Abstract—In the regret-based formulation of multi-armed Bandit (MAB) problems, except in rare instances, much of the literature focuses on arms with independent and identically distributed (i.i.d.) rewards. In this article, we consider the problem of obtaining regret guarantees for MAB problems, in which the rewards of each arm form a Markov chain that may not belong to a single parameter exponential family. To achieve a logarithmic regret in such problems is not difficult: a variation of standard Kullback–Leibler upper confidence bound (KL-UCB) does the job. However, the constants obtained from such an analysis are poor for the following reason: i.i.d. rewards are a special case of Markov rewards and it is difficult to design an algorithm that works well independent of whether the underlying model is truly Markovian or i.i.d. To overcome this issue, we introduce a novel algorithm that identifies whether the rewards from each arm are truly Markovian or i.i.d. using a total variation distance-based test. Our algorithm then switches from using a standard KL-UCB to a specialized version of KL-UCB when it determines that the arm reward is Markovian, thus resulting in low regrets for both i.i.d. and Markovian settings.

Index Terms—Kullback–Leibler upper confidence bound (KL-UCB), multi-armed bandit (MAB), online learning, regret, rested bandit.

I. INTRODUCTION

In a multi-armed Bandit (MAB) problem [1], [2], a player needs to choose arms sequentially to maximize the total reward or minimize the total regret [3]. Unlike independent and identically distributed (i.i.d.) arm reward case [2], [4], [5], [6], [7], [8], [9], there are only a few works that consider bandits with Markovian rewards. When the states of the arms that are not played remain frozen (rested bandit) [10], [11], [12], [13], the states observed in their next selections do not depend on the interval between successive plays of the arms. In the restless [12], [14], [15] case, states of the arms continue to evolve irrespective of the selections by the player. In this article, we focus on the rested bandit setting. Many real-world problems, such as gambling, ad placement, and clinical trials fall into this category [16]. If a slot-machine (viewed as an arm) produces a high reward in a particular play, then the probability that it will produce a high reward in the next play is not very high. Therefore, it is reasonable to assume that the reward distribution depends on the previous outcome and can be modeled through the rested bandit setting. Another example is an online advertising setting to a customer [e.g., video advertisements (ads) with IMdb television, social ads on Facebook], where an agent presents an ad to a customer from a pool of ads and adapts future displayed ads based on the customer’s past responses. Other examples are medical diagnosis problem (repeated interventions of different medications), click-through rate prediction, and rating systems [17], [18].

Existing studies in the literature on rested bandits either consider a particular class of Markov chains (with single-parameter families of transition matrices) [11], [13], or result in large asymptotic regret bounds [10], [12]. In general, however, these Markov chains may not belong to a single parameter family of transition matrices; nevertheless, we would desire low regret. In the setting considered by us, we assume that each arm evolves according to a two-state Markov chain. When an arm is chosen, a finite reward is generated based on the present state of the arm. Imagine a scenario where a company (Netflix, say) recommends a set of products (places movies in the user’s browser) to users (subscribers) to maximize the generated revenue. A user’s response to a recommended product (a movie genre) can be captured using a MAB approach, the company can adapt the set of displayed products [20]. In such a setting, we can obtain a logarithmic regret by a variant of standard Kullback–Leibler upper confidence bound (KL-UCB) [8], [9] (originally designed for i.i.d. rewards). Although the constant obtained from the regret analysis is better than that of [10], the performance is poor for i.i.d. rewards (a special case of Markovian rewards). We address this problem in this article. The main contributions of this article are as follows.

A. Adaptive KL-UCB Algorithm

We propose a novel total variation KL-UCB (TV-KL-UCB) algorithm, which can identify whether the rewards from an arm are truly Markovian or i.i.d. The identification is done using a (time-dependent) TV distance-based test between the empirical estimates of transition probabilities from one state to another. The proposed algorithm switches from the standard sample mean-based KL-UCB to a sample transition probability-based KL-UCB when it determines that the arm reward is Markovian using the test. The sample transition probability-based KL-UCB determines the UCB on the transition probability from the current state of the arm and uses the estimate of the transition probability to the current state while evaluating the mean reward. An arm can be represented uniquely by the transition probability matrix (two parameters), however, requiring a single parameter (mean reward) in an i.i.d setting. Our approach adapts itself from learning two parameters for truly Markovian arms to learning a single parameter for i.i.d. arms, thereby producing low regrets in both cases. An extension to multistate Markovian arms is also described.
B. Upper Bound on Regret

We derive finite time and asymptotic upper bounds on the regrets of TV-KL-UCB (capturing the worst-case regret [19]) for both truly Markovian and i.i.d. rewards. We prove that TV-KL-UCB is order-optimal for Markovian rewards and optimal when all arm rewards are i.i.d.

The standard analysis for regret with KL-UCB (either for i.i.d. rewards [8], or rewards from a Markov transition matrix specified through a single-parameter exponential family [5, 11]) crucially relies on the invertibility of the KL divergence function when applied to empirical estimates. This allows one to translate concentration guarantees for sums of random variables to one for level crossing of the KL divergence function applied to empirical estimates. In our multiparameter estimation setting, this invertibility property no longer directly holds. However, we convert this multiparameter problem into a collection of single parameter problems and derive concentration bounds for individual single parameter problems that are combined to obtain an upper bound on the regret of TV-KL-UCB. To derive our bound, we establish that a certain condition on the TV distance between estimates of transition probabilities (for using a sample mean-based KL-UCB) is satisfied infinitely often if the arm rewards are i.i.d. over time, which in turn implies that the regret due to choosing the incorrect variant of KL-UCB vanishes asymptotically.

Analytical and experimental results establish that TV-KL-UCB performs better than the state-of-the-art algorithms [10, 13] when at least one of the arm rewards is truly Markovian. Moreover, TV-KL-UCB is optimal [8] when all arm rewards are i.i.d.

Related work: While much of the literature on MAB focuses on arms with i.i.d. rewards, arms with Markovian rewards have not been studied extensively. In [2], when the parameter space is dense and can be represented using a single-parameter density function, a lower bound on the regret is derived. Lai and Robbins [2] also proposed policies that asymptotically achieve the lower bound. The work in [2] is extended in [5] for the case when multiple arms can be played at a time. A sample mean-based index policy in [6] achieves a logarithmic regret for a one-parameter family of distributions. The KL-UCB-based index policy [8], [9] is asymptotically optimal for Bernoulli rewards and performs better than the UCB policy [7]. Zheng and Hua [21] provided an overview of the state-of-the-art on Markovian bandits. Under the assumption of single-parameter families of transition matrices, an index policy that matches the corresponding lower bound asymptotically is proposed in [11]. In [10], a UCB policy based on the sample mean reward is proposed. Unlike [11], the analysis in [10] is not restricted to a single-parameter family of transition matrices. Moreover, since the index calculation is based on the sample mean, the policy is significantly simpler than that of [11]. Although order-optimal, the proposed policy may be worse than that in [11] in terms of the constant. In [13], an extension of KL-UCB using the sample mean is proposed for the optimal allocation problem involving multiple plays. Similar to [11], rewards are generated from Markov chains belonging to a one-parameter exponential family.

Unlike i.i.d. rewards, in many practical applications, temporal variation in reward distribution is present [22], [23], [24] ranging from Markovian to general historical-dependency rewards and adversarial rewards [25], [26]. In [27], regret analysis for a large class of MAB problems involving nonstationary rewards is conducted. A relation between the rate of variation in reward (limited by a variation budget unlike the unbounded adversarial setup, however including a large set of nonstationary stochastic MABs) and minimal regret is established in [27]. Low-regret algorithms such as to learn the optimal policy in Markov decision processes (MDPs). UCRL2 [28] switches between computing the optimal policy with the largest optimal gain in each phase for the MDP and implementing the policy, based on a state-action visit criteria. The authors in [29] and [30] improve upon UCRL2 by demonstrating better finite time behavior and weaker dependence on the diameter and state space of the MDP, respectively.

Our contributions: Existing works on rested Markovian bandits either consider only single-parameter families of transition matrices

(one-parameter exponential family of Markov chains) [11, 13] or undertake a sample mean-based approach [10, 12] that is closer in spirit to i.i.d bandits than Markovian bandits. We, for the first time, consider a sample transition probability-based approach combined with a sample mean-based approach in [10] and [12] with a significant improvement in performance than that of [10, 12]. Following [10] and unlike [11, 13], we do not consider any parameterization on the transition probability matrices. The only assumption we require is that the Markov chains have to be irreducible. The key reasons why our approach achieves lower regret than the methods in [10] and [12] are 1) usage of confidence bounds on sample transition probabilities for truly Markovian arms coupled with a sample mean based approach for i.i.d. arms), the methods in [10] and [12] always use sample mean-based indices that may not uniquely represent the arms in a truly Markovian setting, and 2) usage of KL-UCB (based on the Chernoff’s bound) provides a tighter confidence bound than the Hoeffding’s bound for UCB [19].

Our analysis is different from that of the Bernoulli bandit [8] in the following way: 1) we determine the contribution of mixing time components of the underlying Markov chains in the finite-time regret upper bound, and 2) we prove that the appropriate conditions for the online test are satisfied infinitely often for both cases (i.i.d and purely Markovian). Since we do not consider only a one-parameter exponential family of Markov chains, it is unclear whether our algorithm is the best among all algorithms for this setting. To show that our approach is asymptotically optimal, one needs to obtain matching upper and lower bounds, which remains an open problem. However, we note that both our theory and simulations show that our results improve upon the performance of the state-of-the-art algorithms.

II. Problem Formulation and Preliminaries

We assume that we have K arms. The reward from each arm is modeled as a two-state irreducible Markov chain with a state space \( S = \{0, 1\} \). Let the reward obtained when an arm, which is in state \( s \), is played be denoted by \( r(s) = s \) (say). Let the transition probability from state \( s = 0(1) \) to state \( s = 1(0) \) of arm \( i \) be denoted by \( p_{ij} \). Let the stationary distribution of arm \( i \) be \( \pi_i = (\pi_i(s), s \in S) \). Therefore, the mean reward of arm \( i = \mu_i \). The mean reward of an arm \( i = \mu_i \). Let the suboptimality of arm \( i \) be \( \Delta_i = \mu_i - \mu^* \). Let \( R_i(n) \) denote the regret of policy \( \alpha \) up to time \( n \). Hence, \( R_i(n) = \pi_i - \sum_{s=0}^{n-1} r(s) \). Hence, \( \alpha(t) \) denotes the arm that is in state \( s_t(t) \). Let \( \alpha(t) \) be uniformly good if \( R_i(n) = o(n) \) for every \( \beta > 0 \). For two arms \( i \) and \( j \) with associated Markov chains \( M_i \) and \( M_j \), the KL distance between them is \( I(M_i||M_j) = \sum_{s=0}^{n-1} \frac{n}{n} \). Let \( D(A|B) = A \log \frac{A}{B} + (1 - A) \log \frac{1 - A}{1 - B} \). In [11], a lower bound on the regret of any uniformly good policy is derived. For a uniformly good policy \( \alpha \), limit inf \( n \to \infty \) \( \frac{1}{\log n} \geq \sum_{i=0}^{n-1} 2^{M_i|M_j} \). In [11], the above lower bound is derived when the transition functions belong to a single-parameter family. It is straightforward to show that the lower bound holds more generally but since the proof techniques are standard, we present the proof as in [31]. We aim to determine an upper bound on \( R_i(n) \) as a function of \( n \) for a given policy \( \alpha \).

Remark 1: The model considered by us is a first step toward capturing the temporal correlations between decisions in a MAB problem. More complicated models involving more than two states would require us to estimate more parameters, leading to higher complexity which can be determined using statistical learning theory. Because of the tradeoff between model complexity and usefulness, we take into account a tradeoff between model complexity and usefulness, we take into account a tradeoff between model complexity and usefulness, we take into account a tradeoff between model complexity and usefulness, we take into account a tradeoff between model complexity and usefulness, we take into account a
Algorithm 1: TV-KL-UCB Algorithm.

1: Input K (number of arms).
2: Choose each arm once.
3: while TRUE do
4: if \((p_{i|0}(t-1) + \hat{p}_{i|0}(t-1)) > \frac{1}{(t-1)^{1/2}}\) (procedure STP_PHASE) then
5: \(U_i = \sup \left\{ \frac{\hat{p}}{\hat{p} + \hat{p}_{i|0}(t-1)} : D(p_{i|0}(t-1)||\hat{p}) \leq \frac{\log f(t)}{T(t-1)} \right\} \quad (1)\)
6: else
7: \(U_i = \sup \left\{ \frac{\hat{p}_{i|0}(t-1)}{\hat{p}_{i|0}(t-1) + \hat{q}} : D(p_{i|0}(t-1)||\hat{q}) \leq \frac{\log f(t)}{T(t-1)} \right\} \quad (2)\)
8: end if
9: end while
10: Choose \(A_t = \arg \max_i U_i\).

choosing an arm is observable. The cardinality of the state space of the
underlying HMM is typically unknown and hence, acts as an additional
hyperparameter to be estimated.

III. TV-KL-UCB ALGORITHM AND REGRET UPPER BOUND

When the rewards of an arm are i.i.d., the arm can be represented
uniquely using the mean reward. However, in the truly Markovian
reward setting, arm \(i\) can be described uniquely by \(p_{i|0}\) and \(p_{i|1}\).
Using a variation (KL-UCB-MC) [31, Appendix IV] of standard KL-UCB
for i.i.d. rewards [8], [9], one can get a logarithmic regret. The main
idea is to obtain a confidence bound on the estimate of \(p_{i|0}\) (estimate of \(p_{i|1}\)) in state \(0\) (state \(1\)) of arm \(i\) using KL-UCB. For purely Markovian, the resulting
regret is smaller than that of [10], [12]. However, KL-UCB-MC results
in large constants in the regret for i.i.d. rewards. Hence, we introduce
the TV-KL-UCB algorithm which improves over KL-UCB-MC and
performs well in both truly Markovian and i.i.d. settings.

A. TV-KL-UCB Algorithm

The proposed TV-KL-UCB that is based on sample transition probabilities
between different states and sample mean, is motivated from
the KL-UCB algorithm [8]. Let \(T_{i,j}(t)\) denote the number of times arm \(i\) is
selected while it was in state \(j\), till time \(t\). We define \(T_{i}(t) := \sum_{j \in S} T_{i,j}(t)\). Further assume that
\(p_{i|0}(t)\), \(\hat{p}_{i|0}(t)\) and \(\mu(t)\) denote the empirical estimate of \(p_{i|0}\),
empirical estimate of \(p_{i|0}\) and sample mean of arm \(i\) at time \(t\),
respectively. We have, \(\mu(t) := \sum_{(i,o) \in \Xi} (i,o) \mu(i,o)\), \(p_{i|0}(t) := \frac{\sum_{(i,o) \in \Xi} (i,o) \mu(i,o)}{T_{i,0}(t)}\), \(\hat{p}_{i|0}(t) := \frac{T_{i,0}(t)}{T_{i,0}(t) + T_{i,1}(t)}\), where \(T_{i,0}(t)\) and \(T_{i,1}(t)\) denote the number of times arm \(i\) is selected while it was in state \(0\) and made a transition to state \(k\), till time \(t\). Hence, \(T_{i,k}(t) := \sum_{j \in S} T_{i,j,k}(t)\). We compute the TV distance (TV distance), say) between \(p_{i|0}(t) - 1\) and \(\hat{p}_{i|0}(t)\) of arm \(i\) which is \(p_{i|0}(t-1) + \hat{p}_{i|0}(t-1) - 1\). If it is greater than \(\frac{1}{(t-1)^{1/2}}\) (Line 4), then based on the current state of the arm, we calculate
the index of the arm (STP_PHASE). If arm \(i\) is in state \(0\), then the index
is calculated using the confidence bound on the estimate of \(p_{i|0}\) at time \(t\) (Line 5), else using the confidence bound on the estimate of \(p_{i|0}\) (Line 6). \(U_i\) is an overestimate of \(\mu_i\) with high probability because
of the considered confidence bounds (Lines 5 and 6). Therefore, a
suboptimal arm cannot be played too often as \(U_i\) overestimates \(\mu_i\), and
a suboptimal arm \(i\) can be played only if \(U_i > U_j\). The index computation
ensures that an arm is explored more often if it is promising (high
\(\hat{p}_{i|0}(t-1) + \hat{p}_{i|0}(t-1) - 1\) or under-explored (small \(T_i(t)\)). We use the estimate of
the transition probability from the other state while evaluating the
index. However, if \(p_{i|0}(t-1) + \hat{p}_{i|0}(t-1) - 1\) is less than \(\frac{1}{(t-1)^{1/2}}\),
then the index of the arm is calculated (Line 8) using the confidence
bound on the current value of \(\mu\) (SM_PHASE). Then, we play the arm
with the highest index. We take \(f(t) = 1 + t \log t\). The physical
interpretation behind the condition on TV distance is that the KL-UCB
algorithm (which uses sample mean) [8] is asymptotically optimal for
i.i.d. Bernoulli arms. When \(p_{i|0} = 1 - p_{i|1}\) (i.i.d arm), the condition
in Line 8 is satisfied frequently often, and arm \(i\) uses (3) for index
calculation, similar to [8]. Else, the condition in Line 4 is met frequently
often. We formally establish these statements in Proposition 4.

Remark 3: The key differences between TV-KL-UCB and KL-
UCB [8] are: 1) the idea of using confidence bound on one of the transition
probabilities while using the raw value of the other one to
determine the index of an arm is novel and 2) the design of an
online test for detecting whether an arm is purely Markovian or not is new,
allowing simultaneous use of sample transition probability and
sample mean to uniquely represent truly Markovian and i.i.d arms,
respectively.

Remark 4: In general, it may not be possible to write closed-form
expressions for (1)–(3) to determine zeroes of convex and increasing
scalar functions [9]. This can be performed by dichotomous search
or Newton iteration. However, for finitely supported distribution (as
considered by us), it can be reduced to the maximization of a linear
function on the probability simplex under KL distance constraints
to obtain an explicit computational solution [32, Appendix C1].

Remark 5: Instead of KL distance which is a natural choice for
representing the similarity between \(p_{i|0}(t)\) and \(1 - p_{i|0}(t)\), we choose the
TV distance. Hellinger distance can be used as it permits additive sep-
 arity of the estimates: \(p_{i|0}(t)\) and \(p_{i|0}(t)\), which in turn enables the
use of standard concentration inequalities for the proof of asymptotic
upper bound on regret (see [31, Appendix IV]). Note that any distance
metric \(L(\cdot)\) that satisfies \(L(1) \leq TV(1)\) for two probability
2
1
2
1
\(\sum_{i \in S} p_{i|0}(t) + p_{i|0}(t)\) \(\sum_{i \in S} p_{i|0}(t) + p_{i|0}(t)\)

2
1
\(\sum_{i \in S} p_{i|0}(t) + p_{i|0}(t)\) \(\sum_{i \in S} p_{i|0}(t) + p_{i|0}(t)\)

2
1
\(\sum_{i \in S} p_{i|0}(t) + p_{i|0}(t)\) \(\sum_{i \in S} p_{i|0}(t) + p_{i|0}(t)\)

2
1
\(\sum_{i \in S} p_{i|0}(t) + p_{i|0}(t)\) \(\sum_{i \in S} p_{i|0}(t) + p_{i|0}(t)\)

2
1
\(\sum_{i \in S} p_{i|0}(t) + p_{i|0}(t)\) \(\sum_{i \in S} p_{i|0}(t) + p_{i|0}(t)\)

2
1
\(\sum_{i \in S} p_{i|0}(t) + p_{i|0}(t)\) \(\sum_{i \in S} p_{i|0}(t) + p_{i|0}(t)\)

2
1
\(\sum_{i \in S} p_{i|0}(t) + p_{i|0}(t)\) \(\sum_{i \in S} p_{i|0}(t) + p_{i|0}(t)\)

2
1
\(\sum_{i \in S} p_{i|0}(t) + p_{i|0}(t)\) \(\sum_{i \in S} p_{i|0}(t) + p_{i|0}(t)\)
Theorem 2: Let the eigenvalue gap of $i$th arm be $\sigma_i$. The asymptotic upper bound on the regret of TV-KL-UCB is smaller than that of [10] (we call it UCB-SM) always (when $\min \sigma_i \geq \frac{1}{1440}$) for i.i.d. (truly Markovian suboptimal) arms.

Algorithm 2: m-TV-KL-UCB.

1: Choose each arm once.
2: while true do
3: Set flag ← 0.
4: for $s, s' = 0$ to $(x - 1), s \neq s'$ do
5: if $TV(\tilde{p}_i(t-1)|\tilde{p}_i(t-1)) > \frac{1}{(x-1)^{1/4}}$ then
6: Set flag ← 1.
7: end for
8: if (flag == 1) (procedure STEP_PHASE) then
9: while (state of arm $i = s$) then
10: end if
11: else (procedure SM_PHASE) (See Algorithm 1)
12: end if
13: Choose $A_t = \arg \max U_i$.
14: end while

C. Extension to Multistate Markovian Model

In this section, we sketch the modifications required to take into account finite-state space ($X$, say with state space $\{0, 1, \ldots, x\}$ and $|X| > 2$) Markov chains to represent the arm rewards. We assume that $r(s) \in [0, 1]$. Let the transition probability from state $a$ to $b$ of arm $i$ be $p_{ab}^i$. Let $p_i^s = [p_i^s(0), p_i^s(1), \ldots, p_i^s(|X|)]$, $\tilde{p}_i(t) = [\tilde{p}_i^0(t), \tilde{p}_i^1(t), \ldots, \tilde{p}_i^{|X|-1}(t)]$, and $\pi_i(s) = [f_{\mu_i}(p_i^s(0)), \ldots, f_{\mu_i}(p_i^s(|X|))]$. We have, $\mu_i = \sum_{x \in X} r(s) \pi_i(s)$. The resulting m-TV-KL-UCB algorithm for the multistate Markovian model is described in Algorithm 2. Similar to Algorithm 1, it is divided into STEP_PHASE and SM_PHASE depending on whether the following online test condition on the TV distance is satisfied

$$TV(\tilde{p}_i(t)||\tilde{p}_i(t-1)) < \frac{1}{(x-1)^{1/4}} \forall s, s' \in \{0, 1, \ldots, x\}, s \neq s'.$$

If this condition is met (flag = 0) at time $t$, we conclude that the arm $i$ is i.i.d. and choose the SM_PHASE. Else, the STEP_PHASE is chosen.

Arm $i$ is i.i.d. (a special case of Markovian arm) if $p_i = \tilde{p}_i^s \forall s, s' \in X, s \neq s'$ (if the condition (Line 11) is satisfied, then the algorithm enters the SM_PHASE (see Algorithm 1). Else, the STEP_PHASE is chosen as the arm is likely to be a truly Markovian arm. If the current state of the arm is $s$, then we compute $U_i$, following (4), where $\tilde{\mu}_i(s) = \sum_{x \in X} r(s) f_{\mu_i}(\tilde{p}_i^s(t-1)). \ldots, f_{\mu_i}(\tilde{p}_i^{|X|-1}(t-1))$. The rest of the algorithm is identical to Algorithm 1.

IV. EXPERIMENTAL EVALUATION

In this section, we compare the performances of TV-KL-UCB and m-TV-KL-UCB with UCB-SM [10]. We also consider an improvement over [10] (referred to as KL-UCB-SM [8]) by replacing the UCB on the sample mean by KL-UCB. Arm $A_t$ is chosen at time $t$ using the following scheme.

$A_t = \arg \max \{\tilde{\mu} : D(\tilde{\mu}(t-1)||\tilde{\mu}) \leq \frac{\log(1/t)}{T(t-1)} \}$. KL-UCB-SM2 is a modification of the algorithm in [13] when only one arm can be played at a time. We consider various scenarios and average the result over 100 runs. Details of the scenarios are provided in [31].

In the first scenario [Fig. 1(a) and (c)], TV-KL-UCB significantly outperforms other algorithms. As KL-UCB provides a tighter confidence bound than UCB, the amount of exploration reduces. Since TV-KL-UCB identifies the best arm using estimates of $p_i^s$, $p_i^s$ in a truly Markovian setting, it spends lesser time in exploration than KL-UCB-SM and KL-UCB-SM2 that work based on the sample mean. In the second scenario, $p_i^s$ and $p_i^s$ for $i \neq 1$ are close to zero. As a result, the suboptimal arms spend a considerable amount of time in their initial states. If the initial estimates of $p_i^0$ and $p_i^0$ are taken to be zero,
then the index of a suboptimal arm remains 1 if it starts at state 0 [see (1)] until it makes a transition to state 1. Therefore, our scheme may lead to a large regret in the beginning until all arms observe a transition to state 1. To address this issue, we initialize \( p_{10} \) and \( p_{01} \) to 1. We observe in Fig. 1(b) and (d) that TV-KL-UCB performs significantly better than UCB-SM, KL-UCB-SM and KL-UCB-SM2. Note that initial values of \( p_{10} \) and \( p_{01} \) do not play much role in the first scenario since each arm makes transitions from one state to another comparably often. The asymptotic upper bound on the regret of TV-KL-UCB is smaller than that of UCB-SM when \( \min \sigma_i \geq \frac{1}{16a^2} \) (see Theorem 2). We choose parameters to satisfy \( \min \sigma_i < \frac{1}{16a^2} \) and observe in Fig. 1(e) that still the asymptotic upper bound on the regret of TV-KL-UCB is better. For the i.i.d. case, the asymptotic upper bounds on the regrets of both KL-UCB-SM and TV-KL-UCB match the corresponding lower bound [8].

We also compare the performance of m-TV-KL-UCB with other algorithms when each arm is a three-state Markov chain with \( r(0) = 0, r(1) = 1/2 \) and \( r(2) = 1 \). In both scenarios, m-TV-KL-UCB performs better than other algorithms as KL-UCB provides a tighter confidence bound than UCB [see Fig. 1(f)-(i)]. Moreover, in truly Markovian settings, as considered by us, m-TV-KL-UCB picks the correct variant of KL-UCB frequently, contrary to other algorithms, such as KL-UCB-SM and KL-UCB-SM2, that work based on the sample mean. The performance improvement achieved by m-TV-KL-UCB is more than that of TV-KL-UCB as the incorrect variant of KL-UCB leads to more regret when the number of states is more.

V. DISCUSSION AND CONCLUSION

TV-KL-UCB uses the TV distance between empirical estimates of \( p_{01} \) and \( 1-p_{10} \) to switch from STP_PHASE to SM_PHASE. Arm \( i \) can be represented uniquely using \( p_{0i} \) and \( p_{1i} \). However, for i.i.d. rewards, it can be represented uniquely using the mean reward. Therefore, usage of STP_PHASE for i.i.d. arms may lead to a large regret since the additional information (which can be exploited by the sample mean based KL-UCB) is not exploited. We design the algorithm so that it works well for both scenarios. When the arms are i.i.d. (truly Markovian), the SM_PHASE (STP_PHASE) is chosen infinitely often. Theoretical (under a mild assumption) and experimental results show that the asymptotic upper bound on the regret of TV-KL-UCB is lower than that of UCB-SM [10]. In the i.i.d. case, the upper bounds on the regrets of KL-UCB-SM and TV-KL-UCB match the lower bound. For TV-KL-UCB, to overcome the issue of high regret associated with zero initialization, we initialize \( p_{10} \) and \( p_{01} \) to 1.

To conclude, in this article, we have designed algorithms that achieve a significant improvement over state-of-the-art bandit algorithms. The idea is to detect whether the arm reward is truly Markovian or i.i.d. using a TV distance based test. We switch from the standard KL-UCB to a sample transition probability-based KL-UCB when we detect that the arm reward is truly Markovian. Logarithmic upper bounds on the regret of TV-KL-UCB have been derived for i.i.d. and Markovian settings. The upper bound on the regret of TV-KL-UCB matches the lower bound for i.i.d. arm rewards.

APPENDIX A

PROOF OF THEOREM 1

We describe a set of useful results before deriving the asymptotic upper bound on the regret. Proof of Proposition 1 is in [31].

**Proposition 1:** Let \( p, q, \epsilon \in [0, 1] \). The following relations hold.
1. \( D(p||q) \geq 2(p-q)^2 \) ( Pinsker’s inequality).
2. If \( p \leq q - \epsilon \leq q \), then \( D(p||q-\epsilon) \leq D(p||q) - 2\epsilon^2 \).
3. If \( p \leq q + \epsilon \leq q \), then \( D(q||p+\epsilon) \leq D(q||p) - 2\epsilon^2 \).

**Corollary 1:** [19, Corollary 10.4] For any \( a > 0 \), \( P(D(\hat{\mu}||\mu) \geq a, \hat{\mu} \leq \mu) \leq \exp(-na) \) and \( P(D(\hat{\mu}||\mu) \geq a, \hat{\mu} \geq \mu) \leq \exp(-na) \).

Let the transition probabilities from state 0 to state 1 and state 1 to state 0 of a two-state Markov chain \( \{J_t\}_{t\geq 0} \) be \( P_{01} \) and \( P_{10} \), respectively. Let \( P_{00} = 1 - P_{01} \) and \( P_{11} = 1 - P_{10} \). Let the stationary probabilities of states 0 and 1 be \( \pi_0 \) and \( \pi_1 \), respectively. Let \( N_{ij}(t) = \sum_{s=1}^{t} \mathbb{1}_{\{J_s = i\}} \), \( N_i(t) = \sum_{s=1}^{t} \mathbb{1}_{\{J_s = i\}} \), \( N_{ij}(t) \) and \( N_i(t) \) denote the number of times the Markov chain transitions from state \( i \) to state \( j \) till time \( t \). Let \( \hat{P}_{ij}(t) = \frac{N_{ij}(t)}{N_i(t)} \). The proposition presented next establishes that the fraction of visits to any state of a Markov chain is never too away from the stationary probability of the state. This describes the contribution of mixing time to the regret upper bound. Clearly, the upper bound in Proposition 2 is finite and does not contribute to the asymptotic regret bound. The proposition described next depicts that the estimates of the transition probabilities associated with a Markov chain are never
too far from the true transition probabilities. Proposition 4, along with Borel–Cantelli Lemma establishes that appropriate conditions on the TV distance are satisfied sufficiently often for i.i.d. $(P_{01} + P_{10} = 1)$ and truly Markovian arm reward $(P_{01} + P_{10} \neq 1)$, respectively.

**Proposition 2:** Let $C_1 := \{ \frac{\|\Phi(t)\|}{\|\Phi(t)\|} > \epsilon_1 \}$. Then $\sum_{i=0}^{\infty} \mathbb{P}(C_1) \leq \frac{\sum_{i=0}^{\infty} (t+1)^3}{\epsilon_1^4 \epsilon_1^2 (P_{01} + P_{10})^2}$.

**Proposition 3:** Let $D_{01} := \{ |P_{01} - P_{10}| > \epsilon_1 \}$ and $D_{10} := \{ |P_{10} - P_{01}| > \epsilon_1 \}$. Then $\sum_{i=0}^{\infty} \mathbb{P}(D_{01}) \leq \frac{\sum_{i=0}^{\infty} (t+1)^3}{\epsilon_1^4 \epsilon_1^2 (P_{01} + P_{10})^2}$ and $\sum_{i=0}^{\infty} \mathbb{P}(D_{10}) \leq \frac{\sum_{i=0}^{\infty} (t+1)^3}{\epsilon_1^4 \epsilon_1^2 (P_{01} + P_{10})^2}$.

**Proposition 4:** Let $B_{\epsilon} := \{ |P_{01} - P_{10}| < \frac{\epsilon_1^2}{\gamma_1} \}$. Then $P_{01} + P_{10} = 1$, $\sum_{i=0}^{\infty} \mathbb{P}(B_{\epsilon}) \leq \frac{\sum_{i=0}^{\infty} (t+1)^3}{\epsilon_1^4 \epsilon_1^2 (P_{01} + P_{10})^2}$.

Proofs of Propositions 2–4 are provided in Appendix B. The proposition presented next is used to prove that the confidence bounds on transition probabilities associated with the optimal arm are never too far from the respective true transition probabilities. The proposition thereafter is used to establish that the index associated with a suboptimal arm is not often much greater than the index of the optimal arm. Let $p_{12}$, $p_{21}$, $q_1$, $q_2 \in [0, 1]$. Let $Z_i(s)$ be a nonnegative random variable with $Z_i(s) \in [0, 1]$, $s = 1, 2, \ldots, n$. Let $W_1, W_2, \ldots, W_n$ be i.i.d. Bernoulli random variables with mean $\mu \in [0, 1]$, and $Z(s)$ is measurable with respect to $\{W_1, W_2, \ldots, W_n\}$. Of Propositions 5 and 6 are given in [31], Appendix III.

**Proposition 5:**

1. Let $X_1, X_2, \ldots, X_n$ be i.i.d. Bernoulli random variables with mean $p \in [0, 1]$, and $q_\epsilon > 0$. Let $p_{\epsilon} := \frac{1}{n} \sum_{i=1}^{n} X_i$, where $s$ is the number of samples till time $t$. Let $p_{\epsilon} := \max \{ \epsilon \in [0, 1] : D(\hat{p}_{\epsilon}||\epsilon) \leq \frac{\log(t)}{\epsilon} \}$. Then, $\sum_{i=0}^{\infty} \mathbb{P}(p_{\epsilon} < p - \epsilon_2) \leq \frac{4}{\epsilon_2^4}$.

2. Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. Bernoulli random variables with mean $q \in [0, 1]$, $q_\epsilon > 0$. Let $\hat{q}_i := \frac{1}{n} \sum_{i=1}^{n} Y_i$ and $\hat{q}_\epsilon := \min \{ \epsilon \in [0, 1] : D(\hat{q}_i||\epsilon) \leq \frac{\log(t)}{\epsilon} \}$. Then, $\sum_{i=0}^{\infty} \mathbb{P}(\hat{q}_d > q + \epsilon_2) \leq \frac{4}{\epsilon_2^4}$.

**Proposition 6:**

1. $\kappa_1 := \sum_{i=0}^{\infty} \mathbb{P}(D(p_{\epsilon}, q_{\epsilon}) \leq \frac{\log(t)}{\epsilon} + Z_i(s) \leq \frac{\log(t)}{\epsilon})$. Then, $\kappa_1 \leq \frac{1}{\epsilon_1^4 \epsilon_1^2 \epsilon_1^2 (P_{01} + P_{10})^2}$.

2. $\kappa_2 := \sum_{i=0}^{\infty} \mathbb{P}(D(p_{\epsilon}, q_{\epsilon}) \leq \frac{\log(t)}{\epsilon} + Z_i(s) \leq \frac{\log(t)}{\epsilon})$. Then, $\kappa_2 \leq \frac{1}{\epsilon_1^4 \epsilon_1^2 \epsilon_1^2 (P_{01} + P_{10})^2}$.

3. $\kappa_3 := \sum_{i=0}^{\infty} \mathbb{P}(D(\hat{p}_d, \hat{q}_d) \leq \frac{\log(t)}{\epsilon} + Z_2(s) \leq \frac{\log(t)}{\epsilon})$. Then, $\kappa_3 \leq \frac{1}{\epsilon_1^4 \epsilon_1^2 \epsilon_1^2 (P_{01} + P_{10})^2}$.

1) **Truly Markovian Arms:** An arm can either be in STP_PHASE or in SM_PHASE at time $t$ depending on whether the condition on the TV distance is satisfied or not. When all arms are truly Markovian, we establish that the expected number of times the arms are in SM_PHASE, is finite. In other words, the appropriate conditions on the TV distances are satisfied sufficiently often. Let $E_{1, i, t} := \{ \hat{p}_{10}(t - 1) - 1 + \hat{p}_{10}(t - 1) > \frac{1}{\epsilon_1^4 \epsilon_1^2 \epsilon_1^2 (P_{01} + P_{10})^2} \}$. For $t_1, i := \{ \hat{p}_{10}(t - 1) - 1 + \hat{p}_{10}(t - 1) > \frac{1}{\epsilon_1^4 \epsilon_1^2 \epsilon_1^2 (P_{01} + P_{10})^2} \}$, we obtain (using Proposition 4)

$$\kappa_{i-1, i, t} := \frac{1}{\epsilon_1^4 \epsilon_1^2 \epsilon_1^2 (P_{01} + P_{10})^2} + \frac{1}{\epsilon_1^4 \epsilon_1^2 \epsilon_1^2 (P_{01} + P_{10})^2} \leq \frac{1}{\epsilon_1^4 \epsilon_1^2 \epsilon_1^2 (P_{01} + P_{10})^2} \leq \frac{1}{\epsilon_1^4 \epsilon_1^2 \epsilon_1^2 (P_{01} + P_{10})^2}.$$
where the first equality follows from the definition of \( \tilde{p}_{n1}^0(t) \). The second inequality follows from the fact that \( D(x|\hat{p}_{n1}^0(t)) \) is decreasing in \( x \in [\hat{p}_{n1}^0 - \epsilon, \hat{p}_{n1}^0 + \epsilon] \) and \( \hat{p}_{n1}^0 + \epsilon < \frac{(\hat{p}_{n1}^0 - \epsilon \hat{p}_{n1}^0 - \epsilon)}{\hat{p}_{n1}^0 + \epsilon} \). The last inequality follows from Proposition 6. Similarly

\[
\sum_{t=1}^{\infty} \left( A_i = i, \tilde{p}_{n1}^0(t) > \frac{p_{n1}^0(t) \hat{p}_{n1}^0(t)}{p_{n1}^0 + \epsilon}, E_{1,1}^t \right) \leq \left( \frac{\log f(n)}{D(p_{n1}^0 - \epsilon \ln(p_{n1}^0 + \epsilon))} - \tau_{1,i} \right) + \sum_{t=1}^{\infty} \mathbb{P}(F_{1,1}^t)
\]

where the equality follows from (9) and (10). The inequality follows since each of \( P_{n1}^0(t), \hat{P}_{n1}^0(t), \hat{p}(0, t) \) (and \( \hat{p}(t) \)) can take \((t + 1)\) possible values and hence, \( |H| \leq (t + 1)^2 \). Therefore the proof reduces to

\[
\frac{\min \left\{ \frac{y}{x+y}, D(x|P_{n1}^0) \right\}}{\max \left\{ \frac{y}{x+y}, D(y|P_{n1}^0) \right\}} \\
\text{subject to } \frac{y}{x+y} - \pi_0 > \epsilon_1, \frac{x}{x+y} - \pi_1 > \epsilon_1, \frac{x}{y} - \pi_0 > \epsilon_2, \frac{x}{y} - \pi_1 > \epsilon_2.
\]

Let \((x', y')\) be a solution to the above problem. Using Pinsker’s inequality, \( D(x'|P_{n1}^0) \geq 2(x' - P_{n1}^0)^2 \) and \( D(y'|P_{n1}^0) \geq 2(y' - P_{n1}^0)^2 \). We know that \( (x'+y') - P_{n1}^0 + P_{n1}^0 - P_{n1}^0) > \epsilon^2 \) and \( (x'+y') - P_{n1}^0 + P_{n1}^0 - P_{n1}^0) > \epsilon^2 \). Hence, \( (x'+y') - P_{n1}^0 + P_{n1}^0 - P_{n1}^0 > \epsilon_1 \). The result of the proof follows immediately.

\[
\text{Proof of Proposition 3: } \text{We take } C_i := \left\{ \frac{N(t)}{t} - \pi_0 > \epsilon_1 \right\}
\]

\[
\sum_{t=1}^{\infty} \mathbb{P}(D_{0,t}) = \sum_{t=1}^{\infty} \mathbb{P}(D_{0,t}, C_i) + \mathbb{P}(D_{0,t}, C_i)
\]

\[
\leq \sum_{t=1}^{\infty} \mathbb{P}\left( \hat{P}_{01}^0(t) - P_{01}^0 \geq \epsilon_1, \frac{N(t)}{t} - \pi_0 \leq \epsilon_1 \right) + \mathbb{P}(C_i)
\]

\[
\leq \sum_{t=1}^{\infty} \mathbb{P}\left( \hat{P}_{01}^0(t) - P_{01}^0 > \epsilon_1 \right) + \mathbb{P}(C_i)
\]

\[
\leq \sum_{t=1}^{\infty} \mathbb{P}\left( \hat{P}_{01}^0(t) - P_{01}^0 > \epsilon_1 \right) + \mathbb{P}(C_i)
\]

where the third and last inequalities follow from Propositions 1 and 2, respectively. Proof for \( \sum_{t=1}^{\infty} \mathbb{P}(D_{1,t}) \) follows in a similar manner.

\[
\text{Proof of Proposition 4: Case I-} P_{01}^0 = 1: \text{Using triangle inequality, } TV(\hat{P}_{01}^0(t)|1 - P_{01}^0(t)) \leq |\hat{P}_{01}^0(t) - P_{01}^0| + |\hat{P}_{01}^0(t) - P_{01}^0| + |P_{01}^0 - P_{01}^0 - 1|.
\]

Therefore using (5), (6), and (8), \( \mathbb{E}[T(t)_n] \leq \tau_{1,i} + \frac{2n}{t-1} + \frac{2n}{t-1} + 10\sum_{t=1}^{\infty} (t - 1)^3 \exp(-2(t-1) \epsilon_1^2 (p_{n1}^0 + p_{n1}^0)^2) + 10\sum_{t=1}^{\infty} (t - 1)^3 \exp(-2(t-1) \epsilon_1^2 (p_{n1}^0 + p_{n1}^0)^2).
\]

We choose \( \epsilon_1 = \epsilon_2 = \epsilon_3 = \log(\frac{4}{t^{1/2}}) \) to complete the proof. Proof for the other cases follow in a similar manner and is given in [31].

\[
\sum_{t=1}^{\infty} \mathbb{P}\left( TV(\hat{P}_{01}^0(t)|1 - P_{01}^0(t)) \geq \frac{1}{t^{1/2}} \right) \leq \sum_{t=1}^{\infty} \mathbb{P}\left( |\hat{P}_{01}^0(t) - P_{01}^0| + |\hat{P}_{01}^0(t) - P_{01}^0| + |P_{01}^0 - P_{01}^0 - 1| \geq \frac{1}{t^{1/2}} C_t \right) + \mathbb{P}(C_i)
\]
We know that $\sigma_i = p_{i1} + p_{i0}$. The upper bound on the regret of UCB-SM [10] for arms modeled as two-state Markov chains $(r(s) = 0)$ is $\Sigma_{i=1}^{n} \frac{1}{\min \sigma_i}$. Let $L = \min \sigma_i$.  

1) Truly Markovian arms: Using Proposition 1

$$\Delta_i = \frac{1}{2} \left( \frac{p_{i1} p_{i0}}{p_{i0}} - \frac{p_{i1} p_{i0}}{p_{i1}} \right) + \frac{1}{D \left( \frac{p_{i1} p_{i0}}{p_{i0}} \right)}$$

$$\leq \frac{1}{2} \left( \frac{p_{i1} p_{i0}}{p_{i0}} - \frac{p_{i1} p_{i0}}{p_{i1}} \right) + \frac{1}{D \left( \frac{p_{i1} p_{i0}}{p_{i0}} \right)}$$

$$\leq \Delta_i \leq \frac{1}{2} \left( \frac{p_{i1} p_{i0}}{p_{i0}} - \frac{p_{i1} p_{i0}}{p_{i1}} \right) + \frac{1}{D \left( \frac{p_{i1} p_{i0}}{p_{i0}} \right)}$$

$$\leq \Delta_i \leq \frac{1}{2} \left( \frac{p_{i1} p_{i0}}{p_{i0}} - \frac{p_{i1} p_{i0}}{p_{i1}} \right) + \frac{1}{D \left( \frac{p_{i1} p_{i0}}{p_{i0}} \right)}$$

since $\Delta_i = \frac{1}{2} \left( \frac{p_{i1} p_{i0}}{p_{i0}} - \frac{p_{i1} p_{i0}}{p_{i1}} \right)$. The result holds true if $\min \sigma_i \geq \frac{1}{1440}$.  

2) I.I.D. optimal and truly Markovian suboptimal: Proof follows directly from (11), $\mu_i = p_{i1}$ and $1 - \mu_i = p_{i0}$.  

3) Truly Markovian & i.i.d. suboptimal: Using Proposition 1, $\Sigma_{i=1}^{n} \Delta_i \leq \frac{1}{2} \left( \frac{p_{i1} p_{i0}}{p_{i0}} - \frac{p_{i1} p_{i0}}{p_{i1}} \right) + \frac{1}{D \left( \frac{p_{i1} p_{i0}}{p_{i0}} \right)}$.  

4) I.I.D. arms: Since the upper bound on the regret of TV-KL-UCB matches the lower bound, the result follows immediately.  

REFERENCES

[1] H. Robbins, “Some aspects of the sequential design of experiments,” Bull. Amer. Math. Soc., vol. 58, no. 5, pp. 527–535, 1952.

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

[3] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, “The non-stochastic multiarmed bandit problem,” SIAM J. Comput., vol. 32, no. 1, pp. 48–77, 2002.

[4] W. R. Thompson, “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, 1933.

[5] V. Anantharam, P. Varaiya, and J. Walrand, “Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part I: I.I.D. rewards,” IEEE Trans. Autom. Control, vol. 32, no. 11, pp. 968–976, Nov. 1987.

[6] R. Agrawal, “Sample mean based index policies by $o(\log n)$ regret for the multi-armed bandit problem,” Adv. Appl. Probability, vol. 27, no. 4, pp. 1054–1078, 1995.

[7] P. Auer, N. Cesa-Bianchi, and P. Fischer, “Finite-time analysis of the multiarmed bandit problem,” Mach. Learn., vol. 47, no. 2/3, pp. 235–256, 2002.

[8] A. Garivier and O. Cappe, “The KL-UCB algorithm for bounded stochastic bandits and beyond,” in Proc. 24th Annu. Conf. Learn. Theory, 2011, pp. 359–376.

[9] O. Cappe, A. Garivier, O.-A. Maillard, R. Munos, and G. Stoltz, “Kullback–Leibler upper confidence bounds for optimal sequential allocation,” Ann. Statist., vol. 41, no. 3, pp. 1516–1541, 2013.

[10] T. K. D. M. Andrieu, P. V. Mathur, and J. Walrand, “Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part II: Markovian rewards,” IEEE Trans. Autom. Control, vol. 32, no. 11, pp. 977–982, Nov. 1987.

[11] P. Auer, N. Cesa-Bianchi, and G. Lugosi, Prediction, Learning, and Games, Cambridge, U.K.: Cambridge Univ. Press, 2002.

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

[13] T. Graepel, J. Q. Candela, T. Borchert, and R. Herbrich, “Web-scale Bayesian click-through rate prediction for sponsored search advertising in microsoft’s bing search engine,” in Proc. 27th Int. Conf. Mach. Learn., 2010, pp. 13–20.

[14] R. Herbrich, T. Minka, and T. Graepel, “TruetimeTM: A Bayesian skill rating system,” in Proc. Int. Conf. Adv. Neural Inf. Process. Syst., 2006, pp. 593–596.

[15] T. Lattimore and C. Szepesvári, Bandit Algorithms. Cambridge, U.K.: Cambridge Univ. Press, 2018.

[16] V. Avadhanula, R. Colini Baldeschi, S. Leonardi, K. A. Sankaranarayanan, and O. Schrijvers, “Stochastic bandits for multi-platform budget optimization in online advertising,” in Proc. Int. Conf. World Wide Web, 2021, pp. 2805–2817.

[17] J. D. Gittins, “A dynamic allocation index for the sequential design of experiments,” in Progress in Statistics, Amsterdam, The Netherlands: North Holland, 1974, pp. 241–256.

[18] P. Whittle, “Restless bandits: Activity allocation in a changing world,” J. Appl. Probability, vol. 25, pp. 287–298, 1988.

[19] A. Lazaric et al., “Online stochastic optimization under correlated bandit feedback,” in Proc. Int. Conf. Mach. Learn., 2014, pp. 1557–1565.

[20] D. P. Foster and R. Vohra, “Regret in the on-line decision problem,” Games Econ. Behav., vol. 29, no. 1/2, pp. 7–35, 1999.

[21] N. Cesa-Bianchi and G. Lugosi, Prediction, Learning, and Games, Cambridge, U.K.: Cambridge Univ. Press, 2006.

[22] O. Besbes, Y. Gu, and A. Zeevi, “Stochastic multi-armed-bandit problem with non-stationary rewards,” in Proc. Int. Conf. Adv. Neural Inf. Process. Syst., 2014, vol. 27, pp. 199–207.

[23] T. Jakusch, R. Ortner, and N. Auer, “Near-optimal regret bounds for reinforcement learning,” J. Mach. Learn. Res., vol. 11, no. 4, pp. 1563–1600, 2010.

[24] M. G. Azar, I. Osband, and R. Munos, “Minimax regret bounds for reinforcement learning,” in Proc. Int. Conf. Mach. Learn., 2017, pp. 263–272.

[25] R. Fruit, M. Pirrotta, and A. Lazaric, “Near optimal exploration-exploitation in non-communicating Markov decision processes,” in Proc. Adv. Neural Inf. Process. Syst., vol. 31, 2018.

[26] A. Roy, S. Shakkottai, and R. Srikant, “Adaptive KL-UCB based bandit algorithms for Markovian and I.I.D. settings,” 2020, arXiv:2009.06606.

[27] O. Cappe, A. Garivier, R. Munos, and G. Stoltz, “Kullback–Leibler upper confidence bounds for optimal sequential allocation,” Ann. Statist., vol. 41, no. 3, pp. 1516–1541, 2013.

[28] J. Lin, “Divergence measures based on the Shannon entropy,” IEEE Trans. Inf. Theory, vol. 37, no. 1, pp. 145–151, Jan. 1991.

[29] R. Agrawal, “Finite-sample concentration of the multinomial in relative entropy,” IEEE Trans. Inf. Theory, vol. 66, no. 10, pp. 6297–6302, Oct. 2020.