Upper Confidence Interval Strategies for Multi-Armed Bandits with Entropy Rewards

Nir Weinberger*[1] and Michal Yemini*[2]

Abstract

We present the informational multi-armed bandit (IMAB) model for a multi-armed bandit problem with an information-based reward. 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. It is well known that there is no unbiased estimator for the entropy functional with finite variance. Therefore, the expected regret guarantees of the current upper confidence bounds (UCB) algorithm cannot be readily applied. To that end, we propose two UCB-based algorithms 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. We provide performance guarantees by upper bounding the expected regret of each of the algorithms. Additionally, we compare their asymptotic behavior to the Lai-Robbins lower bound for the pseudo regret. Finally, we provide numerical results that illustrate the expected regret of the algorithms we present in this paper.

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. At each round, 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

* Equal contribution.
[1] Nir Weinberger is with the Faculty of Electrical and Computer Engineering, Technion - Israel Institute of Technology. (Email: nirwein@technion.ac.il.)
[2]Michal Yemini is with the Faculty of Electrical and Computer Engineering, Princeton University. (Email: myemini@princeton.edu.)

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 [1], 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 [2]–[5], Markovian dynamics and rewards [6]–[11], and combinatorial bandits with monotone reward functions [12]. 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 *informativness* 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. Clearly, there could be different ways to measure this quantity, and in this work we focus on the natural choice of 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 anomaly detection [13]–[16]. We thus henceforth refer to a MAB problem with entropy rewards as *informational* MAB (IMAB). 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 $-\log p_i(x)$. 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.

A highly successful algorithmic approach in MAB problem 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 an upper confidence bounds (UCB) [17], [18]. The performance guarantees of the expected regret of UCB algorithms are achieved by utilizing concentration inequalities (see primers in [19], [20]), 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 [21]. However, the bias of the plug-in estimator is upper bounded by $\log(1+\frac{|\mathcal{X}|}{n})$ where $|\mathcal{X}|$ is the alphabet size and n is the number of samples used for estimation [21]. 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 [22] resulted in a concentration inequality bound for the plug-in entropy estimator w.r.t. its (biased) mean. Therefore, our first approach for confidence interval to entropy estimators 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 total variation bound on the entropy difference of a pair of PMFs [23], [24]. 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 $\zeta(p)|\mathcal{X}|$). We then show that $\zeta(p)$ itself can be estimated from data, and this estimate can be used in a UCB algorithm in lieu of the true value.

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, as input, a choice of an entropy estimator, and a choise of an upper confidence bound (on that estimator). 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 standard UCB bound (for mean-based rewards), yet only after a large number of rounds $O(\exp\sqrt{|\mathcal{X}|})$ where here $\mathcal{X}$ is the maximal alphabet size of the arms. Then, in Sec. V we present a UCB algorithm which 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 is 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 variance-UCB, 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. In addition, this setting can also be compared with the Lai-Robbins lower bound [1, Thm. 1], which reveals the strength and possible weaknesses of the UCB algorithm. We then extend the analysis to alphabets of arbitrary size. In Sec. VI we provide a few numerical examples, and in Sec. VII we summarize the paper.

Related works: 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 [21]–[23], [25]–[32]. 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 in to decrease the expected regret of several MAB models [33]–[35]. Using the mutual information of the probabilities for arm sampling at two consecutive rounds, information-directed sampling (IDS) outperforms both UBC based algorithms and Thompson sampling [36], [37] in problems with special structures, such as dependent on prior models [33], and bandits with heteroscedastic noise which is arm dependent [34]. Nonetheless, these works utilize the informational measures to create exploration-exploitation trade-offs, rather than the reward structure. 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 [12], [38]. However, the work [38] is limited to a known parametric family of distributions with unknown parameters. Moreover, [12] relies on stochastically dominant confidence bounds that requires monotonically increasing instantaneous reward function, however the self-information $-\log(p_i(x))$ is monotonically decreasing in $p_i(x)$. Furthermore, it is assumed in [12], [38] that the instantaneous reward is directly known, however, in our case the instantaneous reward is unknown to the player and is not a function of the symbol outcome but rather of its probability.

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 $\max a, b := a \vee b$ and $\min a, b := a \wedge b$, as well as $(t)_+ = \max\{t, 0\}$. To focus the reader on the first-order terms we denote the linear-times-polylogarithmic function by

$$\Lambda_k(s) := s \log^k s. \tag{1}$$

For a discrete alphabet $\mathcal{Y}$, we denote the entropy of a PMF p over an alphabet $\mathcal{Y}$ by $H(p) := -\sum_{y\in\mathcal{Y}} p(y) \log p(y)$, where logarithms are arbitrarily taken to the natural base. We denote the *total variation distance* between two PMFs p and q on a finite alphabet $\mathcal{Y}$ by $\mathrm{d_{TV}}(p, q) := \sum_{y\in\mathcal{Y}} |p(y) - q(y)|$. We note in passing that in what follows, as customary, we have opted for the simplicity of the bounds rather than 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 alphabet $\mathcal{X}_i$, such that $p_i(x) := \mathbb{P}[X_i = x]$. We further denote by $p_i = \{p_i(x)\}_{x\in\mathcal{X}_i}$ the PMF of the ith source. We consider 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 t-th symbol $X_i(t)$ from that source. In the context of MAB, each of the sources is refereed 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* $-\log p_i(X_i(t))$, and so the expected reward of sampling only arm i is $-\mathbb{E}[\log p_i(X)] = H(p_i) := H_i$, which is the entropy of the ith source. The goal of the player is to choose the arm with the maximal expected reward, that is, the maximal entropy, $i^* \in \arg\max_{i\in[K]} H_i$. The player, which does not know in advance the probabilities p_i (and so also not the entropy values H_i) estimates the expected reward H_i of each arm from its previous actions and observations.

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. As a measure for the performance of the policies used by the player, we will adopt the standard expected *pseudo-regret* [18, Ch. 1]

$$R(t) := t \cdot H_{i^*} - \sum_{i\in[K]} \mathbb{E}(N_i(t)) \cdot H_i. \tag{2}$$

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

$$R(t) = \sum_{i\in[K]:\Delta_i>0} \mathbb{E}(N_i(t))\Delta_i. \tag{3}$$

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 [18, Sec. 2.2] [39, 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 $\boldsymbol{Y} := \{Y_\ell\}_{\ell\in[n]}$ be n IID samples from a PMF p over a finite alphabet $\mathcal{Y}$. Suppose that there exists an entropy estimator $\hat{H}(\boldsymbol{Y}, n)$ and an upper confidence interval function $\mathrm{UCB}(\boldsymbol{Y}, n, \delta)$ (for $\delta \in (0, 1)$) for which the upper confidence bound

$$H(p) \leq \hat{H}(\boldsymbol{Y}, n) + \mathrm{UCB}(\boldsymbol{Y}, \delta, n) \tag{4}$$

holds with probability larger than $1 - \delta$. Note that both the estimator $\hat{H}(\boldsymbol{Y}, n)$ and the confidence interval $\mathrm{UCB}(\boldsymbol{Y}, \delta, n)$ may depend on the observed source symbols $\boldsymbol{Y}$. 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 interval for each of the arms. The played arm is then the one maximizing the estimated entropy plus the confidence interval, 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 $\{\mathcal{X}_i\}_{i\in[K]}$;
- A sequence of entropy estimators $\{\hat{H}(\cdot, n)\}_{n\in\mathbb{N}_+}$ and a sequence of upper confidence interval functions $\{\mathrm{UCB}(\cdot, \cdot, n)\}_{n\in\mathbb{N}_+}$;
- 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, $\boldsymbol{X}_i(t)$ is the set of samples available to the player at round t from the ith arm.

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

Algorithm 1 A general UCB-entropy algorithm

procedure AN UPPER CONFIDENCE BOUND ALGORITHM(K, $\{\mathcal{X}_i\}_{i\in[K]}$, $\hat{H}(\cdot,n)$,UCB$(\cdot,\cdot,n)$, α, $\delta_\alpha(t)$)
  set $\boldsymbol{X}_i(0)=\phi$ and $N_i(0)=0$ for all $i\in[K]$
    ▷ The observation set of each arms is empty at round $t=0$
  for $t=1,2,\ldots$ **do**
    play $I(t)\in\arg\max_{i\in[K]}\{\hat{H}(\boldsymbol{X}_i(t-1),N_i(t-1)))+\mathrm{UCB}(\boldsymbol{X}_i(t-1),\delta_\alpha(t),N_i(t-1))\}$
    set $\boldsymbol{X}_{I(t)}(t)=\boldsymbol{X}_{I(t)}(t-1)\cup X_{I(t)}(t)$ and $N_{I(t)}(t)=N_{I(t)}(t-1)+1$
      ▷ The observation of the chosen arm is added to the set of observations
    set $\boldsymbol{X}_i(t)=\boldsymbol{X}_i(t-1)$ and $N_i(t)=N_i(t-1)$ for all $i\in[K]\backslash I(t)$
      ▷ The observation set of others arms is unchanged
  end for
  return $\{N_i(t)\}_{i\in[K],\,t\in\mathbb{N}_+}$
    ▷ The number of times each arm $i\in[K]$ have been played up to each round $t\in\mathbb{N}_+$
end procedure

Thm.	Alphabet	$\hat{H}(\boldsymbol{Y},n)$	$\delta_\alpha(t)$	UCB$(\boldsymbol{Y},\delta,n)$	PMF based UCB	Regret of arm i
Thm. 2	discrete, finite	$H(\hat{p}(n))+B(n)$	$t^{-\alpha}$	(6)	no	(9)
Thm. 5	binary	$H(\hat{p}(n))$	$6t^{-\alpha}$	(11)	yes	(17)
Thm. 8	binary	$H(\hat{p}(n))$	$4t^{-\alpha}$	(19)	yes	(23)
Thm. 10	discrete, finite	$H(\hat{p}(n))$	$t^{-\alpha}$	(26)	yes	(28)

TABLE I
SUMMARY OF INPUTS FOR ALGORITHM 1 AND RESULTS.

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 [22], yet suffers from a negative bias [21]. Thus, a natural method of obtaining an upper confidence bound is by a correction of this bias. Specifically, let $\boldsymbol{Y}=\{Y_\ell\}_{\ell\in[n]}$ be as sequence of IID samples from some distribution p over the alphabet $\mathcal{Y}$, and let (with a slight abuse of notation) $\hat{p}(n)=\frac{1}{n}\{\hat{p}(y,n)\}_{y\in\mathcal{Y}}$ be the empirical mean of the n samples, where $\hat{p}(y,n):=\sum_{\ell=1}^{n}\mathbb{1}\{Y_\ell=y\}$ for all $y\in\mathcal{Y}$. Then, the plug-in estimator $H(\hat{p}(n))$ is biased, and, as was proved in [21],

$$H(p)-B(n)\leq\mathbb{E}[H(\hat{p}(n))]\leq H(p), \tag{5}$$

where for $B(n) := \log(1 + \frac{|\mathcal{Y}|-1}{n})$ for $n \geq 1$. Therefore, the bias-corrected estimator $H(\hat{p}(n)) + B(n)$ has a nonnegative bias. Let

$$\mathrm{UCB}_{\text{bias}}(\delta, n) := B(n) + \sqrt{\frac{2\log^2(n)}{n} \log\left(\frac{2}{\delta}\right)}. \tag{6}$$

The concentration result of the plug-in estimator from [22, p. 168] implies the following confidence interval bound:

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

$$|H(\hat{p}(n)) + B(n) - H(p)| \leq \mathrm{UCB}_{\text{bias}}(\delta, n), \tag{7}$$

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

$$\Gamma_{\text{bias}}(\alpha, \beta, \mathcal{Y}, \Delta, t) := \max\left\{\frac{|\mathcal{Y}|-1}{e^{\beta\cdot\Delta/2} - 1},\ 15 \cdot \Lambda_2\left(\frac{8 \cdot \log(2t^\alpha)}{(1-\beta)^2\Delta^2}\right)\right\}. \tag{8}$$

Theorem 2. *Assume that Algorithm 1 is run with a bias-corrected plug-in entropy estimator $\hat{H}(\boldsymbol{Y}, n) \equiv H(\hat{p}(n)) + B(n)$, and upper confidence interval $\mathrm{UCB}(\boldsymbol{Y}, \delta, n) \equiv \mathrm{UCB}_{\text{bias}}(\delta, n)$ with $\delta \equiv \delta_\alpha(t) = t^{-\alpha}$ and $\alpha > 2$. Let $\beta \in (0,1)$ be given. Then, the pseudo-regret is bounded as*

$$R(t) \leq \sum_{i \in [K]:\Delta_i > 0} \left[\Gamma_{\text{bias}}(\alpha, \beta, \mathcal{X}_i, \Delta_i, t) \cdot \Delta_i + \frac{2(\alpha-1)}{\alpha - 2} \cdot \Delta_i\right]. \tag{9}$$

For example, one can arbitrary choose $\beta = 1/2$, and then there exists a t_0, which only depends on $\{\mathcal{X}_i\}_{i\in[K]}$ and α and hence is known in advance, so that the regret bound holds for all $t \geq t_0$. We may note that the bound on the regret of Algorithm 1 with a bias corrected entropy estimator in (9) is comprised of a few terms. However, there is only a single term which blows-up as $\Delta_i \downarrow 0$ for some $i \in [K]$, which is given by

$$\frac{c_1 \cdot \log(t)}{\Delta_i} \cdot \log^2\left(\frac{c_2 \cdot \log(t)}{\Delta_i^2}\right), \tag{10}$$

for some constants c_1, c_2. Thus, the regret scales as $\tilde{O}(\frac{\log(t)}{\Delta_i})$, where the only difference from the standard UCB [18, 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 Δt is always an upper bound on the pseudo-regret, we may obtain the $\tilde{O}(\min\{\frac{1}{\Delta}\}, \Delta t) = \tilde{O}(\sqrt{t})$, which roughly matches the gap-independent

bound in the standard MAB problem.

Nonetheless, from a different perspective, assuming that the gaps are all constants, then if $\log^2(t) = \tilde{O}(|\mathcal{X}_i|)$ the regret will be determined by the first term in (8), and so the regret bound is large as long as $t = O(\max_{i\in[K]} \exp(\sqrt{|\mathcal{X}_i|}))$. In the next section we develop a UCB algorithm which 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 $\mathcal{X}_i$. A natural question is therefore, 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 variance-UCB [40], [41] 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 UCB. 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 consider Bernoulli arms for which $p_i(1) = \mathbb{P}[X_i = 1] \ll 1$ for some arm $i \in [K]$. The entropy of this arms is much smaller than the maximal possible value of $\log|\mathcal{X}_i| = \log 2$. A multiplicative Chernoff's inequality (or Bernstein's inequality, see Lemma 13) results a confidence interval of $O(\sqrt{\frac{p_i(1)\log(1/\delta)}{n}})$ 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 $O(\sqrt{\frac{\log(1/\delta)}{n}})$ 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 can also be compared to the Lai-Robbins impossibility result [1, Theorem 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 ones with large entropy. The PMF of these arms is at 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[1] $\zeta(p) \in [0,1]$, 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 $\zeta_i|\mathcal{X}_i|$, which demonstrates the utility of this functional.

A. The Bernoulli Case

In this section, we consider the Bernoulli case, in which $\mathcal{X}_i = \{0,1\}$ for all the K arms, $i \in [K]$. For brevity we use $h_b(p) := -p\log p - (1-p)\log(1-p)$ 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 proposed UCB algorithm and its regret analysis are based on the following confidence interval function

$$\mathrm{UCB}_{\text{ber}}(q,\delta,n) := \sqrt{\frac{12q\log(\frac{6}{\delta})}{n}}\log\left(\frac{n}{q\log(\frac{6}{\delta})}\right) + \frac{18\log(\frac{6}{\delta})\log(n)}{n}, \tag{11}$$

and the corresponding confidence interval bound for the plug-in entropy estimator:

Proposition 3. *Let $\boldsymbol{Y} = \{Y_\ell\}_{\ell\in[n]}$ be IID from a Bernoulli with parameter $p = \mathbb{P}[Y_i = 1]$, and let $\hat{p}(n) = \frac{1}{n}\sum_{\ell=1}^{n}\mathbb{1}\{Y_\ell = 1\}$ be the empirical probability of '1'. Let $\delta \in [0, \frac{1}{2}]$ be given. If $n \geq 200\cdot\log(\frac{4}{\delta})$ then*

$$|h_b(\hat{p}(n)) - h_b(p)| \leq \mathrm{UCB}_{\text{ber}}(\hat{p}(n),\delta,n), \tag{12}$$

with probability larger than $1-\delta$.

Remark 4. The confidence interval of Proposition 3 follows from the relation

$$|H(p) - H(q)| \leq \mathrm{d_{TV}}(p,q)\log\left(\frac{|\mathcal{Y}|}{\mathrm{d_{TV}}(p,q)}\right), \tag{13}$$

where $\mathrm{d_{TV}}(p,q)$ is the total variation distance between PMFs p and q defined on a common alphabet $\mathcal{Y}$, and that holds as long as $\mathrm{d_{TV}}(p,q) \leq \frac{1}{2}$ [42, Lemma 2.7]. We remark that (13) is not the sharpest known bound, and, e.g., it also holds that [23, Theorem 6]

$$|H(p) - H(q)| \leq \mathrm{d_{TV}}(p,q)\log\left(|\mathcal{Y}|-1\right) + h_b\left(\mathrm{d_{TV}}(p,q)\right), \tag{14}$$

[1]We define this functional explicitly in (25).

(see also an additional refinement in [24]). In our proofs in the rest of the section such a bound is utilized in the regime $\mathrm{d_{TV}}(p,q) = o(1)$, for which both (13) and (14) are of the same order of $\Theta\left(\mathrm{d_{TV}}(p,q)\log\frac{|\mathcal{Y}|}{\mathrm{d_{TV}}(p,q)}\right)$. Thus, we exclusively use the simpler bound (13).

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

$$\Gamma_{\text{ber}}(\alpha,\beta,q,\Delta,t) := \max\left\{6\cdot\Lambda_1\left(\frac{36\alpha\log(t)}{(1-\beta)\Delta}\right), \frac{5120q\alpha\log(t)}{\beta^2\Delta^2}\cdot\log^2\left(\frac{48}{\beta^2\Delta^2}\right), \frac{88\sqrt{\alpha\log(t)}}{\beta\Delta}\cdot\log\left(\frac{48}{\beta^2\Delta^2}\right)\right\}, \quad (15)$$

and use the notation $\hat{p}(\boldsymbol{Y},n)$ for the empirical probability of $'1'$ of $\boldsymbol{Y}=\{Y_\ell\}_{\ell\in[n]}$.

Theorem 5. *Assume that $\mathcal{X}_i=\{0,1\}$ for all $i\in[K]$. Further assume that Algorithm 1 is run with the plug-in entropy estimator $\hat{H}(\boldsymbol{Y},n)\equiv H(\hat{p}(\boldsymbol{Y},n))$ and upper confidence interval*

$$\mathrm{UCB}(\boldsymbol{Y},\delta,n)\equiv\mathrm{UCB}_{\text{ber}}(\hat{p}(\boldsymbol{Y},n),\delta,n), \quad (16)$$

(as defined in (11)) with $\delta\equiv\delta_\alpha(t)=6t^{-\alpha}$ with $\alpha>2$. Then,

$$R(t)\le\sum_{i\in[K]:\Delta_i>0}\inf_{\beta\in(0,1)}\Gamma_{\text{ber}}(\alpha,\beta,p_i(1),\Delta_i,t)\cdot\Delta_i+\frac{16(\alpha-1)}{\alpha-2}\cdot\Delta_i. \quad (17)$$

Let us inspect the regret bound of (17) of Theorem 5 in the regime of small gaps. By inserting the definition of the $\Gamma_{\text{ber}}(\cdot)$, the dominating term as $\Delta_i\downarrow 0$ is on the order of $\tilde{O}(\frac{p_i\log(t)}{\Delta_i})$ and all other terms are $O(\log^c(\frac{1}{\Delta_i}))$. Thus, e.g., in case $\Delta_i=\Theta(p_i)$ (as $t\to\infty$), then the regret is only $O(\log(t)\cdot\log^c(\frac{1}{\Delta_i}))$. This is a similar behavior to the variance-UCB algorithm [40], [41] with standard bounded rewards. Nonetheless, in general, the bound of Theorem 5 is not optimal. Indeed, recall that in the standard Bernoulli MAB problem, say with two arms ($K=2$), the regret bound depends on the difference $p_1(1)-p_2(1)$ between the $'1'$-probability of each arms, 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 [1] on the pseudo-regret. We follow the clear statement made in [39, Theorems 2.14 and 2.16]:

Theorem 6 (Lai-Robbins lower bound)**.** *Consider the IMAB problem with K arms. A problem instance $\mathcal{I}$ is the collection $\{p_i\}_{i\in[K]}$ with $p_i\equiv p_i(1)\in[0,1/2)$. Suppose that a IMAB algorithm is such that*

$R(t) = O(C_{\mathcal{I},a} t^a)$ for each problem instance I and $a > 0$. Fix an arbitrary problem instance $\mathcal{I}$. Then,

$$\liminf_{t\to\infty} \frac{R(t)}{\log(t)} \geq \sum_{i\in[K]:\Delta_i>0} \frac{\Delta_i}{\mathrm{D}_{\mathrm{KL}}(p_i||p_{i^*})}, \tag{18}$$

where $\mathrm{D}_{\mathrm{KL}}(p||q) := p\log(p/q) + (1-p)\log((1-p)/(1-q))$ is the binary Kullback-Leibler divergence, and $\Delta_i = \max_{j\in[K]} h_b(p_j) - h_b(p_i)$.[2]

The proof of Theorem 6 for the IMAB problem follows in almost exactly the same manner as in the proof of the standard Lai-Robbins lower bound for Bernoulli bandits, and can be found in [1, Theorem 1] [18, Theorem 2.2] [39, Ch. 2].

Strictly speaking, Theorem 6 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 5, while considering the effect of varying the gap. To simplify the next discussion, we will assume $K = 2$ with $p_2 < p_1 < \frac{1}{2}$. Let $\Delta \equiv \Delta_2 = h_b(p_1) - h_b(p_2)$. In the standard bandit case, one approximates $\mathrm{D}_{\mathrm{KL}}(p_2||p_1) = \Theta(\Delta^2)$ (e.g., using Pinsker's inequality and its reverse version), and then the lower bound is $\Omega(\frac{1}{\Delta}\log(t))$. This lower bound is roughly achieved by the basic UCB algorithm [18, Theorem 2.1] (and the variance-UCB algorithm leads to data-dependent improvement in the constant). Before comparing this result to the upper bound of Theorem 5, we note that the latter has extra multiplicative logarithmic factor, which can be as large as $\Theta(\log(\frac{\log(t)}{\Delta}))$. 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 $p_2 = p - \Lambda$ with p fixed and $\Lambda \downarrow 0$, then $\Delta = h_b(p_1) - h_b(p_2) = \Theta(\Lambda)$ and $\mathrm{D}_{\mathrm{KL}}(p_2||p_1) = \Theta(\Lambda^2)$, and the ratio in the lower bound is $\Theta(\frac{\log t}{\Lambda})$ as in standard bandits. This roughly matches the upper bound of Theorem 5 on the pseudo-regret achieved by the algorithm, and no significant improvements are anticipated.

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 be a difficult instance of the problem. More explicitly, let us assume that both $p_1 = \frac{1}{2} - \Lambda$ and $p_2 = \frac{1}{2} - 2\Lambda$. Then, both $\Delta = h_b(p_1) - h_b(p_2) = \Theta(\Lambda^2)$ and $\mathrm{D}_{\mathrm{KL}}(p_2||p_1) = \Theta(\Lambda^2)$, and the ratio in the lower bound is asymptotically $\Theta(\log t)$ even if $\Lambda \downarrow 0$ and so also $\Delta \downarrow 0$. In this regime, the regret of Algorithm 1 is upper bounded as $\tilde{O}(\frac{\log t}{\Delta})$ whereas the lower bound is only $\Theta(\log t)$, and so the bounds do not match. However, as we next show, the fact

[2] If $q = 0$ or $q = 1$ and $p \neq q$ then $\mathrm{D}_{\mathrm{KL}}(p||q) := \infty$.

that the binary entropy function at $1/2$ is quadratic leads to an ameliorated confidence interval bound. Specifically, consider the following confidence interval

$$\mathrm{UCB}_{\text{ber}}^{(1/2)}(q,\delta,n) := 7\left|\frac{1}{2}-q\right|\cdot\sqrt{\frac{\log(\frac{4}{\delta})}{n}}+\frac{9\log(\frac{4}{\delta})}{n}. \tag{19}$$

We now have the following confidence interval bound.

Proposition 7. *Let $\boldsymbol{Y}=\{Y_\ell\}_{\ell\in[n]}$ be IID from a Bernoulli with parameter $p=\mathbb{P}[Y_i=1]$, and let $\hat{p}(n)=\frac{1}{n}\sum_{\ell=1}^{n}\mathbb{1}\{Y_\ell=1\}$ be the empirical probability of $'1'$. Assume that $p\in[\frac{2}{5},\frac{1}{2}]$, and that $n\geq 60\log(\frac{4}{\delta})$. let $\delta\in[0,\frac{1}{2}]$ be given. Then*

$$|h_b(\hat{p}(n))-h_b(p)|\leq \mathrm{UCB}_{\text{ber}}^{(1/2)}(\hat{p}(n),\delta,n) \tag{20}$$

with probability larger than $1-\delta$, and

$$|h_b(\hat{p}(n))-h_b(p)|\leq \mathrm{UCB}_{\text{ber}}^{(1/2)}(p,\delta,n) \tag{21}$$

with probability larger than $1-\delta$.

Note that for simplicity of exposition, and since the regime of interest is of arm probabilities close to $1/2$, we have restricted p_i to $[\frac{2}{5},\frac{1}{2}]$. With this result, the following regret bound can be easily derived using the same methods used in the proof of Theorem 5, and so its proof is omitted. For simplicity, we only state it for $K=2$ arms.

Theorem 8. *Assume that $\mathcal{X}_i=\{0,1\}$ and that $p_i\in[\frac{2}{5},\frac{1}{2}]$ 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 $\hat{H}(\boldsymbol{Y},n)\equiv H(\hat{p}(\boldsymbol{Y},n))$ and upper confidence interval*

$$\mathrm{UCB}(\boldsymbol{Y},\delta,n)\equiv \mathrm{UCB}_{\text{ber}}^{(1/2)}(\hat{p}(\boldsymbol{Y},n),\delta,n) \tag{22}$$

(as defined in (19)) with $\delta\equiv\delta_\alpha(t)=4t^{-\alpha}$ with $\alpha>2$. Then,

$$R(t)\leq \frac{784\left(\frac{1}{2}-p_2\right)^2\alpha\log(t)}{\Delta}+60\alpha\log(t)+\frac{8(\alpha-1)}{\alpha-2}\cdot\Delta. \tag{23}$$

In the regime above, $(\frac{1}{2}-p_2)^2=\Lambda^2=\Theta(\Delta)$ 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 5 by taking the minimal confidence bound out of those used in Theorem 5 and that of Theorem 8. This will result the minimum of the regret upper bounds of both

theorems.

Finally, we consider the other extreme regime for the Bernoulli probabilities, to wit, the small value regime, in which the derivative of the binary entropy function is unbounded. 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 $\mathrm{D}_{\mathrm{KL}}(p_2||p_1) = \Theta(\gamma)$ (adopting the convention that $0 \log(0/q) = 0$ in the definition of the KL divergence, which satisfies a continuity). So, the lower bound is $\Omega(\log \frac{1}{\gamma} \cdot \log t)$. According to Theorem 5, the pseudo-regret bound achieved by the algorithm is $\tilde{O}\left(\frac{\log t}{\Delta}\right) = \tilde{O}(\frac{\log t}{\gamma \log \frac{1}{\gamma}})$ and this guarantee is orderwise larger (by a factor $\Omega(1/\gamma)$) than the lower bound. A similar observation is obtained if one compares a gap-independent lower bound with the gap-independent regret bound that Theorem 5 implies, given by

$$R(t) = O\left(\min\left\{\frac{\log t}{\gamma \log \frac{1}{\gamma}}, \Delta t\right\}\right) = O\left(\min\left\{\frac{\log t}{\gamma \log \frac{1}{\gamma}}, \gamma \log \frac{1}{\gamma} t\right\}\right) = O\left(\sqrt{t \log t}\right) \tag{24}$$

where the maximum regret is achieved for the gap $\gamma = \Theta(\frac{1}{\sqrt{t \log t}})$ (as can be shown by equating both terms and evaluating the order of the solution). This result does not match the gap-independent lower bound of $R(t) = \Omega(\log(t))$, which can be established by modifying, e.g., the argument in [39, 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 ϵ-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 $\Theta(\epsilon \cdot t)$ 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 $R(t) = \Omega\log(t)$. 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 $\mathrm{D}_{\mathrm{KL}}(p_1(1)||p_2(1)) = \Theta(\gamma)$ rather than quadratic, while the gap is only logarithmically above linear $\Theta(\gamma \log(1/\gamma))$). We conjecture that there exists an algorithm which achieves this lower bound, yet this is left for future research.

B. The General Alphabet Case

In this section we extend the data-dependent UCB bound of the previous section to larger alphabets. To this end, let p and q be two probability mass functions over an alphabet $\mathcal{Y}$. We consider the distribution-dependent functional

$$\zeta(p) := 1 - \sum_{y \in \mathcal{Y}} p^2(y), \tag{25}$$

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 $\zeta(p) \approx H^{(2)}(p) \ll 1$ too. As we shall see, $\zeta(p_i)|\mathcal{X}_i|$ is a measure of the effective alphabet size of the ith arm. In addition, $\zeta(p_i)$ can also be accurately estimated from the data, and thus can be used by player in determining its confidence interval.

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

$$\begin{aligned}&\mathrm{UCB}_{\mathrm{tv}}(\zeta, \mathcal{Y}, \delta, n) := \\ &\quad 3\sqrt{\frac{\zeta|\mathcal{Y}|}{n}} \log\left(\frac{n|\mathcal{Y}|}{36\zeta}\right) + \frac{3}{2}\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{n}} \log\left(\frac{n|\mathcal{Y}|^2}{9}\right) + \frac{2|\mathcal{Y}|^{1/2}\log^{1/4}\left(\frac{2}{\delta}\right)\log\left(n|\mathcal{Y}|^{2/3}\right)}{n^{3/4}},\end{aligned} \tag{26}$$

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

Proposition 9. *Let $\boldsymbol{Y} = \{Y_\ell\}_{\ell \in [n]}$ 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 $\hat{p}(n, y) = \frac{1}{n}\sum_{\ell=1}^{n} \mathbb{1}\{Y_\ell = y\}$ be the empirical PMF of $\boldsymbol{Y}$. Let $\delta \in [0, 0.2]$ be given. Then, if $n \geq 112 \cdot \log\left(\frac{2}{\delta}\right)$ then it holds that*

$$|H(\hat{p}(n)) - H(p)| \leq \mathrm{UCB}_{\mathrm{tv}}(\hat{\zeta}(n), \mathcal{Y}, \delta, n), \tag{27}$$

with probability larger than $1 - \delta$.

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

$$\begin{aligned}&\Gamma_{\mathrm{tv}}(\alpha, \zeta(p_i), \Delta_i, t) = \max\left\{288\frac{\zeta(p_i)}{|\mathcal{Y}|}\Lambda_1^2\left(\frac{2|\mathcal{Y}|}{3\Delta_i}\right),\ 36230\frac{\alpha^{1/3}\log^{1/3}(t)}{|\mathcal{Y}|^{2/3}}\Lambda_1^{4/3}\left(\frac{2|\mathcal{Y}|}{3\Delta_i}\right),\right. \\ &\left.\frac{135}{|\mathcal{Y}|^2}\Lambda_2\left(\frac{9|\mathcal{Y}|^2\alpha\log(t)}{\Delta_i^2}\right),\ \frac{3}{|\mathcal{Y}|^{2/3}}\Lambda_{4/3}\left(\frac{27|\mathcal{Y}|^{4/3}\alpha^{1/3}\log^{1/3}(t)}{\Delta_i^{4/3}}\right), 30\cdot\alpha\log\left(2^{1/\alpha}t\right), 119\zeta(p_i)|\mathcal{Y}|\right\}.\end{aligned} \tag{28}$$

Theorem 10. *Assume that Algorithm 1 is run with the plug-in entropy estimator*

$$\hat{H}(\boldsymbol{Y}, n) \equiv H(\hat{p}(\boldsymbol{Y}, n)), \tag{29}$$

and upper confidence interval

$$\mathrm{UCB}(\boldsymbol{Y}, \delta, n) \equiv \mathrm{UCB}_{\mathrm{tv}}(\hat{\zeta}(\boldsymbol{Y}, n), \mathcal{Y}, \delta, n), \tag{30}$$

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

$$R(t) \leq \sum_{i \in [K]: \Delta_i > 0} \Gamma_{\mathrm{tv}}(\alpha, \zeta(p_i), \Delta_i, t) \cdot \Delta_i + \frac{4(\alpha-1)}{\alpha-2} \cdot \Delta_i. \tag{31}$$

To show the improvement of using $\mathrm{UCB}_{\mathrm{tv}}(\zeta, \mathcal{Y}, \delta, n)$ of (26) over $\mathrm{UCB}_{\mathrm{bias}}(\delta, n)$ of (6) we observe that for a non-asymptotic time t that is upper bounded by a polynomial function of $|\mathcal{X}_i|$ the regret bound of Theorem 10 scales as

$$\tilde{O}\left(\frac{\zeta(p_i)|\mathcal{X}_i|+1}{\Delta_i} + \frac{|\mathcal{X}_i|^{2/3}}{\Delta_i^{1/3}} + \zeta(p_i)|\mathcal{X}_i|\Delta_i + \Delta_i\right), \tag{32}$$

where the $\tilde{O}(\cdot)$ 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 $|\mathcal{X}_i|^{2/3} \vee \zeta(p_i)|\mathcal{X}_i|$ which can be much smaller then the $|\mathcal{X}_i|$ dependence obtained in Theorem 2 for the biased-based UCB. 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 $\tilde{O}(\sqrt{(\zeta(p_i)|\mathcal{X}_i|+1)t} \vee \sqrt{|\mathcal{X}_i|} \cdot t^{1/4})$ which is mild in the alphabet size whenever $\zeta(p_i)|\mathcal{X}_i|$ is small.

VI. Numerical Results

This section presents numerical results that illustrate the average total regret achieved by Algorithm 1 for the entropy estimators upper and upper confidence bounds we utilize for Algorithm 1. We examined the following eight 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 2×10^6 rounds. Additionally, for each setup, we averaged the total regret across 100 Monte Carlo realizations.

Figure 1 presents the numerical results for the binary alphabet, i.e., Setups 1-4. Additionally, Figure 2 presents the numerical results for the ternary alphabet, i.e., Setups 5-8. The line *'UCB-bias'* depicts the average total regret of bias correction estimator and confidence interval used in (6). The line *'UCB-TV'* denotes the PMF-based confidence intervals that are used with conjunction of the plug-in entropy estimator. In the case of the a binary alphabet, we take the minimum between the confidence interval

	Alphabet	PMF	Entropy [nats]
Setup 1	binary	$p_1(0) = 1 - p_1(1),\ p_1(1) = 0.025$ $p_2(0) = 1 - p_2(1),\ p_2(1) = 10^{-4}$	$H_1 = 0.1169$ $H_2 = 0.0010$
Setup 2	binary	$1 - p_1(0) = p_1(1),\ p_1(1) = 0.25$ $p_2(0) = 1 - p_2(1),\ p_2(1) = 10^{-4}$	$H_1 = 0.5623$ $H_2 = 0.0010$
Setup 3	binary	$p_1(0) = 1 - p_1(1),\ p_1(1) = 0.25$ $p_2(0) = 1 - p_2(1),\ p_2(1) = 10^{-2}$	$H_1 = 0.5623$ $H_2 = 0.0560$
Setup 4	binary	$p_1(0) = 1 - p_1(1),\ p_1(1) = 0.3$ $p_2(0) = 1 - p_2(1),\ p_2(1) = 0.15$	$H_1 = 0.6109$ $H_2 = 0.4227$
Setup 5	ternary	$p_1(0) = p_1(1) = 0.0125,\ p_1(2) = 1 - p_1(0) - p_1(1)$ $p_2(0) = p_2(1) = 5 \times 10^{-5},\ p_2(2) = 1 - p_2(0) - p_2(1)$	$H_1 = 0.1342$ $H_2 = 0.0011$
Setup 6	ternary	$p_1(0) = p_1(1) = 0.125,\ p_1(2) = 1 - p_1(0) - p_1(1)$ $p_2(0) = p_2(1) = 5 \times 10^{-5},\ p_2(2) = 1 - p_2(0) - p_2(1)$	$H_1 = 0.7356$ $H_2 = 0.0011$
Setup 7	ternary	$p_1(0) = p_1(1) = 0.125,\ p_1(2) = 1 - p_1(0) - p_1(1)$ $p_2(0) = p_2(1) = 5 \times 10^{-2},\ p_1(2) = 1 - p_2(0) - p_2(1)$	$H_1 = 0.7356$ $H_2 = 0.0629$
Setup 8	ternary	$p_1(0) = p_1(1) = 0.15,\ p_1(2) = 1 - p_1(0) - p_1(1)$ $p_2(0) = p_2(1) = 0.075,\ p_2(2) = 1 - p_2(0) - p_2(1)$	$H_1 = 0.8188$ $H_2 = 0.5267$

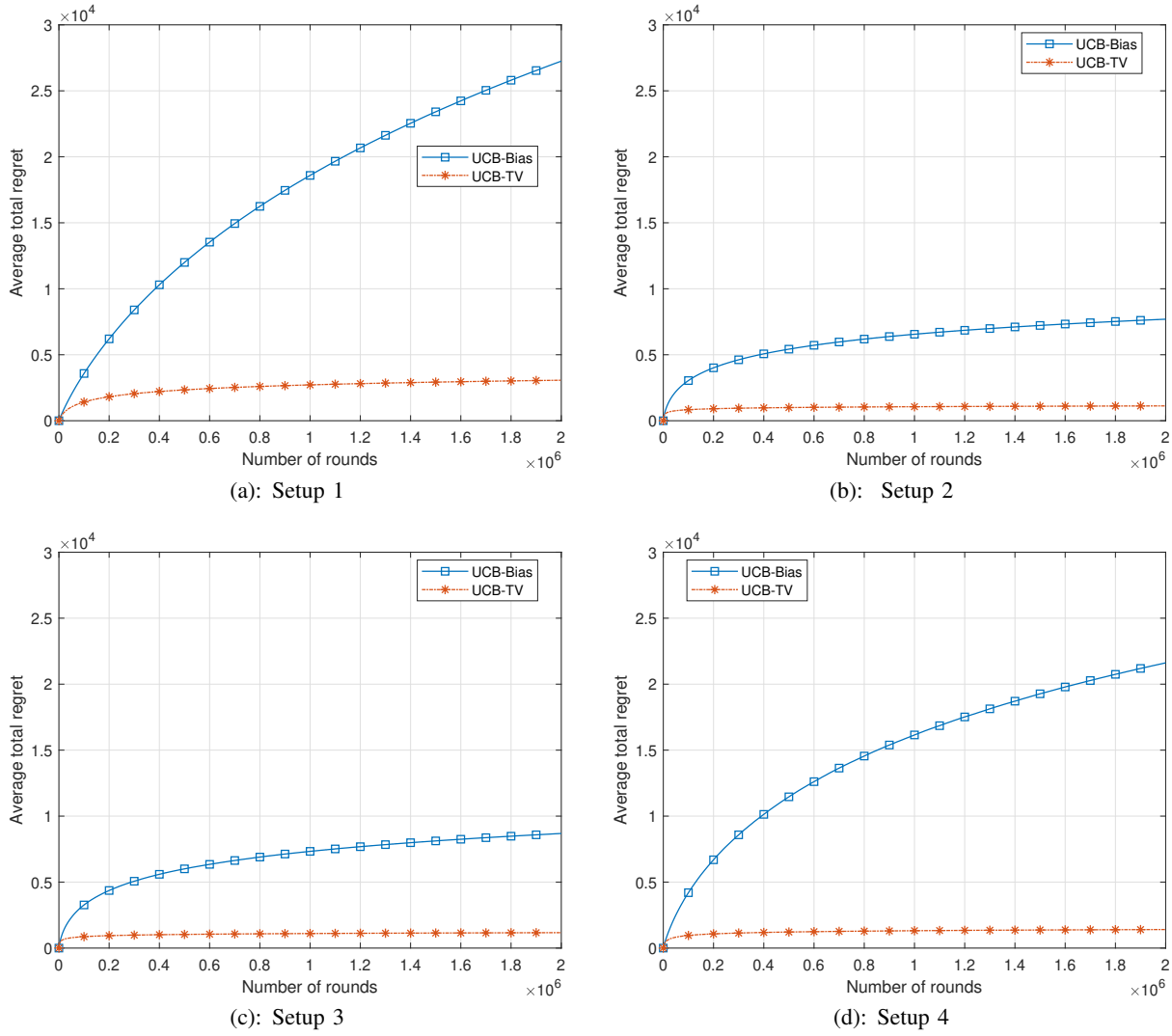
TABLE II
NUMERICAL RESULT SETUPS.

(11) and the confidence interval (19). In the case of a larger alphabet, we use the general alphabet confidence interval (26).

It is evident from Figure 1 that the Bernoulli PMF-based confidence intervals (11) and (19) provide significant reduction in the average total regret in comparison to the combination of the bias correction estimator and confidence interval used in (6). Furthermore, we can see that as expected, the bias correction approach suffers from significant increased regret values as the probability of drawing the symbol '1', i.e., $p(1)$, gets closer to the boundary points of the interval $\left[0, \frac{1}{2}\right]$. The Bernoulli PMF-based confidence intervals exhibit robustness in these regimes and have not suffered from such stark degradation in performance.

As we increase the alphabet size from two (binary) to three (ternary), Figure 2 shows that the general PMF-based confidence interval (26) does not perform as well as the binary ones, i.e., (11) and (19). This occurs since the general PMF-based confidence interval (26) targets scenarios where $\zeta(p_i) \cdot |\mathcal{X}_i|$ is sufficiently small. Indeed, this is illustrated in Figure 2(a), where the performance of the PMF-based confidence interval is comparable to that of the bias correction scheme (6).

In addition to Setups 1-8 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 5×10^{-3}, these probabilities are chosen randomly; the probability of the last



(a): Setup 1 (b): Setup 2

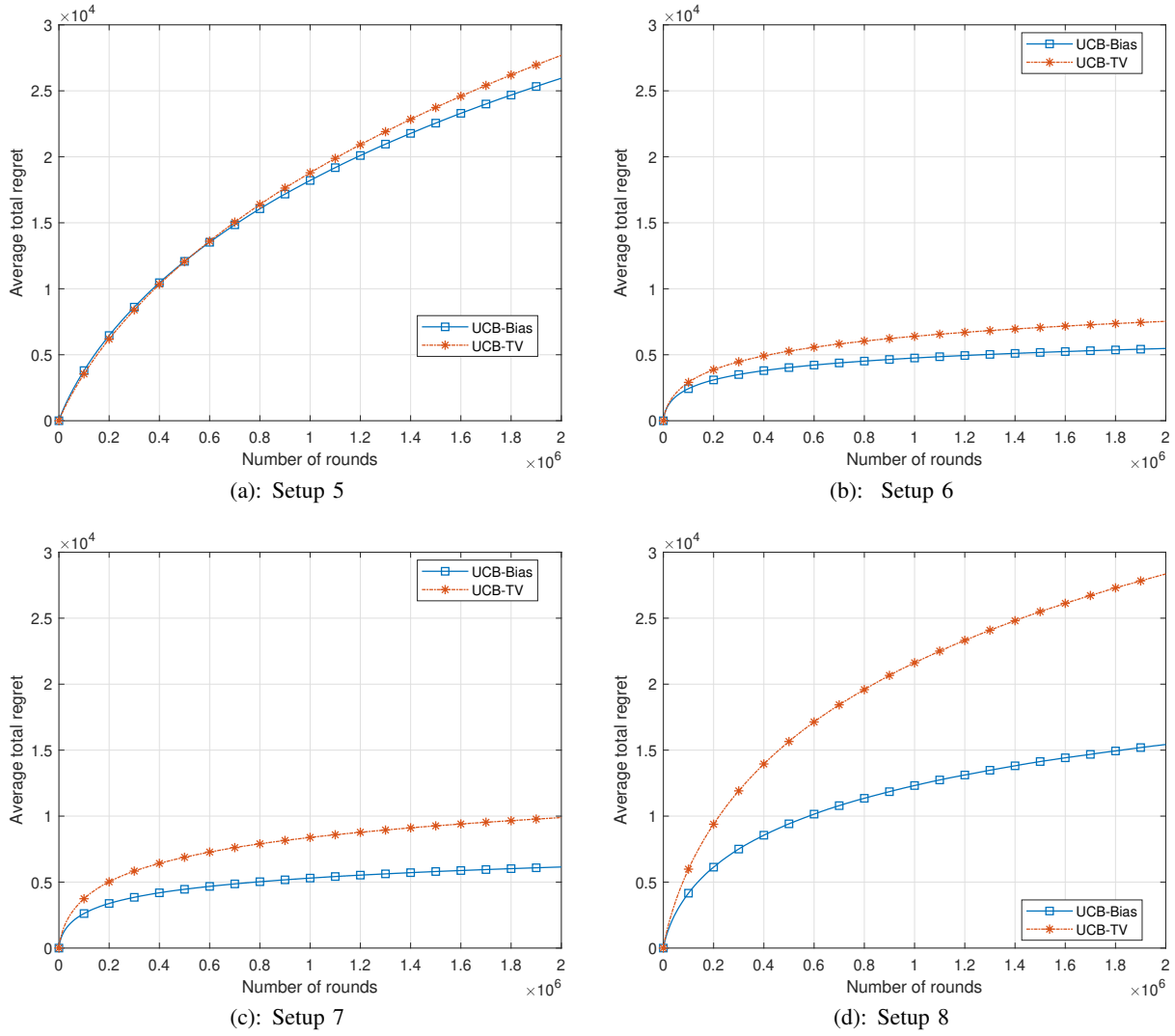
(c): Setup 3 (d): Setup 4

Fig. 1. Average total regret as a function of the number of rounds for a two-armed bandit model for arms with a binary alphabet.

symbol is $1-5\times10^{-3}$. For second arm, the total probability of the first 10^4-1 symbols is $\times10^{-4}$, these probabilities are chosen randomly; the probability of the last symbol is $1-10^{-4}$. Thus, it must be for all generated PFMs that $\zeta_1 \leq 0.01$ and $\zeta_2 \leq 2\times10^{-4}$. We refer to this setup as Setup 9. Figure 3 demonstrates the reduction in the average regret that the general PMF-based confidence interval (26) leads to in a non-asymptotic time regime with a large alphabet size and small total variance values.

VII. Summary and Future Research

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 that it is inefficient for when the entropy is very low compared to the log-alphabet size. For this regime, we have proposed a



(a): Setup 5 (b): Setup 6

(c): Setup 7 (d): Setup 8

Fig. 2. Average total regret as a function of the number of rounds for a two-armed bandit model for arms with a ternary alphabet.

UCB algorithm that is based on data-dependent UCB, which significantly improves upon the bias UCB algorithm. Nonetheless, there is still a gap between our upper bound on its regret and standard impossibility lower bounds. It is thus an open problem to close this gap. In addition, in this paper we have assumed that the alphabet of each arms is known in advance to the player. In practice, the alphabet of the arm may be very large compared to the support of the PMF, and it is of interest to develop UCB algorithms for this case. This can be achieved by incorporating support-size estimators into the confidence interval [29], [43], [44], and is left for future research.

ACKNOWLEDGEMENT

The first author thanks Itay Ron for multiple discussions at early stages of this research.

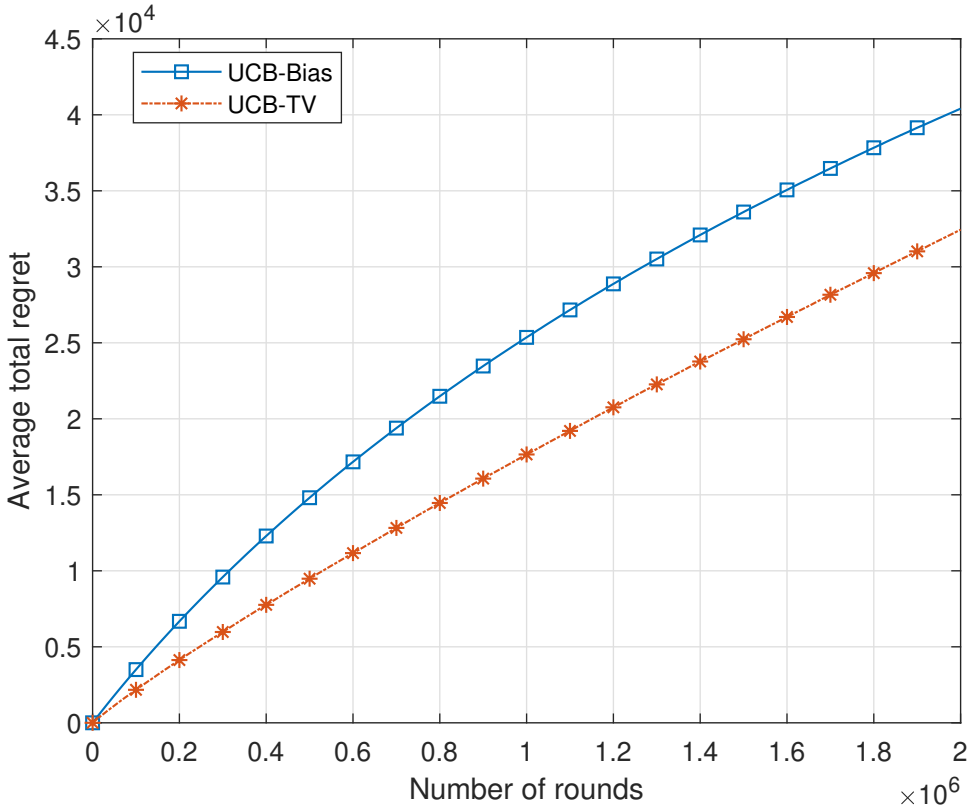


Fig. 3. Average total regret as a function of the number of rounds for a two-armed bandit model for arms with an alphabet size of 10^4 (Setup 9).

APPENDIX A
INVERTING POLYLOGARITHMIC FUNCTIONS OVER LINEAR FUNCTIONS

Lemma 11. *Let $r \in [1,2]$ be given. There exists a constant $c_r > 0$ so that if $x \geq c_r \log^r(1/y)/y = \Lambda_r(1/y)$ then $\frac{\log^r x}{x} \leq y$. 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 $x \to \frac{\log^r x}{x}$ has a unique maximum at $x = e^r$, and its maximal value is $\left(\frac{r}{e}\right)^r$ (which is less than 1 for any $r \in [1,2]$). So, $\frac{\log^r x}{x}$ is monotonic strictly decreasing for $x \geq e^r$. If $y \geq \left(\frac{r}{e}\right)^r$ then $\frac{\log x}{x} \leq y$ for all $x \in \mathbb{R}_+$ and the claim of the lemma trivially holds. Otherwise, if $y \in [0, \left(\frac{r}{e}\right)^r]$ then setting $x = c_r \log^r(1/y)/y$ results

$$\begin{aligned} \frac{\log^r x}{x} &= \frac{y \log^r\left(\frac{c_r \log^r(1/y)}{y}\right)}{c_r \log^r(1/y)} \\ &\leq y \cdot \left[\frac{\log(c_r) + r\log\log(1/y) + \log(1/y)}{c_r^{1/r} \cdot \log(1/y)}\right]^r \end{aligned} \tag{33}$$

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

For any given power r, the term inside the square brackets can be made arbitrarily small by taking $c_r \uparrow \infty$, and specifically, can be made less than 1, which results $\frac{\log^r x}{x} \leq y$ for the aforementioned choice of x, with some numerical constant c_r. The minimal constant can be find by checking (33) 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, $\frac{\log x}{x}/y = \omega(1)$. □

APPENDIX B
PROOFS FOR SECTION IV

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

$$\begin{aligned}
&\mathbb{P}\left(H(\hat{p}(n)) + B(n) - H(p) > \epsilon\right) \\
&= \mathbb{P}\left(H(\hat{p}(n)) - \mathbb{E}(H(\hat{p}(n))) > \epsilon + \underbrace{H(p) - \mathbb{E}(H(\hat{p}(n)))}_{\geq 0} - B(n) \right) \\
&\overset{(a)}{\leq} \mathbb{P}\left(H(\hat{p}(n)) - \mathbb{E}(H(\hat{p}(n))) > \epsilon - B(n)\right) \\
&\overset{(b)}{\leq} \mathbb{1}[\epsilon \leq B(n)] + \mathbb{1}[\epsilon > B(n)] \cdot 2\exp\left[-\frac{n}{2} \left(\frac{\epsilon - B(n)}{\log(n)} \right)^2 \right],
\end{aligned} \tag{34}$$

where (a) follows from the bound on the bias in (5), and (b) follows from [22, p. 168]. For the lower confidence bound it holds similarly that

$$\begin{aligned}
&\mathbb{P}\left(H(\hat{p}(n)) + B(n) - H(p) < -\epsilon\right) \\
&= \mathbb{P}\left(H(\hat{p}(n)) - \mathbb{E}(H(\hat{p}(n))) < -\epsilon + \underbrace{H(p) - \mathbb{E}(H(\hat{p}(n))) - B(n)}_{\leq 0} \right) \\
&\overset{(a)}{\leq} \mathbb{P}\left(H(\hat{p}(n)) - \mathbb{E}(H(\hat{p}(n))) < -\epsilon\right) \\
&\overset{(b)}{\leq} \exp\left[-\frac{n}{2} \left(\frac{\epsilon}{\log(n)} \right)^2 \right].
\end{aligned} \tag{35}$$

. Combining (34) and (35) shows that

$$\mathbb{P}\left(|H(\hat{p}(n)) + B(n) - H(p)| > \epsilon\right) \leq \mathbb{1}\left[\epsilon \leq B(n)\right] + \mathbb{1}\left[\epsilon > B(n)\right] \cdot 2\exp\left[-\frac{n}{2}\left(\frac{\epsilon - B(n)}{\log(n)}\right)^2\right]$$
$$= 2\exp\left[-\frac{n}{2}\frac{(\epsilon - B(n))_+^2}{\log^2(n)}\right]. \tag{36}$$

for every $n \geq 2$ and $\epsilon > 0$. Setting the RHS of (36) to δ and simplifying leads to the claimed result. □

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 12. *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 $\mathrm{UCB}_{\text{bias}}(t^{-\alpha}, n) \leq \Delta/2$.*

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

$$B(n) \leq \beta \cdot \Delta/2, \tag{37}$$

and

$$\sqrt{\frac{2\log^2(n)}{n}\log\left(\frac{2}{\delta}\right)} \leq (1-\beta)\cdot\Delta/2, \tag{38}$$

holds. The first condition (37) is equivalent to

$$n \geq \frac{|\mathcal{Y}|-1}{e^{\beta\cdot\Delta/2} - 1}, \tag{39}$$

and the second condition (38) is equivalent to

$$\frac{\log^2(n)}{n} \leq \frac{(1-\beta)^2\Delta^2}{8\log(2t^\alpha)}. \tag{40}$$

According to Lemma 11 this holds if

$$n \geq \frac{120\cdot\log(2t^\alpha)\cdot\log^2\left(\frac{8\log(2t^\alpha)}{(1-\beta)^2\Delta^2}\right)}{(1-\beta)^2\Delta^2} = 15\cdot\Lambda_2\left(\frac{8\cdot\log(2t^\alpha)}{(1-\beta)^2\Delta^2}\right), \tag{41}$$

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

With this result at hand we may prove Theorem 2.

Proof of Thm. 2. The proof follows the analysis of [18, Proof of Theorem 2.1], with required modifications

to entropy rewards structure. At round t, the player chooses a suboptimal i arm with $\Delta_i > 0$ if

$$\hat{H}(\boldsymbol{X}_{i^*}(t-1), N_{i^*}(t-1))) + \mathrm{UCB}_{\text{bias}}(\delta_\alpha(t), N_{i^*}(t-1)) \\ \leq \hat{H}(\boldsymbol{X}_i(t-1), N_i(t-1))) + \mathrm{UCB}_{\text{bias}}(\delta_\alpha(t), N_i(t-1)).$$

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:

$$\hat{H}(\boldsymbol{X}_{i^*}(t-1), N_{i^*}(t-1))) + \mathrm{UCB}_{\text{bias}}(\delta_\alpha(t), N_{i^*}(t-1)) \leq H_{i^*}. \tag{42}$$

II. The entropy of arm i is significantly overestimated:

$$\hat{H}(\boldsymbol{X}_i(t-1), N_i(t-1))) > H_i + \mathrm{UCB}_{\text{bias}}(\delta_\alpha(t), N_i(t-1)). \tag{43}$$

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

$$\mathrm{UCB}_{\text{bias}}(\delta_\alpha(t), N_i(t-1)) > \Delta_i/2. \tag{44}$$

If all three events I-III are false, then

$$\begin{aligned} &\hat{H}(\boldsymbol{X}_{i^*}(t-1), N_{i^*}(t-1))) + \mathrm{UCB}_{\text{bias}}(\delta_\alpha(t), N_{i^*}(t-1)) \\ &> H_{i^*} = H_i + \Delta_i \\ &\geq H_i + 2 \cdot \mathrm{UCB}_{\text{bias}}(\delta_\alpha(t), N_i(t-1)) \\ &\geq \hat{H}(\boldsymbol{X}_i(t-1), N_i(t-1))) + \mathrm{UCB}_{\text{bias}}(\delta_\alpha(t), N_i(t-1)), \end{aligned} \tag{45}$$

which contradicts the assumption that Algorithm 1 chooses $I_t = i$ at the tth 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 $N_i(t) \geq \Gamma_{\text{bias}}(\alpha, \beta, \mathcal{X}_i, \Delta_i, t)$ then event III does not occur. so,

$$\mathbb{E}(N_i(t)) = \mathbb{E}\left(\sum_{\tau=1}^{t} \mathbb{1}[I(\tau) = i]\right) \\ \leq \Gamma_{\text{bias}}(\alpha, \beta, \mathcal{X}_i, \Delta_i, t) + \sum_{\tau=\Gamma_{\text{bias}}(\alpha,\beta,\mathcal{X}_i,\Delta_i,t)+1}^{t} \mathbb{E}\left(\mathbb{1}[\text{I or II is true a round } \tau]\right).$$

$$\leq \Gamma_{\text{bias}}(\alpha, \beta, \mathcal{X}_i, \Delta_i, t) + \sum_{\tau=1}^{t} \left[\mathbb{P}\left(\text{I is true at round } \tau\right) + \mathbb{P}\left(\text{II is true at round } \tau\right)\right]. \quad (46)$$

For any $\tau \leq t$, the first probability in (46) is upper bounded as

$$\begin{aligned}
&\mathbb{P}\left(\text{I is true at round } \tau\right) \\
&\overset{(a)}{\leq} \sum_{n=1}^{\tau} \mathbb{P}\left(\hat{H}(\{X_i(\ell)\}_{\ell\in[n]}, n)) + \text{UCB}_{\text{bias}}(\delta_\alpha(\tau)), n) \leq H_i\right) \\
&\overset{(b)}{\leq} \tau \cdot \delta_\alpha(\tau) = \frac{1}{\tau^{\alpha-1}},
\end{aligned}$$

where (a) follows from the union bound, and (b) from the definition of the upper confidence interval $\text{UCB}(\delta, n)$. The second probability in (46) is similarly upper bounded. Inserting these bounds back to the sum in (46) it then follows that

$$\begin{aligned}
&\sum_{\tau=1}^{t} \left[\mathbb{P}\left(\text{I is true at round } \tau\right) + \mathbb{P}\left(\text{II is true at round } \tau\right)\right] \\
&\leq 2\sum_{\tau=1}^{t} \frac{1}{\tau^{\alpha-1}} \leq 2\sum_{t=1}^{\infty} \frac{1}{\tau^{\alpha-1}} \\
&\leq 2\left[1 + \int_1^{\infty} \frac{1}{\tau^{\alpha-1}} \mathrm{d}t\right] = \frac{2(\alpha-1)}{\alpha-2}. \quad (47)
\end{aligned}$$

Substituting the upper bounds in the last two displays back to (46), and using the resulting bound in $R(t) = \sum_{i\in[K]:\Delta_i>0} \mathbb{E}(N_i(t))\Delta_i$ then concludes the proof. □

APPENDIX C
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 13. *In the setting of Proposition 3, each of the following events holds with probability larger than* $1 - \delta$*:*

$$|p - \hat{p}(n)| \leq \sqrt{\frac{3p\log(\frac{2}{\delta})}{n}}, \quad (48)$$

$$p \leq 2\hat{p}(n) + \frac{12\log(\frac{1}{\delta})}{n}, \quad (49)$$

and

$$\hat{p}(n) \leq 2p + \frac{3\log(\frac{1}{\delta})}{n}. \quad (50)$$

Proof. We will use the relative (multiplicative) Chernoff bound multiple times. This bound states that [45, Thm. 4.4]

$$\mathbb{P}\left[|\hat{p}(n)-p|\geq \xi p\right]=\mathbb{P}\left[\hat{p}(n)-p\geq \xi p\right]+\mathbb{P}\left[\hat{p}(n)-p\leq -\xi p\right]\leq 2e^{-\frac{\xi^2 pn}{3}}, \tag{51}$$

for any $\xi\in[0,1]$ (and it holds for the pair of one-sided deviations each without the 2 pre-factor). Setting $\xi=\sqrt{\frac{3\log(\frac{2}{\delta})}{pn}}$ in (51) immediately leads to (48). Next, if $p>\frac{12\log(\frac{1}{\delta})}{n}$ then

$$\mathbb{P}\left[p\geq 2\hat{p}(n)\right]=\mathbb{P}\left[\hat{p}(n)-p\leq -\frac{1}{2}p\right]\overset{(a)}{\leq} e^{-\frac{pn}{12}}\overset{(b)}{\leq}\delta, \tag{52}$$

where (a) is by setting $\xi=1/2$ in the one-sided version of (51), and (b) utilizes the assumption on p. Thus, with probability larger than $1-\delta$ it holds that

$$p\leq 2\hat{p}(n)\vee\frac{12\log(\frac{1}{\delta})}{n}, \tag{53}$$

which can be loosened to (49). Finally, If $p>\frac{3\log(\frac{1}{\delta})}{n}$ then

$$\mathbb{P}\left[\hat{p}(n)>2p\right]=\mathbb{P}\left[\hat{p}(n)-p\geq \xi p\right]\overset{(a)}{\leq} e^{-\frac{pn}{3}}\overset{(b)}{\leq}\delta, \tag{54}$$

where (a) is by setting $\xi=1$ in the one-sided version of (51), and (b) utilizes the assumption on p. Thus, with probability larger than $1-\delta$ it holds that

$$\hat{p}(n)\leq 2p\vee\frac{3\log(\frac{1}{\delta})}{n}, \tag{55}$$

which can be loosened to (50). □

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 $\mathrm{d_{TV}}(p,\hat{p}(n))\leq\frac{1}{2}$ then [42, Lemma 2.7] implies that

$$\begin{aligned}|h_b(\hat{p}(n))-h_b(p)|&\leq\sqrt{\frac{12p\log(\frac{2}{\delta})}{n}}\log\left(\sqrt{\frac{4n}{12p\log(\frac{2}{\delta})}}\right)\\&=-2\cdot\Lambda_1\left(\frac{\mathrm{d_{TV}}(p,\hat{p}(n))}{2}\right),\end{aligned}$$

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

$$\mathrm{d_{TV}}(p,q)=2|p(1)-q(1)|, \tag{56}$$

and so by (48) and (49) from Lemma 13 it holds that

$$\mathrm{d_{TV}}(p,q) \leq \sqrt{\frac{12p\log(\frac{2}{\delta})}{n}}, \tag{57}$$

and

$$p \leq 2\hat{p}(n) + \frac{12\log(\frac{1}{\delta})}{n}, \tag{58}$$

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 (57), when substituted with the upper bound on p in (58), is less than e^{-1}, to wit

$$\sqrt{\frac{12\left[2\hat{p}(n)+\frac{12\log(\frac{1}{\delta})}{n}\right]\log(\frac{2}{\delta})}{n}} \leq e^{-1}. \tag{59}$$

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

$$|h_b(\hat{p}(n)) - h_b(p)| \leq \sqrt{\frac{12\hat{p}(n)\log(\frac{2}{\delta})}{n}}\log\left(\frac{n}{\hat{p}(n)\log(\frac{2}{\delta})}\right) \tag{60}$$

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

$$|h_b(\hat{p}(n)) - h_b(p)| \leq \frac{18\log(\frac{2}{\delta})\log(n)}{n} \tag{61}$$

(after, again, slightly deteriorating the constants). To account for both cases, we sum the two deviation terms. Finally, to obtain (11), we replace δ with 2δ. □

Next, we turn to the proof of Theorem 5, which is based on a lemma analogous to Lemma 12. To this end, we further denote a simplified version of $\Gamma_{\text{ber}}(\cdot)$ from (15), defined as

$$\tilde{\Gamma}_{ber}(\alpha,\beta,q,\Delta,t) := \max\left\{2\cdot\Lambda_1\left(\frac{36\alpha\log(t)}{(1-\beta)\Delta}\right), \frac{960q\alpha\log(t)}{\beta^2\cdot\Delta^2}\cdot\log^2\left(\frac{48}{\beta^2\cdot\Delta^2}\right)\right\}. \tag{62}$$

Lemma 14. *With* $\delta \equiv \delta_\alpha(t) = 4t^{-\alpha}$, $\beta \in (0,1)$ *and* $\alpha > 2$, *if* $n \geq \tilde{\Gamma}_{\text{ber}}(\alpha,\beta,q,\Delta,t)$ *then* $\mathrm{UCB_{ber}}(q,\delta_\alpha(t),n) \leq \Delta/2$ *where* $\mathrm{UCB_{ber}}(\cdot)$ *is as defined in (11).*

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, $\mathrm{UCB}_{\text{ber}}(q, 4t^{-\alpha}, n) \leq \Delta/2$ if both

$$\sqrt{\frac{12q\alpha\log(t)}{n}}\log\left(\frac{n}{q\alpha\log(t)}\right) \leq \beta\cdot\Delta/2, \tag{63}$$

and

$$\frac{18\alpha\log(t)\log(n)}{n} \leq (1-\beta)\cdot\Delta/2, \tag{64}$$

hold. The first condition is satisfied if

$$\frac{\log^2\left(\frac{n}{q\log(\frac{6}{\delta})}\right)}{\frac{n}{q\log(\frac{6}{\delta})}} \leq \frac{\beta^2\cdot\Delta^2}{48}, \tag{65}$$

for which Lemma 11 implies that this condition is satisfied if

$$n \geq \frac{960q\alpha\log(t)}{\beta^2\cdot\Delta^2}\cdot\log^2\left(\frac{48}{\beta^2\cdot\Delta^2}\right). \tag{66}$$

The second term is satisfied if

$$\frac{\log(n)}{n} \leq \frac{(1-\beta)\Delta}{36\alpha\log(t)}. \tag{67}$$

for which Lemma 11 implies that this is satisfied if

$$n \geq \frac{72\alpha\log(t)}{(1-\beta)\Delta}\log\left(\frac{36\alpha\log(t)}{(1-\beta)\Delta}\right) = 2\cdot\Lambda_1\left(\frac{36\alpha\log(t)}{(1-\beta)\Delta}\right), \tag{68}$$

(recall the notation (1)). The claim of the lemma then follows from the definition of $\tilde{\Gamma}_{\text{ber}}(\cdot)$ in (62). □

We may now prove Theorem 5.

Proof of Theorem 5. 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 $\mathrm{UCB}(\boldsymbol{Y}, \delta, n)$ and $\mathrm{UCB}_{\text{ber}}(\hat{p}(\boldsymbol{Y}, n), \delta, n)$ 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

$$\hat{H}(\boldsymbol{X}_{i^*}(t-1), N_{i^*}(t-1))) + \mathrm{UCB}(\boldsymbol{X}_{i^*}(t-1), \delta_\alpha(t), N_{i^*}(t-1)) \\ \leq \hat{H}(\boldsymbol{X}_i(t-1), N_i(t-1))) + \mathrm{UCB}(\boldsymbol{X}_i(t-1), \delta_\alpha(t), N_i(t-1)).$$

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

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

$$\hat{H}(\boldsymbol{X}_{i^*}(t-1), N_{i^*}(t-1))) + \mathrm{UCB}(\boldsymbol{X}_{i^*}(t-1), \delta_\alpha(t), N_{i^*}(t-1)) \leq H_{i^*}, \tag{69}$$

or

$$\hat{p}\left(\boldsymbol{X}_{i^*}(t-1), N_{i^*}(t-1)\right) - \frac{1}{2}p_{i^*} \leq -\frac{6\log(1/\delta_\alpha(t))}{N_{i^*}(t-1)}. \tag{70}$$

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

$$\hat{H}(\boldsymbol{X}_i(t-1), N_i(t-1))) > H_i + \mathrm{UCB}(\boldsymbol{X}_i(t-1), \delta_\alpha(t), N_i(t-1)), \tag{71}$$

or

$$\hat{p}\left(\boldsymbol{X}_i(t-1), N_i(t-1)\right) - 2p_i \geq \frac{3\log(1/\delta_\alpha(t))}{N_i(t-1)}. \tag{72}$$

III'. The upper confidence interval, which is based on an overestimation of $\hat{p}(\boldsymbol{X}_i(t-1), N_i(t-1))$ is significantly larger than the gap

$$\mathrm{UCB}_{\text{ber}}\left(2p_i + \frac{3\log(1/\delta_\alpha(t))}{N_i(t-1)}, \delta_\alpha(t), N_i(t-1)\right) > \frac{\Delta_i}{2}, \tag{73}$$

or

$$N_i(t-1) \leq 200\alpha\log(t). \tag{74}$$

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

$$\begin{aligned}
&\hat{H}(\boldsymbol{X}_{i^*}(t-1), N_{i^*}(t-1))) + \mathrm{UCB}(\boldsymbol{X}_{i^*}(t-1), \delta_\alpha(t), N_{i^*}(t-1)) \\
&\geq H_{i^*} = H_i + \Delta_i \\
&\geq H_i + 2\mathrm{UCB}_{\text{ber}}\left(2p_i + \frac{3\log(1/\delta_\alpha(t))}{N_i(t-1)}, \delta_\alpha(t), N_i(t-1)\right) \\
&\overset{(*)}{\geq} H_i + 2\mathrm{UCB}\left(\boldsymbol{X}_i(t-1), \delta_\alpha(t), N_i(t-1)\right) \\
&\geq \hat{H}(\boldsymbol{X}_i(t-1), N_i(t-1))) + \mathrm{UCB}\left(\boldsymbol{X}_i(t-1), \delta_\alpha(t), N_i(t-1)\right),
\end{aligned}$$

where in $(*)$ we have used the current assumption that $N_i(t-1) \geq 200\alpha\log(t)$, which assures that $\mathrm{UCB}_{\text{ber}}(q, \delta_\alpha(t), N_i(t-1))$ is a monotonically non-decreasing function of q. Thus, in this case Algorithm 1 will not choose $I_t = i$ at the tth round; a contradiction.

By Lemma 14 if

$$N_i(t-1) \geq \tilde{\Gamma}_{\text{ber}}\left(\alpha, \beta, 2p_i + \frac{3\log(1/\delta_\alpha(t))}{N_i(t-1)}, \Delta_i, t\right), \tag{75}$$

then first part of the event III' does not occur. By the definition of $\tilde{\Gamma}_{\text{ber}}(\cdot)$ in (62), and by setting $\delta_\alpha(t) = 4t^{-\alpha}$, the RHS in the last equation is upper bounded as

$$\max\left\{2\cdot\Lambda_1\left(\frac{36\alpha\log(t)}{(1-\beta)\Delta_i}\right),\right.$$
$$\left.\frac{2560p_i\alpha\log(t)}{\beta^2\cdot\Delta_i^2}\cdot\log^2\left(\frac{48}{\beta^2\cdot\Delta_i^2}\right)+\frac{3840\alpha\log(t)}{\beta^2\cdot\Delta_i^2 N_i(t-1)}\cdot\log^2\left(\frac{48}{\beta^2\cdot\Delta_i^2}\right)\right\}.$$

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

$$N_i(t-1)\geq$$
$$\max\left\{2\cdot\Lambda_1\left(\frac{36\alpha\log(t)}{(1-\beta)\Delta_i}\right),\frac{5120p_i\alpha\log(t)}{\beta^2\cdot\Delta_i^2}\cdot\log^2\left(\frac{48}{\beta^2\cdot\Delta_i^2}\right),\frac{88\sqrt{\alpha\log(t)}}{\beta\cdot\Delta_i}\cdot\log\left(\frac{48}{\beta^2\cdot\Delta_i^2}\right)\right\}$$
$$=\Gamma_{\text{ber}}(\alpha,\beta,p_i,\Delta_i,t). \quad (76)$$

The second part of event III' does not occur if $N_i(t-1)\geq 200\alpha\log(t)$, which is already covered by the condition in (76) if we increase the pre-constant of the second term to 6, which is the definition of $\Gamma_{\text{ber}}(\cdot)$ used in (15).

The analysis then follows as in the proof of Theorem 2, by using Lemma 13 and Proposition 3 to bound the probabilities of the events in I' and II. Note that the condition $N_i(t-1)\geq 200\cdot\log(\frac{4}{\delta})=200\alpha\log(t)$ required for the confidence bound to hold with high probability is already satisfied in (76). □

Proof of Proposition 7. By Taylor approximation at the point p, for any $q\in[0,\frac{1}{2}]$

$$h_b(q)=h_b(p)+h_b'(p)(q-p)+\frac{h_b''(\xi)}{2}(q-p)^2, \quad (77)$$

where $\xi\in[p,q]\cup[q,p]$. From Lemma 13, it holds with probability larger than $1-2\delta$ that both $p\leq 2\hat{p}(n)+\frac{12\log(\frac{1}{\delta})}{n}$ and $|p-\hat{p}(n)|\leq\sqrt{\frac{3p\log(\frac{2}{\delta})}{n}}$. Under this event, since $n\geq 60\log(\frac{2}{\delta})$ was assumed, it holds that $\hat{p}(n)\geq\frac{1}{10}$. For $q\in[\frac{2}{5},\frac{1}{2}]$ it can be easily verified that

$$|h_b'(q)|=\left|\log\frac{1-q}{q}\right|\leq 5\left(\frac{1}{2}-q\right), \quad (78)$$

and for any $q\in[\frac{1}{10},\frac{1}{2}]$ it holds that $|h_b''(q)|\leq 12$. Hence, by (77), and under the high probability event

$$|h_b(\hat{p}(n))-h_b(p)|\leq 5\left|\frac{1}{2}-p\right||\hat{p}(n)-p|+6(\hat{p}(n)-p)^2$$

$$\leq 7\left|\frac{1}{2}-p\right|\sqrt{\frac{\log(\frac{2}{\delta})}{n}}+\frac{9\log(\frac{2}{\delta})}{n}.$$

The proof of (21) is completed by replacing δ with 2δ. The proof of (20) is similar, with a Taylor approximation for p around $\hat{p}(n)$. □

APPENDIX D
PROOFS FOR SECTION V-B

The proof of Proposition 9 relies on a confidence interval bound for the entropy which is based on empirical version of $\zeta(p)$. We begin with the following bound.

Lemma 15. *Consider the setting of Proposition 9. Then, for any* $\delta \in (0,1)$

$$\mathrm{d_{TV}}(p,\hat{p}(n)) \leq \sqrt{\frac{4\zeta(p)|\mathcal{Y}|+\log\left(\frac{1}{\delta}\right)}{n}}, \tag{79}$$

with probability larger than $1-\delta$.

Proof. The total variation $\mathrm{d_{TV}}(p,\hat{p}(n))$ satisfies a bounded difference inequality with constant $1/n$ as a function of $(Y_1,\ldots,Y_n)$, and so by McDiarmid's inequality [46, Theorem 3.11]

$$\mathbb{P}\left[\left|\mathrm{d_{TV}}(p,\hat{p}(n))-\mathbb{E}\left[\mathrm{d_{TV}}(p,\hat{p}(n))\right]\right|\geq\epsilon\right]\leq e^{-2n\epsilon^2}. \tag{80}$$

Recall that $\hat{p}(n,y)=\frac{1}{n}\sum_{\ell=1}^{n}\mathbb{1}\{Y_i=y\}$. We next upper bound the expected value $\mathbb{E}[\mathrm{d_{TV}}(p,\hat{p}(n))]$ as follows:

$$\begin{aligned}
\mathbb{E}\left[\mathrm{d_{TV}}(p,\hat{p}(n))\right] &= \sum_{y\in\mathcal{Y}}\mathbb{E}\left[|p(y)-\hat{p}(n,y)|\right]\\
&\leq \sum_{y\in\mathcal{Y}}\sqrt{\mathbb{E}\left[(p(y)-\hat{p}(n,y))^2\right]}\\
&= \sum_{y\in\mathcal{Y}}\sqrt{\frac{2}{n}p(y)(1-p(y))}\\
&\leq |\mathcal{Y}|\sqrt{\frac{1}{|\mathcal{Y}|}\sum_{y\in\mathcal{X}}\frac{2}{n}p(y)(1-p(y))}\\
&= \sqrt{\frac{2|\mathcal{Y}|}{n}}\sqrt{\sum_{y\in\mathcal{Y}}p(y)(1-p(y))}\\
&= \sqrt{\frac{2\zeta(p)|\mathcal{Y}|}{n}},
\end{aligned}$$

where the two inequalities follow from Jensen's inequality. Setting $e^{-2n\epsilon^2} = \delta$ directly leads to

$$\mathrm{d_{TV}}(p, \hat{p}(n)) \leq \sqrt{\frac{2\zeta(p)|\mathcal{Y}|}{n}} + \sqrt{\frac{1}{2n}\log\left(\frac{1}{\delta}\right)}, \tag{81}$$

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 $\mathrm{d_{TV}}(p, \hat{p}(n))$, 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 16. *Under the setting of Lemma 15, let the plug-in estimator of $\zeta(p)$ be given by $\hat{\zeta}(n) \equiv \hat{\zeta}(\boldsymbol{Y}, n) := 1 - \sum_{y\in\mathcal{Y}} \hat{p}^2(n, y)$. Then, for any $\delta \in (0, 1)$ it holds that*

$$\hat{\zeta}(n) - \sqrt{\frac{18\log\left(\frac{1}{\delta}\right)}{n}} - \frac{1}{n} \leq \zeta(p) \leq \hat{\zeta}(n) + \sqrt{\frac{18\log\left(\frac{1}{\delta}\right)}{n}}, \tag{82}$$

with probability larger than $1 - \delta$.

Proof. Since $|(p(\boldsymbol{Y}, n) \pm \frac{1}{n})^2 - p(\boldsymbol{Y}, n)^2| \leq \frac{3}{n}$ for any $p(\boldsymbol{Y}, n) \in [0, 1]$, the plug-in estimator $\hat{\zeta}(n) \equiv \hat{\zeta}(\boldsymbol{Y}, n)$ satisfies a bounded difference inequality with constant $6/n$ as a function of $(Y_1, \ldots, Y_n)$, and so by McDiarmid's inequality [46, Theorem 3.11]

$$\mathbb{P}\left[\left|\hat{\zeta}(\boldsymbol{Y}, n) - \mathbb{E}\left[\hat{\zeta}(\boldsymbol{Y}, n)\right]\right| \geq \epsilon\right] \leq e^{-\frac{n\epsilon^2}{18}}. \tag{83}$$

The plug-in estimator $\hat{\zeta}(n)$ is biased, and easily seen to satisfy $\mathbb{E}\left[\hat{\zeta}(n)\right] = \zeta(p) + \frac{\zeta(p)}{n}$. The result follows since $\zeta(p) \in [0, 1]$. □

We combine Lemma 15 and Lemma 16 to obtain a confidence interval bound which can be computed by the player according to its empirical data.

Lemma 17. *Under the setting of Lemma 15, any $\delta \in (0, 1/e)$ it holds that*

$$\mathrm{d_{TV}}(p, \hat{p}(n)) \leq \sqrt{\frac{4\hat{\zeta}(n)|\mathcal{Y}|}{n}} + \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{n}} + \frac{5|\mathcal{Y}|^{3/4}}{n^{3/4}}\log^{1/4}\left(\frac{2}{\delta}\right), \tag{84}$$

with probability larger than $1 - \delta$.

Proof. By combining Lemma 15 and Lemma 16, and a union bound, it holds with probability larger than $1-2\delta$ that

$$\begin{aligned}\mathrm{d_{TV}}(p,\hat{p}(n)) &\leq \sqrt{\frac{4\zeta(p)|\mathcal{Y}|+\log\left(\frac{1}{\delta}\right)}{n}} \\ &\leq \sqrt{\frac{4\left[\hat{\zeta}(n)+\sqrt{\frac{18\log\left(\frac{1}{\delta}\right)}{n}}\right]|\mathcal{Y}|+\log\left(\frac{1}{\delta}\right)}{n}} \\ &\leq \sqrt{\frac{4\hat{\zeta}(n)|\mathcal{Y}|}{n}}+\left[\frac{288|\mathcal{Y}|^2\log\left(\frac{1}{\delta}\right)}{n^3}\right]^{1/4}+\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{n}},\end{aligned}$$

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 δ with 2δ. □

With these results at hand we may prove Proposition 9.

Proof of Proposition 9. As in the proof of Proposition 3 if $\mathrm{d_{TV}}(p,\hat{p}(n))\leq\frac{1}{2}$ then

$$|H(\hat{p}(n))-H(p)|\leq-|\mathcal{Y}|\cdot\Lambda_1\left(\frac{\mathrm{d_{TV}}(p,\hat{p}(n))}{|\mathcal{Y}|}\right). \tag{85}$$

From Lemma 17

$$\mathrm{d_{TV}}(p,\hat{p}(n))\leq\sqrt{\frac{4\hat{\zeta}(n)|\mathcal{Y}|}{n}}+\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{n}}+\frac{5|\mathcal{Y}|^{1/2}}{n^{3/4}}\log^{1/4}\left(\frac{2}{\delta}\right):=a_1+a_2+a_3, \tag{86}$$

with probability larger than $1-\delta$, where $\{a_i\}_{i\in[3]}$ 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

$$\frac{\mathrm{d_{TV}}(p,\hat{p}(n))}{|\mathcal{Y}|}\leq\sqrt{\frac{4\hat{\zeta}(n)}{n|\mathcal{Y}|}}+\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{n|\mathcal{Y}|^2}}+\frac{5}{n^{3/4}|\mathcal{Y}|^{1/2}}\log^{1/4}\left(\frac{2}{\delta}\right)\leq\frac{1}{e}. \tag{87}$$

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 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 $\mathrm{d_{TV}}(p,\hat{p}(n))$ with the high probability upper bound (84) of Lemma 17. We may consider three cases, according to which of the terms is the largest.

- If $\max_{i\in[3]}a_i=a_1$ then the upper bound (84) is less than $3a_1=\sqrt{\frac{36\hat{\zeta}(n)|\mathcal{Y}|}{n}}$. By the monotonicity

property, (85) results

$$|H(\hat{p}(n)) - H(p)| \leq \sqrt{\frac{36\hat{\zeta}(n)|\mathcal{Y}|}{n}} \log\left(\sqrt{\frac{n|\mathcal{Y}|}{36\hat{\zeta}(n)}}\right) \leq 3\sqrt{\frac{\hat{\zeta}(n)|\mathcal{Y}|}{n}} \log\left(\frac{n|\mathcal{Y}|}{36\hat{\zeta}(n)}\right). \tag{88}$$

- If $\max_{i\in[3]} a_i = a_2$ then the upper bound (84) is less than $3a_2 = \sqrt{\frac{9\log\left(\frac{2}{\delta}\right)}{n}}$. By the monotonicity property, (85) results

$$|H(\hat{p}(n)) - H(p)| \leq \sqrt{\frac{9\log\left(\frac{2}{\delta}\right)}{n}} \log\left(\sqrt{\frac{n|\mathcal{Y}|^2}{9\log\left(\frac{2}{\delta}\right)}}\right) \leq \frac{3}{2}\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{n}} \log\left(\frac{n|\mathcal{Y}|^2}{9}\right). \tag{89}$$

- If $\max_{i\in[3]} a_i = a_3$ then the upper bound (84) is less than $3a_3 = \frac{15|\mathcal{Y}|^{3/4}}{n^{3/4}} \log^{1/4}\left(\frac{2}{\delta}\right)$. By the monotonicity property, (85) results

$$\begin{aligned} |h_b(H(n)) - H(p)| &\leq \frac{15|\mathcal{Y}|^{1/2}\log^{1/4}\left(\frac{2}{\delta}\right)}{n^{3/4}} \log\left(\frac{n^{3/4}|\mathcal{Y}|^{1/2}}{15\log^{1/4}\left(\frac{2}{\delta}\right)}\right) \\ &\leq \frac{2|\mathcal{Y}|^{1/2}\log^{1/4}\left(\frac{2}{\delta}\right)\log\left(n|\mathcal{Y}|^{2/3}\right)}{n^{3/4}}. \end{aligned}$$

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 10, which is based on a lemma analogous to Lemma 14. To this end, we further denote a simplified version of $\Gamma_{\text{tv}}(\cdot)$ from (28) defined as

$$\tilde{\Gamma}_{\text{tv}}(\alpha, \zeta, \Delta, t) :=$$
$$\max\left\{144\frac{\zeta}{|\mathcal{Y}|}\Lambda_1^2\left(\frac{2|\mathcal{Y}|}{3\Delta}\right),\ \frac{135}{|\mathcal{Y}|^2}\Lambda_2\left(\frac{9|\mathcal{Y}|^2\alpha\log(t)}{\Delta^2}\right),\ \frac{3}{|\mathcal{Y}|^{2/3}}\Lambda_{4/3}\left(\frac{27|\mathcal{Y}|^{4/3}\alpha^{1/3}\log^{1/3}(t)}{\Delta^{4/3}}\right)\right\}. \tag{90}$$

Lemma 18. *For $\delta \equiv \delta_\alpha(t) = 2t^{-\alpha}$ and $\alpha > 2$, if $n \geq \tilde{\Gamma}_{\text{tv}}(\alpha, \zeta, \Delta, t)$ then $\text{UCB}_{\text{tv}}(\zeta, \delta, \mathcal{Y}, n) \leq \Delta/2$, where $\text{UCB}_{\text{tv}}(\zeta, \delta, \mathcal{Y}, n)$ is as defined in (26).*

Proof. We may assume[3] that $n \geq e$. Then, $\text{UCB}_{\text{tv}}(\zeta, \mathcal{Y}, \delta, n) \leq \Delta/2$ if all three conditions hold[4]

$$3\sqrt{\frac{\zeta|\mathcal{Y}|}{n}} \log\left(\frac{n|\mathcal{Y}|}{36\zeta}\right) \leq \Delta/6, \tag{91}$$

[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 β (see the proof of Lemma 14).

and

$$\frac{3}{2}\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{n}}\log\left(\frac{n|\mathcal{Y}|^2}{9}\right) \leq \Delta/6, \tag{92}$$

as well as

$$\frac{2|\mathcal{Y}|^{1/2}\log^{1/4}\left(\frac{2}{\delta}\right)\log\left(n|\mathcal{Y}|^{2/3}\right)}{n^{3/4}} \leq \Delta/6. \tag{93}$$

For the first condition, we write it equivalently as

$$\frac{\log\left(\sqrt{\frac{n|\mathcal{Y}|}{36\zeta}}\right)}{\sqrt{\frac{n|\mathcal{Y}|}{36\zeta}}} \leq \frac{3\Delta}{2|\mathcal{Y}|}. \tag{94}$$

Lemma 11 then implies that this condition is satisfied if

$$n \geq 144\frac{\zeta}{|\mathcal{Y}|}\Lambda_1^2\left(\frac{2|\mathcal{Y}|}{3\Delta}\right). \tag{95}$$

For the second condition, we write it equivalently as

$$\frac{\log^2\left(\frac{n|\mathcal{Y}|^2}{9}\right)}{\frac{n|\mathcal{Y}|^2}{9}} \leq \frac{\Delta^2}{9|\mathcal{Y}|^2\log\left(\frac{2}{\delta}\right)}. \tag{96}$$

Lemma 11 implies that this condition is satisfied if

$$n \geq \frac{135}{|\mathcal{Y}|^2}\Lambda_2\left(\frac{9|\mathcal{Y}|^2\log\left(\frac{2}{\delta}\right)}{\Delta^2}\right). \tag{97}$$

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

$$\frac{\log^{4/3}\left(n|\mathcal{Y}|^{2/3}\right)}{n|\mathcal{Y}|^{2/3}} \leq \frac{\Delta^{4/3}}{27|\mathcal{Y}|^{4/3}\log^{1/3}\left(\frac{2}{\delta}\right)}. \tag{98}$$

Lemma 11 implies that this condition is satisfied if

$$n \geq \frac{3}{|\mathcal{Y}|^{2/3}}\Lambda_{4/3}\left(\frac{27|\mathcal{Y}|^{4/3}\log^{1/3}\left(\frac{2}{\delta}\right)}{\Delta^{4/3}}\right). \tag{99}$$

The claim of the lemma then follows from the definition of $\tilde{\Gamma}_{\text{tv}}(\cdot)$ in (90). □

We may now prove Theorem 10.

Proof of Theorem 10. The proof begins as the proof of Theorem 5. We then define the events:

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

$$\hat{H}(\boldsymbol{X}_{i^*}(t-1), N_{i^*}(t-1))) + \text{UCB}(\boldsymbol{X}_{i^*}(t-1), \delta_\alpha(t), N_{i^*}(t-1)) \leq H_{i^*}, \tag{100}$$

or

$$\hat{\zeta}\left(\boldsymbol{X}_{i^*}(t-1), N_{i^*}(t-1)\right) - \zeta(p_{i^*}) \leq -\sqrt{\frac{18\log\left(\frac{1}{\delta}\right)}{N_{i^*}(t-1)}}. \tag{101}$$

II". Either the entropy of arm i is significantly overestimated

$$\hat{H}(\boldsymbol{X}_i(t-1), N_i(t-1))) > H_i + \mathrm{UCB}(\boldsymbol{X}_i(t-1), \delta_\alpha(t), N_i(t-1)), \tag{102}$$

or

$$\hat{\zeta}\left(\boldsymbol{X}_i(t-1), N_i(t-1)\right) - \zeta(p_i) \geq \sqrt{\frac{18\log\left(\frac{1}{\delta}\right)}{N_i(t-1)}} + \frac{1}{N_i(t-1)}. \tag{103}$$

III". The upper confidence interval, which is based on an overestimation of $\hat{\zeta}(\boldsymbol{X}_i(t-1), N_i(t-1))$ is significantly larger than the gap:

$$\mathrm{UCB}_{\text{tv}}\left(\zeta(p_i) + \sqrt{\frac{18\log\left(\frac{1}{\delta}\right)}{N_i(t-1)}}, \delta_\alpha(t), N_i(t-1)\right) > \frac{\Delta_i}{2}, \tag{104}$$

or

$$N_i(t-1) \leq \max\left\{30 \cdot \log\left(\frac{2}{\delta}\right), 119\zeta(p_i)|\mathcal{Y}|\right\}. \tag{105}$$

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

$$\begin{aligned}
&\hat{H}(\boldsymbol{X}_{i^*}(t-1), N_{i^*}(t-1))) + \mathrm{UCB}(\boldsymbol{X}_{i^*}(t-1), \delta_\alpha(t), N_{i^*}(t-1)) \\
&\geq H_{i^*} = H_i + \Delta_i \\
&\geq H_i + 2\mathrm{UCB}_{\text{tv}}\left(\zeta(p_i) + \sqrt{\frac{18\log\left(\frac{1}{\delta}\right)}{N_i(t-1)}}, \delta_\alpha(t), N_i(t-1)\right) \\
&\overset{(*)}{\geq} H_i + 2\mathrm{UCB}\left(\boldsymbol{X}_i(t-1), \delta_\alpha(t), N_i(t-1)\right) \\
&\geq \hat{H}(\boldsymbol{X}_i(t-1), N_i(t-1))) + \mathrm{UCB}\left(\boldsymbol{X}_i(t-1), \delta_\alpha(t), N_i(t-1)\right),
\end{aligned}$$

where in $(*)$ we have used the current assumption that (105) does not hold, which assures that $\mathrm{UCB}_{\text{tv}}(\zeta, \delta_\alpha(t), N_i(t-1))$ is a monotonically non-decreasing function of ζ. Thus, in this case Algorithm 1 will not choose $I_t = i$ at the tth round; a contradiction.

By Lemma 18 if

$$N_i(t-1) \geq \tilde{\Gamma}_{\text{tv}}\left(\alpha, \zeta(p_i) + \sqrt{\frac{18\log\left(\frac{1}{\delta}\right)}{N_i(t-1)}}, \Delta_i, t\right), \tag{106}$$

then event III" does not occur. By the definition of $\tilde{\Gamma}_{\text{tv}}(\cdot)$ in (90), and by setting $\delta_\alpha(t) = t^{-\alpha}$, the RHS

in the last equation is upper bounded as

$$\max\left\{144\frac{\zeta(p_i)}{|\mathcal{Y}|}\Lambda_1^2\left(\frac{2|\mathcal{Y}|}{3\Delta_i}\right)+576\frac{\sqrt{18\log\left(\frac{1}{\delta}\right)}}{|\mathcal{Y}|\sqrt{N_i(t-1)}}\Lambda_1^2\left(\frac{2|\mathcal{Y}|}{3\Delta_i}\right)\right.$$
$$\left., \frac{135}{|\mathcal{Y}|^2}\Lambda_2\left(\frac{9|\mathcal{Y}|^2\alpha\log(t)}{\Delta_i^2}\right), \frac{3}{|\mathcal{Y}|^{2/3}}\Lambda_{4/3}\left(\frac{27|\mathcal{Y}|^{4/3}\alpha^{1/3}\log^{1/3}(t)}{\Delta_i^{4/3}}\right)\right\}.$$

This can be guaranteed by requiring that $N_i(t-1)$ is larger than twice of each of the additive components of the first term, as well as larger than each of the second and third terms. To conclude, a sufficient condition for the event III" not to occur is that

$$N_i(t-1)\geq\max\left\{288\frac{\zeta(p_i)}{|\mathcal{Y}|}\Lambda_1^2\left(\frac{2|\mathcal{Y}|}{3\Delta_i}\right), 36230\frac{\alpha^{1/3}\log^{1/3}(t)}{|\mathcal{Y}|^{2/3}}\Lambda_1^{4/3}\left(\frac{2|\mathcal{Y}|}{3\Delta_i}\right)\right.$$
$$\left., \frac{135}{|\mathcal{Y}|^2}\Lambda_2\left(\frac{9|\mathcal{Y}|^2\alpha\log(t)}{\Delta_i^2}\right), \frac{3}{|\mathcal{Y}|^{2/3}}\Lambda_{4/3}\left(\frac{27|\mathcal{Y}|^{4/3}\alpha^{1/3}\log^{1/3}(t)}{\Delta_i^{4/3}}\right)\right\}. \quad (107)$$

This condition, along with the second part of event III" is then used to define $\Gamma_{\text{ber}}(\alpha,\zeta(p_i),\Delta_i,t)$ in (28). The analysis then follows as in the proof of Theorem 5, by using Lemma 17 and Proposition 9 to bound the probabilities of the events in I" and II". □

REFERENCES

[1] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[2] P. Rusmevichientong and J. N. Tsitsiklis, "Linearly parameterized bandits," *Mathematics of Operations Research*, vol. 35, no. 2, pp. 395–411, 2010.

[3] M. K. Hanawal, A. Leshem, and V. Saligrama, "Efficient algorithms for linear polyhedral bandits," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 4796–4800.

[4] V. Dani, T. P. Hayes, and S. M. Kakade, "Stochastic linear optimization under bandit feedback," in *Conference on Learning Theory (COLT)*, 2008.

[5] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in *Advances in Neural Information Processing Systems 24*, 2011, pp. 2312–2320.

[6] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part II: Markovian rewards," *IEEE Transactions on Automatic Control*, vol. 32, no. 11, pp. 977–982, November 1987.

[7] 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.

[8] P. L. Bartlett and A. Tewari, "REGAL: a regularization based algorithm for reinforcement learning in weakly communicating MDPs," in *UAI*, 2009.

[9] R. Fruit, M. Pirotta, and A. Lazaric, "Near optimal exploration-exploitation in non-communicating Markov decision processes," in *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, ser. NIPS'18, 2018, pp. 2998–3008.

[10] M. Yemini, A. Leshem, and A. Somekh-Baruch, "The restless hidden markov bandit with linear rewards and side information," *IEEE Transactions on Signal Processing*, vol. 69, pp. 1108–1123, 2021.

[11] T. Gafni, M. Yemini, and K. Cohen, "Learning in restless bandits under exogenous global markov process," 2021.

[12] W. Chen, W. Hu, F. Li, J. Li, Y. Liu, and P. Lu, "Combinatorial multi-armed bandit with general reward functions," in *Advances in Neural Information Processing Systems*, vol. 29. Curran Associates, Inc., 2016.

[13] M. Haidar Sharif and C. Djeraba, "An entropy approach for abnormal activities detection in video streams," *Pattern Recognition*, vol. 45, no. 7, pp. 2543–2561, 2012.

[14] C. Callegari, S. Giordano, and M. Pagano, "Entropy-based network anomaly detection," in *2017 International Conference on Computing, Networking and Communications (ICNC)*, 2017, pp. 334–340.

[15] R. Hu, H. Pham, P. Buluschek, and D. Gatica-Perez, "Elderly people living alone: Detecting home visits with ambient and wearable sensing," in *Proceedings of the 2nd International Workshop on Multimedia for Personal Health and Health Care, MMHealth@MM 2017, Mountain View, CA, USA, October 23 - 27, 2017*. ACM, 2017, pp. 85–88.

[16] 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, 2020.

[17] P. Auer, "Using confidence bounds for exploitation-exploration trade-offs," *Journal of Machine Learning Research (JMLR)*, vol. 3, pp. 397–422, 2002.

[18] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *arXiv preprint arXiv:1204.5721*, 2012.

[19] D. Dubhashi and A. Panconesi, *Concentration of Measure for the Analysis of Randomized Algorithms*, 1st ed. Cambridge University Press, 2009.

[20] M. Raginsky and I. Sason, "Concentration of measure inequalities in information theory, communications, and coding," *Foundations and TrendsÂ® in Communications and Information Theory*, vol. 10, no. 1-2, pp. 1–246, 2013.

[21] L. Paninski, "Estimation of entropy and mutual information," *Neural Comput.*, vol. 15, no. 6, pp. 1191–1253, Jun. 2003.

[22] A. Antos and I. Kontoyiannis, "Convergence properties of functional estimates for discrete distributions," *Random Structures & Algorithms*, vol. 19, no. 3-4, pp. 163–193, 2001.

[23] S.-W. Ho and R. W. Yeung, "The interplay between entropy and variational distance," *IEEE Transactions on Information Theory*, vol. 56, no. 12, pp. 5906–5929, 2010.

[24] I. Sason, "Entropy bounds for discrete random variables via maximal coupling," *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 7118–7131, 2013.

[25] P. Grassberger, "Entropy estimates from insufficient samplings," *CoRR*, vol. abs/0307138, 2008.

[26] I. Nemenman, "Coincidences and estimation of entropies of random variables with large cardinalities," *Entropy*, vol. 13, no. 12, pp. 2013–2023, 2011.

[27] A. Chao, Y. T. Wang, and L. Jost, "Entropy and the species accumulation curve: a novel entropy estimator via discovery rates of new species," *Methods in Ecology and Evolution*, vol. 4, no. 11, pp. 1091–1100, 2013.

[28] E. W. Archer, I. M. Park, and J. W. Pillow, "Bayesian entropy estimation for binary spike train data using parametric prior knowledge," in *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[29] P. Valiant and G. Valiant, "Estimating the unseen: Improved estimators for entropy and other properties," in *Advances in Neural Information Processing Systems*, vol. 26, 2013.

[30] T. Schürmann, "A Note on Entropy Estimation," *Neural Computation*, vol. 27, no. 10, pp. 2097–2106, 10 2015.

[31] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Minimax estimation of functionals of discrete distributions," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2835–2885, 2015.

[32] ——, "Maximum likelihood estimation of functionals of discrete distributions," *IEEE Transactions on Information Theory*, vol. 63, no. 10, pp. 6774–6798, 2017.

[33] D. Russo and B. Van Roy, "Learning to optimize via information-directed sampling," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014. [Online]. Available: https://proceedings.neurips.cc/paper/2014/file/301ad0e3bd5cb1627a2044908a42fdc2-Paper.pdf

[34] J. Kirschner and A. Krause, "Information directed sampling and bandits with heteroscedastic noise," in *Proceedings of the 31st Conference On Learning Theory*, ser. Proceedings of Machine Learning Research, vol. 75, ul 2018, pp. 358–384.

[35] J. Kirschner, T. Lattimore, C. Vernade, and C. Szepesvari, "Asymptotically optimal information-directed sampling," in *Proceedings of Thirty Fourth Conference on Learning Theory*, ser. Proceedings of Machine Learning Research, vol. 134, Aug 2021, pp. 2777–2821.

[36] W. R. Thompsom, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3-4, pp. 285–294, 12 1933.

[37] D. Russo and B. V. Roy, "An information-theoretic analysis of Thompson sampling," *Journal of Machine Learning Research*, vol. 17, no. 68, pp. 1–30, 2016. [Online]. Available: http://jmlr.org/papers/v17/14-087.html

[38] A. Gopalan, S. Mannor, and Y. Mansour, "Thompson sampling for complex online problems," *31st International Conference on Machine Learning, ICML 2014*, vol. 1, pp. 169–186, 01 2014.

[39] A. Slivkins, "Introduction to multi-armed bandits," *arXiv preprint arXiv:1904.07272*, 2019.

[40] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2, pp. 235–256, 2002.

[41] J.-Y. Audibert, R. Munos, and C. Szepesvári, "Exploration–exploitation tradeoff using variance estimates in multi-armed bandits," *Theoretical Computer Science*, vol. 410, no. 19, pp. 1876–1902, 2009.

[42] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge, U.K.: Cambridge University Press, 2011.

[43] A. Orlitsky, A. T. Suresh, and Y. Wu, "Optimal prediction of the number of unseen species," *Proceedings of the National Academy of Sciences*, vol. 113, no. 47, pp. 13 283–13 288, 2016.

[44] Y. Wu and P. Yang, "Chebyshev polynomials, moment matching, and optimal estimation of the unseen," *The Annals of Statistics*, vol. 47, no. 2, pp. 857 – 883, 2019.

[45] M. Mitzenmacher and E. Upfal, *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press, 2017.

[46] R. van Handel, "Probability in high dimension," Princeton Uiversity New Jersey, Tech. Rep., 2014.