
\]

\(^3\)This assumption can be easily achieved if the player plays each arm 3 times at the beginning of Algorithm 1.

\(^4\)Since there are three terms involved, we do not over-complicate the analysis with additional parameter \(\beta\) (see the proof of Lemma 18).
max \left\{ 144 \frac{\zeta(p_i)}{|Y|} \Lambda_i^2 \left( \frac{2|Y|}{3 \Delta_i} \right) + 576 \frac{18 \log \left( \frac{1}{\epsilon} \right)}{|Y|^2} \sqrt{N_i(t - 1)} \Lambda_i^2 \left( \frac{2|Y|}{3 \Delta_i} \right), \right.
\left. \frac{135}{|Y|^2} \Lambda_i^2 \left( \frac{9|Y|^2 \alpha \log(t)}{\Delta_i^2} \right), \frac{3}{|Y|^2/3} \Lambda_i^{4/3} \left( \frac{27|Y|^{4/3} \alpha^{1/3} \log^{1/3}(t)}{\Delta_i^{4/3}} \right) \right\}. \tag{132}

N_i(t - 1) \geq \max \left\{ 288 \frac{\zeta(p_i)}{|Y|} \Lambda_i^2 \left( \frac{2|Y|}{3 \Delta_i} \right), 36230 \frac{\alpha^{1/3} \log^{1/3}(t)}{|Y|^2/3} \Lambda_i^{4/3} \left( \frac{2|Y|}{3 \Delta_i} \right), \right.
\left. \frac{135}{|Y|^2} \Lambda_i^2 \left( \frac{9|Y|^2 \alpha \log(t)}{\Delta_i^2} \right), \frac{3}{|Y|^2/3} \Lambda_i^{4/3} \left( \frac{27|Y|^{4/3} \alpha^{1/3} \log^{1/3}(t)}{\Delta_i^{4/3}} \right) \right\}. \tag{133}

where the last inequality follows since \( \log(1 - x) \leq -x \).

**Lemma 24:** Let \( n \geq 1 \), then
\[
\mathbb{P} \left( |\hat{S}(Y, n) - E(\hat{S}(Y, n))| \geq \epsilon \right) \leq \exp \left( -2 \epsilon^2 \right). \tag{145}
\]

**Proof:** This follows directly from McDiarmid’s inequality and since changing the outcome of one sample changes \( \hat{S}(Y, n) \) by at most one.

We proceed to prove Lemma 12.

**Proof of Proposition 12:** We wish to bound the probability \( \mathbb{P}(|S(Y, n) - S(p)| > \tau) \). By the triangle inequality
\[
\mathbb{P}(|S(Y, n) - S(p)| > \tau) \leq \mathbb{P} \left( |S(Y, n) - \mathbb{E}(\hat{S}(Y, n))| + |\mathbb{E}(\hat{S}(Y, n)) - S(p)| > \tau \right). \tag{146}
\]

Thus, we next upper bound the gap \( |\mathbb{E}(\hat{S}(Y, n)) - S(p)| \) and the probability \( \mathbb{P}(|\hat{S}(Y, n) - \mathbb{E}(\hat{S}(Y, n))| > \tilde{\tau}) \). By Lemma 23
\[
|\mathbb{E}(\hat{S}(Y, n)) - S(p)| \leq S(p) \exp \left( -\frac{n}{K} \right). \tag{147}
\]

Additionally, we can upper bound the probability \( \mathbb{P}(|\hat{S}(Y, n) - \mathbb{E}(\hat{S}(Y, n))| > \tau - S(p) \exp \left( -\frac{n}{K} \right) \) using Lemma 24. We then substitute \( \tau = \sqrt{\frac{1}{2} \log \left( \frac{1}{\delta} \right) + S(p) \exp \left( -\frac{n}{K} \right)} \) to conclude the proof.

We are now ready to prove Proposition 13.

**Proof of Proposition 13:**

From Proposition 12, it holds that the event
\[
\mathcal{E} := \left\{ \hat{S}(Y, n) + \sqrt{\frac{1}{2} \log \left( \frac{1}{\delta} \right)} \geq 1 - e^{-n/2} \right\} \tag{148}
\]
occurs with probability \( \mathbb{P}[\mathcal{E}] \geq 1 - \delta \). Thus,
\[
\mathbb{P} \left[ |H(\hat{p}(n)) - H(p)| \geq B_{SE}(Y, \delta, n) \right] \leq \mathbb{P} \left[ |H(\hat{p}(n)) - H(p)| \geq B_{SE}(Y, \delta, n) \right].
\]
Next, we present an upper bound on the number maximal number of rounds, i.e., n, such that \( \text{UCD}_{SE}(Y, t^{-\alpha}, n) \leq \Delta/2 \).

**Lemma 25:** Let a support \( S \) such that \( |S| = S \leq \kappa \) be given, let a gap \( \Delta \in (0, \log(|S|)) \) be given, and let \( \delta = t^{-\alpha} \).

Then, for any \( \beta \in (0, 1) \), if \( n \geq \Gamma_{SE}(\alpha, \beta, S, \kappa, \Delta, t) \) then \( \text{UCD}_{SE}(Y, t^{-\alpha}, n) \leq \Delta/2 \) for every \( Y \in S^n \).

**Proof:** By the definition of \( \text{UCD}_{SE}(Y, \delta, n) \) we have that if

\[
\text{BSE}(Y, \delta, n) \leq \beta \cdot \frac{\Delta}{2},
\]

and

\[
\sqrt{\frac{2 \log^2(n)}{n} \log\left(\frac{2}{\delta}\right)} \leq (1 - \beta) \cdot \frac{\Delta}{2},
\]

then \( \text{UCD}_{SE}(Y, \delta, n) \leq \Delta_i/2 \).

First, we analyze the first condition, i.e., (150). Since \( \hat{S}(Y, n) \leq S \), it is fulfilled whenever

\[
\log\left(1 + \frac{1}{n} \left[(S + \frac{1}{2} \log\left(\frac{1}{\delta}\right)) \cdot (1 - e^{-\frac{\alpha}{2}})^{-1} - 1\right]\right)
\]

\[
\leq \beta \cdot \frac{\Delta}{2}.
\]

Further algebra yields that \( \text{BSE}(Y, \delta, n) \leq \beta \cdot \frac{\Delta}{2} \) for all n such that

\[
\frac{1}{e^{\beta \Delta/2} - 1} \left(S + \frac{1}{2} \log\left(\frac{1}{\delta}\right)\right) \leq \left(n + \frac{1}{e^{\beta \Delta/2} - 1}\right) (1 - e^{-\frac{\alpha}{2}}).
\]

Next, utilizing the bound \( 1 - e^{-x} \geq \frac{x}{2} \) which holds for every \( x \in [0, 1] \), we replace the condition (153) with the following stricter condition

\[
\frac{1}{e^{\beta \Delta/2} - 1} \left(S + \frac{1}{2} \log\left(\frac{1}{\delta}\right)\right) \leq \left(n + \frac{1}{e^{\beta \Delta/2} - 1}\right) \left(1 - e^{-\frac{\alpha}{2}}\right).
\]

Substituting \( \delta = t^{-\alpha} \) and replacing the RHS of (154), i.e. \( n + \frac{1}{e^{\beta \Delta/2} - 1} \frac{\alpha}{2} \), with the stricter condition \( \frac{n}{\kappa} \) on \( n \) yields that (154) is fulfilled whenever

\[
n \geq \frac{\sqrt{2} (S + \sqrt{\frac{\alpha}{2} \log(t)})}{e^{\beta \Delta/2} - 1}, \quad \text{and} \quad \frac{n}{\kappa} \leq 1.
\]

Note that since the RHS of (153) is monotonically increasing with \( n \), it is sufficient to find \( n_0 \) that fulfills (155) to conclude that every \( n \geq n_0 \) fulfills (153) too.

Now, we consider the case where \( n \geq \kappa \). In this scenario \( 1 - e^{-\frac{\alpha}{2}} \geq 1 - e^{-1} \geq 1/2 \). Plugging this bound, we have that (153) is fulfilled in this case whenever

\[
n \geq 2 \cdot \frac{S + \sqrt{\alpha \log(t)/2}}{e^{\beta \Delta/2} - 1}. \quad \text{Overall, we have that (153) is fulfilled for}
\]

\[
n \geq 2 \cdot \left\lfloor \max \left\{ \kappa, \frac{S + \sqrt{\alpha \log(t)/2}}{e^{\beta \Delta/2} - 1} \right\} \cdot \frac{S + \sqrt{\alpha \log(t)/2}}{e^{\beta \Delta/2} - 1} \right\rfloor.
\]

Since \( \sqrt{\alpha \log(t)/2} < \kappa \), we can simplify the bound and conclude that (153) is fulfilled by all \( n \) such that

\[
n \geq \frac{2 \sqrt{\kappa} \left(\sqrt{S + (\alpha \log(t)/2)^{\frac{1}{2}}}\right)}{\min\{e^{\beta \Delta/2} - 1, \sqrt{e^{\beta \Delta/2} - 1}\}}.
\]

Now, we focus on the second condition, i.e., (151). Observe that the second condition is exactly (60) which is investigated in the proof of Lemma 16. Recall the notation (1), it follows that the second condition holds for all \( n \) such that

\[
n \geq 15 \cdot \Lambda_2 \left(\frac{8 \cdot \log(2^{\alpha})}{(1 - \beta^2 \Delta^2)}\right).
\]
II". The entropy of arm $i$ is significantly overestimated:

$$H_{\text{SE}}(X_i(t-1), \delta_i(t), N_i(t-1)) > H_i + UCD_{\text{SE}}(X_i(t-1), \delta_i(t), N_i(t-1)).$$

(160)

III". The upper confidence interval is significantly larger than the gap

$$UCD_{\text{SE}}(X_i(t-1), \delta_i(t), N_i(t-1)) > \Delta_i/2.$$  (161)

As in the proof of Theorem 2, if all three events $I"$-$III"$ are false, then

$$H_{\text{SE}}(X_i(t-1), N_i(t-1)) + UCD_{\text{SE}}(X_i(t-1), \delta_i(t), N_i(t-1)) > H_i = H_i + \Delta_i$$

$$\geq H_i + 2 \cdot UCD_{\text{SE}}(X_i(t-1), \delta_i(t), N_i(t-1))$$

$$\geq H_{\text{SE}}(X_i(t-1), N_i(t-1)) + UCD_{\text{SE}}(X_i(t-1), \delta_i(t), N_i(t-1)),$$

(162)

which contradicts the assumption that Algorithm 1 chooses $I_i = i$ at the $t$th round.

By Lemma 25, if $N_i(t-1) \geq \Gamma_i(N, S_i, \kappa, \Delta, t)$ then

$$UCD_{\text{SE}}(X_i(t-1), t^{-\alpha}, N_i(t-1)) \leq \Delta_i/2$$

for every $X_i(t-1) \in S_i^{N_i}(t-1)$, that is, the event III" does not occur. The analysis then follows as in the proof of Theorem 2. Specifically, we upper bound the probability that the events $I$" and II" occur based on Proposition 13 as follows:

$$\sum_{\tau=1}^{t} \mathbb{P}(I" \text{ is true at round } \tau) + \mathbb{P}(\text{II" is true at round } \tau) \leq 2 \sum_{\tau=1}^{t} \frac{2}{\alpha-1} \leq \frac{4(\alpha - 1)}{\alpha - 2}.$$  (163)

\section*{ACKNOWLEDGMENT}

The authors would like to thank the anonymous reviewers for their careful reading of the article and their constructive suggestions that helped to improve the manuscript, and for pointing out the possibility of developing a KL-UCB algorithm for the IMAB problem. In addition, the author Nir Weinberger would also like to thank Itay Ron for multiple discussions at the early stages of this research.

\section*{REFERENCES}

[1] N. Weinberger and M. Yemini, “Upper confidence interval strategies for multi-armed bandits with entropy rewards,” in Proc. IEEE Int. Symp. Inf. Theory (ISIT), Jun. 2022, pp. 1647–1652.

[2] T. L. Lai and H. Robbins, “Asymptotically efficient adaptive allocation rules,” Adv. Appl. Math., vol. 6, no. 1, pp. 4–22, Mar. 1985.

[3] P. Rusmevichientong and J. N. Tsitsiklis, “Linearly parameterized bandits,” Math. Oper. Res., vol. 35, no. 2, pp. 395–411, May 2010.

[4] M. K. Hanawal, A. Leshem, and V. Saligrama, “Efficient algorithms for linear polyhedral bandits,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2016, pp. 4796–4800.

[5] V. Dani, T. P. Hayes, and S. M. Kakade, “Stochastic linear optimization under bandit feedback,” in Proc. Conf. Learn. Theory (COLT), 2008, pp. 355–366.

[6] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, “Improved algorithms for linear stochastic bandits,” in Proc. Adv. Neural Inf. Process. Syst., vol. 24, 2011, pp. 2312–2320.

[7] V. Anantharam, P. Varaiya, and J. Walrand, “Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays—Part II: Markovian rewards,” IEEE Trans. Autom. Control, vol. AC-32, no. 11, pp. 977–982, Nov. 1987.

[8] T. Jaksch, R. Ortner, and P. Auer, “Near-optimal regret bounds for reinforcement learning,” J. Mach. Learn. Res., vol. 11, pp. 1563–1600, Aug. 2010.

[9] P. L. Bartlett and A. Tewari, “REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs,” in Proc. UAI, 2009, pp. 35–42.

[10] R. Fruit, M. Pirotta, and A. Lazaric, “Near optimal exploitation-exploitation in non-communicating Markov decision processes,” in Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS), 2018, pp. 2998–3008.

[11] M. Yemini, A. Leshem, and A. Somekh-Baruch, “The restless hidden Markov bandit with linear rewards and side information,” IEEE Trans. Signal Process., vol. 69, pp. 1108–1123, 2021.

[12] T. Gafni, M. Yemini, and K. Cohen, “Restless multi-armed bandits under exogenous global Markov process,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2022, pp. 5218–5222.

[13] W. Chen, W. Hu, F. Li, J. Li, Y. Liu, and P. Lu, “Sequential multi-armed bandit with general reward functions,” in Proc. Adv. Neural Inf. Process. Syst., vol. 29. Red Hook, NY, USA: Curran Associates 2016, pp. 1–9.

[14] T. M. Cover, “Universal gambling schemes and the complexity measures of Kolmogorov and Chaitin,” Stanford Univ., Stanford, CA, USA, Tech. Rep. 12, 1974.

[15] T. Cover and R. King, “A convergent gambling estimate of the entropy of English,” IEEE Trans. Inf. Theory, vol. IT-24, no. 4, pp. 413–421, Jul. 1978.

[16] H. Dehghan and H. Ghasemian, “Measurement of uncertainty by the entropy: Application to the classification of MSS data,” Int. J. Remote Sens., vol. 27, no. 18, pp. 4005–4014, Sep. 2006.

[17] A. Posadas, J. Morales, and A. Posadas-Genaro, “Earthquakes and entropy: Characterization of occurrence of earthquakes in southern Spain and Alboran Sea,” Chaos, Interdiscip. J. Nonlinear Sci., vol. 31, no. 4, Apr. 2021, Art. no. 043124.

[18] M. Haider Sharif and C. Djeraba, “An entropy approach for abnormal activities detection in video streams,” Pattern Recognit., vol. 45, no. 7, pp. 2543–2561, Jul. 2012.

[19] R. Hu, H. Pham, P. Bulushech, and D. Gatica-Perez, “Elderly people living alone: Detecting home visits with ambient and wearable sensing,” in Proc. 2nd Int. Workshop Multimedia Pers. Health Health Care, Mountain View, CA, USA, Oct. 2017, pp. 85–88.

[20] A. Howedi, A. Lotfi, and A. Pourabdollah, “An entropy-based approach for anomaly detection in activities of daily living in the presence of a visitor,” Entropy, vol. 22, no. 8, p. 845, Jul. 2020.

[21] C. Callegari, S. Giordano, and M. Pagano, “Entropy-based network anomaly detection,” in Proc. Int. Conf. Comput., Netw. Commun. (ICNC), Jan. 2017, pp. 334–340.

[22] R. Zhou, R. Cai, and G. Tong, “Applications of entropy in finance: A review,” Entropy, vol. 15, no. 12, pp. 4909–4931, Nov. 2013.

[23] T. M. Cover and J. A. Thomas, Elements of Information Theory. Hoboken, NJ, USA: Wiley, 2006.

[24] A. Howedi, B. S. Roy, A. Candela, and D. Wettergreen, “Science-aware exploration using entropy-based planning,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Sep. 2017, pp. 3819–3825.

[25] A. Dai, S. Papatheodorou, N. Funk, D. Tzoumanikas, and S. Leutenegger, “Fast frontier-based information-driven autonomous exploration with an MAV,” in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), May 2020, pp. 9570–9576.

[26] M. Scott and K. Jerath, “Multi-robot exploration and coverage: Entropy-based adaptive maps with adjacency control laws,” in Proc. Ann. Amer. Control Conf. (ACC), Jun. 2018, pp. 4403–4408.

[27] E. K. Morris et al., “Choosing and using diversity indices: Insights for ecological applications from the German biodiversity exploratories,” Ecol. Evol., vol. 4, no. 18, pp. 3514–3524, Sep. 2014.

[28] W. B. Sherwin, “Entropy and information approaches to genetic diversity and its expression: Genomic geography,” Entropy, vol. 12, no. 7, pp. 1765–1798, Jul. 2010.

[29] P. Auer, “Using confidence bounds for exploitation-exploration tradeoffs,” J. Mach. Learn. Res., vol. 3, pp. 397–422, Nov. 2002.
