On the Combinatorial Multi-Armed Bandit Problem with Markovian Rewards

Yi Gai*, Bhaskar Krishnamachari* and Mingyan Liu‡
*Ming Hsieh Department of Electrical Engineering
University of Southern California, Los Angeles, CA 90089, USA
‡Department of Electrical Engineering and Computer Science
University of Michigan, Ann Arbor, MI 48109, USA
Email: {ygai,bkrishna}@usc.edu; mingyan@eecs.umich.edu

***Abstract*— We consider a combinatorial generalization of the classical multi-armed bandit problem that is defined as follows. There is a given bipartite graph of M users and $N \geq M$ resources. For each user-resource pair (i, j), there is an associated state that evolves as an aperiodic irreducible finite-state Markov chain with unknown parameters, with transitions occurring each time the particular user i is allocated resource j. The user i receives a reward that depends on the corresponding state each time it is allocated the resource j. The system objective is to learn the best matching of users to resources so that the long-term sum of the rewards received by all users is maximized. This corresponds to minimizing regret, defined here as the gap between the expected total reward that can be obtained by the best-possible static matching and the expected total reward that can be achieved by a given algorithm. We present a polynomial-storage and polynomial-complexity-per-step matching-learning algorithm for this problem. We show that this algorithm can achieve a regret that is uniformly logarithmic in time and polynomial in the number of users and resources, under certain conditions on the component Markov chains. This formulation is broadly applicable to scheduling and switching problems in networks and significantly extends prior results in the area.**

I. Introduction

Multi-armed bandit problems provide a fundamental approach to learning under stochastic rewards, and find rich applications in a wide range of networking contexts, from Internet advertising [1] to medium access in cognitive radio networks [2]–[4]. In the simplest, classic non-Bayesian version of the problem, studied by Lai and Robbins [5], there are K independent arms, each generating stochastic rewards that are i.i.d. over time. The player is unaware of the parameters for each arm, and must use some policy to play the arms in such a way as to maximize the cumulative expected reward over the long term. The policy's performance is measured in terms of its "regret", defined as the gap between the the expected reward that could be obtained by an omniscient user that knows the parameters for the stochastic rewards generated by each arm and the expected cumulative reward of that policy. It is of interest to characterize the growth of regret with respect to time as well as with respect to the number of arms/players. Intuitively, if the regret grows sublinearly over time, the time-averaged regret tends to zero.

There is inherently a tradeoff between exploration and exploitation in the learning process in a multi-armed bandit problem: on the one hand all arms need to be sampled periodically by the policy used, to ensure that the "true" best arm is found; on the other hand, the policy should play the arm that is considered to be the best often enough to accumulate rewards at a good pace.

In this paper, we formulate a novel combinatorial generalization of the multi-armed bandit problem that allows for Markovian rewards and propose an efficient policy for it. In particular, there is a given bipartite graph of M users and $N \geq M$ resources. For each user-resource pair (i, j), there is an associated state that evolves as an aperiodic irreducible finite-state Markov chain with unknown parameters, with transitions occurring each time the particular user i is allocated resource j. The user i receives a reward that depends on the corresponding state each time it is allocated the resource j. A key difference from most prior works is that each user can potentially see a different reward process for the same resource. If we therefore view each possible matching of users to resources as an arm, then we have a super-exponential number of arms with dependent rewards. Thus, this new formulation is significantly more challenging than the traditional multi-armed bandit problems.

Because our formulation allows for user-resource matching, it is broadly applicable to many networking settings such as switching in routers (where inputs need to be matched to outputs), and frequency scheduling in wireless networks (where nodes need to be allocated to channels). For instance, our work can be applied directly to the channel allocation problem in cognitive radio networks considered in [2] for the case when the rewards for each user-channel pair come from a discrete set.

Our main contribution in this work is the design of a novel policy for this problem that we refer to Matching Learning for Markovian Rewards (MLMR). Since we treat each possible matching of users to resources as an arm, the number of arms in our formulation grows super-exponentially. However, MLMR uses only polynomial storage, and requires only polynomial computation at each step. We analyze the regret for this policy with respect to the best possible static matching, and show that it is uniformly logarithmic over time

and polynomial in the number of users and resources.

The rest of the paper is organized as follows. In section II we present our work in the context of prior results on multi-armed bandits. In section III we present the problem formulation. In section IV we present a polynomial-storage polynomial-time-per-step learning policy, which we refer to as MLMR. We analyze the regret for this policy in section V and show that it yields a bound on the regret that is uniformly logarithmic over time and polynomial in the number of users and resources under certain conditions on the Markov chains describing the state evolution for the arms. We present some examples and simulations in section VI, and conclude with some comments and ideas for future work in section VII.

II. Prior Work

The problem we consider in this paper is different from prior work for two key reasons. We treat rewards that are dependent across a super-exponential number of arms whose states evolve in a non-i.i.d. Markovian fashion over time. We summarize below prior work, which has treated a) independent and temporally i.i.d. rewards, or b) independent and Markovian state-based rewards, or c) non-independent arms with temporally i.i.d.

A. Independent arms with temporally i.i.d. rewards

The work by Lai and Robbins [5] assumes K independent arms, each generating rewards that are i.i.d. over time from a given family of distributions with an unknown real-valued parameter. For this problem, they present a policy that provides an expected regret that is $O(K \log n)$, i.e. linear in the number of arms and asymptotically logarithmic in n. Anantharam *et al*. extend this work to the case when M simultaneous plays are allowed [6]. The work by Agrawal [7] presents easier to compute policies based on the sample mean that also has asymptotically logarithmic regret. The paper by Auer *et al*. [8] that considers arms with non-negative rewards that are i.i.d. over time with an arbitrary un-parameterized distribution that has the only restriction that it have a finite support. Further, they provide a simple policy (referred to as UCB1), which achieves logarithmic regret uniformly over time, rather than only asymptotically. Our work utilizes the Chernoff-Hoeffding-bound-based approach to regret analysis pioneered by Auer *et al*..

Some recent work has shown the design of distributed multiuser policies providing asymptotically logarithmic regret, for the context of cognitive radio networks [3], [4].

B. Independent arms with Markovian rewards

There has been relatively less work on multi-armed bandits with Markovian rewards. Anantharam *et al*. [9] wrote one of the earliest papers with such a setting. They proposed a policy to pick m out of the N arms each time slot and prove the lower bound and the upper bound on regret. However, the rewards in their work are assumed to be generated by rested Markov chains with transition probability matrices defined by a single parameter θ with identical state spaces. Also, the result for the upper bound is achieved only asymptotically.

For the case of single users and independent arms, a recent work by Tekin and Liu [10] has extended the results in [9] to the case with no requirement for a single parameter and identical state spaces across arms. They propose to use UCB1 from [8] for the multi-armed bandit problem with Markovian rewards and prove a logarithmic upper bound on the regret under some conditions on the Markov chain. We use elements of the proof from [10] in this work, which is however quite different in its combinatorial matching formulation (which allows for dependent arms).

C. Dependent arms with temporally i.i.d. rewards

The paper by Pandey *et al*. [1] divides arms into clusters of dependent arms, each providing binary rewards, but they do not present any theoretical analysis on the expected regret. In [11], the reward from each arm is modeled as the sum of a linear combination of a set of static random numbers and a zero-mean random variable that is i.i.d. over time and independent across arms. This is quite different from our model of rewards.

Our work in this paper is closest to and builds on the recent work which introduced combinatorial multi-armed bandits [2]. The formulation in [2] has the restriction that the reward process must be i.i.d. over time. A polynomial storage matching learning algorithm is presented in [2] that yields regret that is polynomial in users and resources and uniformly logarithmic in time for the case of i.i.d. rewards. Although i.i.d. rewards are a special case of Markovian state-based rewards, one reason this work is not a strict generalization of [2] is our assumption that the number of possible states, and hence the support of the reward distribution on each arm, is finite (whereas [2] allows for continuous reward distributions with bounded support).

III. Problem Formulation

We consider a bipartite graph with M users and $N \geq M$ resources predefined by some application, e.g. a wireless network with M transmitters and N channels. Time is slotted and is indexed by n. At each decision period (also referred to interchangeably as time slot), each of the M users is assigned a resource with some policy.

For each user-resource pair (i, j), there is an associated state that evolves as an aperiodic irreducible finite-state Markov chain with unknown parameters. When user i is assigned resource j, assuming there are no other conflicting users assigned this resource, i is able to receives a reward that depends on the corresponding state each time it is allocated the resource j. We denote the state space as $S_{i,j} = \{z_1, z_2, \ldots, z_{|S_{i,j}|}\}$. The state of the Markov chain for each user-resource pair (i, j) evolves only when resource j is allocated to user i. We assume the Markov chains for different user-resource pairs are mutually independent. The reward got by user i while allocated resource j on state $z \in S_{i,j}$ is denoted as $\theta_z^{i,j}$, which is also unknown to the users. We denote $\mathbf{P}_{i,j} = \{p_{i,j}(z_a, z_b)\}_{z_a, z_b \in S_{i,j}}$

as the transition probability matrix for the Markov chain (i,j). Denote $\pi_z^{i,j}$ as the steady state distribution for state z. The mean reward got by user i on resource j is denoted as $\mu_{i,j}$. Then we have $\mu_{i,j} = \sum_{z \in S_{i,j}} \theta_z^{i,j} \pi_z^{i,j}$. The set of all mean rewards is denoted as $\boldsymbol{\mu} = \{\mu_{i,j}\}$.

We denote $Y_{i,j}(n)$ as the actual reward obtained by a user i if it is assigned resource j at time n. We assume that $Y_{i,j}(n) = \theta_{i,j}^{z(n)}$, if user i is the only occupant of resource j at time n where $z(n)$ is the state of Markov chain associated with (i,j) at time n. Else, if multiple users are allocated resource j, then we assume that, due to interference, at most one of the conflicting users j' gets reward $Y_{i,j'}(n) = \theta_{i,j'}^{z'(n)}$ where $z'(n)$ is the state of Markov chain associated with (i,j') at time n, while the other users on the resources $j \neq j'$ get zero reward, i.e., $Y_{i,j}(n) = 0$. This interference model covers scenarios in many networking settings, such as the perfect collision model in which none of the conflicting users derive any benefit and CSMA with perfect sensing in which exactly one of the conflicting user derives benefit from the channel.

A deterministic policy $\alpha(n)$ at each time is defined as a map from the observation history $\{\mathbf{O}_t\}_{t=1}^{n-1}$ to a vector of resources $\mathbf{o}(n)$ to be selected at period n, where $\mathbf{O}_t$ is the observation at time t; the i-th element in $\mathbf{o}(n)$, $o_i(n)$, represents the resource allocation for user i. Then the observation history $\{\mathbf{O}_t\}_{t=1}^{n-1}$ in turn can be expressed as $\{o_i(t), Y_{i,o_i(t)}(t)\}_{1 \leq i \leq M, 1 \leq t < n}$.

Due to the fact that allocating more than one user to a resource is always worse than assigning each a different resource in terms of sum-throughput, we will focus on collision-free policies that assign all users distinct resources, which we will refer to as a permutation or matching. There are $P(N,M)$ such permutations.

We formulate our problem as a combinatorial multi-armed bandit, in which each arm corresponds to a matching of the users to resources. We can represent the arm corresponding to a permutation k ($1 \leq k \leq P(N,M)$) as the index set $\mathcal{A}_k = \{(i,j) : (i,j) \text{ is in permutation } k\}$. The stochastic reward for choosing arm k at time n under policy α is then given as

$$Y_{\alpha(n)}(n) = \sum_{(i,j) \in \mathcal{A}_{\alpha(n)}} Y_{i,j}(n) = \sum_{(i,j) \in \mathcal{A}_{\alpha(n)}} \theta_{i,j}^{z_{\alpha(n)}}.$$

Note that different from most prior work on multi-armed bandits, this combinatorial formulation results in dependence across arms that share common components.

A key metric of interest in evaluating a given policy for this problem is *regret*, which is defined as the difference between the expected reward that could be obtained by the best-possible static matching, and that obtained by the given policy. It can be expressed as:

$$\begin{aligned} R^{\alpha}(n) &= n\mu^* - E_{\alpha}[\sum_{t=1}^{n} Y_{\alpha(t)}(t)] \\ &= n\mu^* - E_{\alpha}[\sum_{t=1}^{n} \sum_{(i,j) \in \mathcal{A}_{\alpha(t)}} \theta_{i,j}^{z_{\alpha(t)}}], \end{aligned} \tag{1}$$

where $\mu^* = \max_k \sum_{(i,j) \in \mathcal{A}_k} \mu^{i,j}$, the expected reward of the optimal arm, is the expected sum-weight of the maximum weight matching of users to resources with $\mu_{i,j}$ as the weight.

We are interested in designing policies for this combinatorial multi-armed bandit problem with Markovian rewards that perform well with respect to regret. Intuitively, we would like the regret $R^{\alpha}(n)$ to be as small as possible. If it is sub-linear with respect to time n, the time-averaged regret will tend to zero.

IV. Matching Learning for Markovian Rewards

A straightforward idea for the combinatorial multi-armed bandit problem with Markovian rewards is to treat each matching as an arm, apply UCB1 policy (given by Auer *et al*. [8]) directly, and ignores the dependencies across the different arms. For each arm k, two variables are stored and updated: the time average of all the observation values of arm k and the number of times that arm k has been played up to the current time slot. The UCB1 policy makes decisions based on this information alone.

However, there are several problems for applying UCB1 directly in the above setting. We note that UCB1 requires both the storage and computation time that are linear in the number of arms. Since the number of arms in this formulation grows as $P(N,M)$, it is highly unsatisfactory. Also, the upper-bound of regret given in [10] will not work anymore since the rewards across arms are not independent anymore and the states of an arm may involve even when this arm is not played. No analysis result on the upper-bound of regret can be applied directly in this setting to our best knowledge.

So we are motivated to propose a policy which more efficiently stores observations from correlated arms and exploits the correlations to make better decisions. Our key idea is to use two M by N matrices, $(\hat{\theta}_{i,j})_{M \times N}$ and $(n_{i,j})_{M \times N}$, to store the information for each user-resource pair, rather than for each arm as a whole. $\hat{\theta}_{i,j}$ is the average (sample mean) of all the observed values of resource j by user i up to the current time slot (obtained through potentially different sets of arms over time). $n_{i,j}$ is the number of times that resource j has been assigned to user i up to the current time slot.

At each time slot n, after an arm k is played, we get the observation of $Y_{i,j}(n)$ for all $(i,j) \in \mathcal{A}_k$. Then $(\hat{\theta}_{i,j})_{M \times N}$ and $(n_{i,j})_{M \times N}$ (both initialized to 0 at time 0) are updated as follows:

$$\hat{\theta}_{i,j}(n) = \begin{cases} \frac{\hat{\theta}_{i,j}(n-1) n_{i,j}(n-1) + Y_{i,j}(n)}{n_{i,j}(n-1)+1} & \text{, if } (i,j) \in \mathcal{A}_k \\ \hat{\theta}_{i,j}(n-1) & \text{, else} \end{cases} \tag{2}$$

$$n_{i,j}(n) = \begin{cases} n_{i,j}(n-1) + 1 & \text{, if } (i,j) \in \mathcal{A}_k \\ n_{i,j}(n-1) & \text{, else} \end{cases} \tag{3}$$

Note that while we indicate the time index in the above updates for notational clarity, it is not necessary to store the matrices from previous time steps while running the algorithm.

Our proposed policy, which we refer to as Matching Learning for Markovian Rewards, is shown in Algorithm 1.

Algorithm 1 Matching Learning for Markovian Rewards (MLMR)

// INITIALIZATION
for $p = 1$ to M **do**
  for $q = 1$ to N **do**
    $n = (M-1)p + q$;
    Play any permutation k such that $(p, q) \in \mathcal{A}_k$;
    Update $(\hat{\theta}_{i,j})_{M\times N}$, $(n_{i,j})_{M\times N}$ accordingly.
  end for
end for
// MAIN LOOP
while 1 **do**
  $n = n + 1$;
  Solve the Maximum Weight Matching problem (e.g., using the Hungarian algorithm [12]) on the bipartite graph of users and resources with edge weights $\left(\hat{\theta}_{i,j} + \sqrt{\frac{L \ln n}{n_{i,j}}}\right)_{M\times N}$ to play arm k that maximizes

$$\sum_{(i,j)\in\mathcal{A}_k} \left(\hat{\theta}_{i,j} + \sqrt{\frac{L \ln n}{n_{i,j}}}\right) \tag{4}$$

  where L is a positive constant.
  Update $(\hat{\theta}_{i,j})_{M\times N}$, $(n_{i,j})_{M\times N}$ accordingly.
end while

V. ANALYSIS OF REGRET

We summarizes some notation we use in the description and analysis of our MLMR policy in Table I.

The regret of a policy for a multi-armed bandit problem is traditionally upper-bounded by analyzing the expected number of times that each non-optimal arm is played and then taking the summation over of this expectation times the reward difference between an the optimal arm and a non-optimal arm all non-optimal arms. Although we could use this approach to analyze the MLMR policy, we notice that the upper-bound for regret consequently obtained is quite loose, which is linear in the number of arms, $P(N, M)$. Instead, we present here a novel analysis for a tighter analysis of the MLMR policy. Our analysis shows an upper bound of the regret that is polynomial in M and N, and uniformly logarithmic over time.

Following lemmas are needed for our main results in Theorem 1:

Lemma 1: (Lemma 2.1 from [9]) $\{X_n, n = 1, 2, \ldots\}$ is an irreducible aperiodic Markov chain with state space S, transition matrix P, a stationary distribution $\pi_z, \forall z \in S$, and an initial distribution $\mathbf{q}$. Denote F_t as the σ-algebra generated by $X_1, X_2, \ldots, X_t$. Let G be a σ-algebra independent of $F = \vee_{t\geq 1} F_t$. Let τ be a stopping time with respect to the increasing family of σ-algebra $G \vee F_t, t \geq 1$. Define $N(z, \tau)$ such that $N(z, \tau) = \sum_{t=1}^{\tau} I(X_t = z)$. Then,

$$|E[N(z, \tau) - \pi_z E[\tau]]| \leq C_P, \tag{5}$$

for all $\mathbf{q}$ and all τ such that $E[\tau] < \infty$. C_P is a constant that

N :	number of resources.
M :	number of users, $M \leq N$.
k :	index of a parameter used for an arm, $1 \leq k \leq P(N, M)$.
i, j :	index of a parameter used for user i, resource j.
$\mathcal{A}_k$:	$\{(i, j) : (i, j)$ is in permutation $k\}$
$\mathcal{K}_{i,j}$:	$\{\mathcal{A}_k : (i, j) \in \mathcal{A}_k\}$
$*$:	index indicating that a parameter is for the optimal arm.
$n_{i,j}$:	number of times that resource j has been matched with user i up to the current time slot.
$\hat{\theta}_{i,j}$:	average (sample mean) of all observed values of resource j by user i up to current time slot.
n_i^k:	$n_{i,j}$ such that $(i, j) \in \mathcal{A}_k$ at current time slot.
$S_{i,j}$:	state space of the Markov chain for user-resource pair (i, j).
$\mathbf{P}_{i,j}$:	transition matrix of the Markov chain associated with user-resource pair (i, j).
$\pi_z^{i,j}$:	steady state distribution for state z of the Markov chain associated with (i, j).
$\theta_z^{i,j}$:	reward got by user i while access resource j on state $z \in S_{i,j}$
$\mu_{i,j}$:	$\sum_{z\in S^i} \theta_z^{i,j} \pi_z^{i,j}$, the mean reward for user i using resource j
μ^k:	$\sum_{(i,j)\in\mathcal{A}_k} \mu^{i,j}$
μ^*:	$\max_k \sum_{(i,j)\in\mathcal{A}_k} \mu^{i,j}$
Δ_k:	$\mu^* - \mu_k$.
$\Delta_{\min}$:	$\min_k \Delta_k$.
$\Delta_{\max}$:	$\max_k \Delta_k$.
$\pi_{\min}$:	$\min_{1\leq i\leq M, 1\leq j\leq N, z\in S_{i,j}} \pi_z^{i,j}$.
$s_{\max}$:	$\max_{1\leq i\leq M, 1\leq j\leq N} \lvert S_{i,j}\rvert$.
$s_{\min}$:	$\min_{1\leq i\leq M, 1\leq j\leq N} \lvert S_{i,j}\rvert$.
$\theta_{\max}$:	$\max_{1\leq i\leq M, 1\leq j\leq N, z\in S_{i,j}} \theta_z^{i,j}$.
$\theta_{\min}$:	$\min_{1\leq i\leq M, 1\leq j\leq N, z\in S_{i,j}} \theta_z^{i,j}$.
$\epsilon_{i,j}$:	eigenvalue gap, defined as $1 - \lambda_2$, where λ_2 is the second largest eigenvalue gap of $\mathbf{P}_{i,j}$.
$\epsilon_{\max}$:	$\max_{1\leq i\leq M, 1\leq j\leq N} \epsilon_{i,j}$.
$\epsilon_{\min}$:	$\min_{1\leq i\leq M, 1\leq j\leq N} \epsilon_{i,j}$.
$T_k(n)$:	number of times arm k has been played by MLMR in the first n time slots.
$\hat{\overline{\theta}}_k(n)$:	$\sum_{(i,j)\in\mathcal{A}_k} \hat{\theta}_{i,j}(n)$. It is the summation of all the average observation values in arm k at time n.
$\hat{\theta}^k_{i,n_i^k}$:	$\hat{\theta}_{i,j}(n)$ such that $(i, j) \in \mathcal{A}_k$ and $n_{i,j}(n) = n_i^k$.
$\hat{\overline{\theta}}_{k,n_1^k,\ldots,n_M^k}$:	$\sum_{i=1}^{M} \hat{\theta}_{k,n_i^k}$.

TABLE I
NOTATION

depends on P.

Lemma 2: (Corollary 1 from [10]) Let $\pi_{\min}$ be the minimum value among the stationary distribution, which is defined as $\pi_{\min} = \min_{z \in S} \pi_z$. Then $C_P \leq 1/\pi_{\min}$.

Lemma 3: For user-resource matching, if the state of reward associated with each user-resource pair (i,j) is given by a Markov chain, denoted $\{X_1^{i,j}, X_2^{i,j}, \ldots\}$, satisfying the properties of Lemma 1, then the regret under policy α is bounded by:

$$R^{\alpha}(n) \leq \sum_{k=1}^{P(N,M)} (\mu^* - \mu^k) E_{\alpha}[T_k^{\alpha}(n)] + C_{\mathbf{S},\mathbf{P},\Theta}, \quad (6)$$

where $C_{\mathbf{S},\mathbf{P},\Theta}$ is a constant that depends on all the state spaces $\{S_{i,j}\}_{1\leq i\leq M, 1\leq i\leq N}$, transition probability matrices $\{\mathbf{P}_{i,j}\}_{1\leq i\leq M, 1\leq i\leq N}$ and the rewards set $\{\theta_{i,j}^z, z \in S_{i,j}\}_{1\leq i\leq M, 1\leq i\leq N}$.

Proof:

$\forall 1 \leq i \leq M$, $1 \leq j \leq N$, define $G_{i,j} = \vee_{k\neq i, l\neq j} F_{k,l}$ where $F_{k,l} = \vee_{t\geq 1} F_t^{i,j}$, which applies to the Markove chain $\{X_1^{i,j}, X_2^{i,j}, \ldots\}$. We note that the Markov chains of different user-resource pairs are mutually independent, so $\forall i, j$, $G^{i,j}$ is independent of $F_{i,j}$. $F_{i,j}$ satisfies the conditions in Lemma 1. Note that $T_{i,j}^{\alpha}(n)$ is a stopping time with respect to $\{G_{i,j} \vee F_n^{i,j}, n > 1\}$.

Since the state of a Markove chain evolves only when it is observed, $X_1^{i,j}, \ldots, X_{T_{i,j}^{\alpha}(n)}^{i,j}$ represents the successive states of the Markov chain up to n when assigning resource j to user i.Then the total reward obtained under policy α up to time n is given by:

$$\sum_{t=1}^{n} Y_{\alpha(t)}(t) = \sum_{j=1}^{N}\sum_{i=1}^{M}\sum_{l=1}^{T_{i,j}^{\alpha}(n)}\sum_{z\in S_{i,j}} \theta_z^{i,j} I(X_l^{i,j} = z). \quad (7)$$

Note that $\forall i = 1, \ldots, M, T_k^{\alpha}(n) = T_k^{\alpha(n),i}$ where $T_k^{\alpha(n),i}$ is the number of times up to n that the i-th component has been observed while playing arm k, and there exist one resource index j such that $(i,j) \in \mathcal{A}_k$. So, we have:

$$\begin{aligned}
&\sum_{k=1}^{P(N,M)} \mu^k E_{\alpha}[T_k^{\alpha}(n)] \\
&= \sum_{k=1}^{P(N,M)}\sum_{i=1}^{M} \mu_i^k E_{\alpha}[T_k^{\alpha}(n)] \\
&= \sum_{k=1}^{P(N,M)}\sum_{i=1}^{M} \mu_i^k E_{\alpha}[T_k^{\alpha,i}(n)] \\
&= \sum_{j=1}^{N}\sum_{i=1}^{M} \mu_{i,j} \sum_{\mathcal{A}_k \in \mathcal{K}_{i,j}} E_{\alpha}[T_k^{\alpha,i}(n)] \\
&= \sum_{j=1}^{N}\sum_{i=1}^{M} \mu_{i,j} E_{\alpha}[T_{i,j}^{\alpha}(n)] \\
&= \sum_{j=1}^{N}\sum_{i=1}^{M}\sum_{z\in S_{i,j}} \theta_z^{i,j}\pi_z^{i,j} E_{\alpha}[T_{i,j}^{\alpha}(n)]
\end{aligned}$$

Hence,

$$\begin{aligned}
&|R^{\alpha}(n) - \sum_{k=1}^{P(N,M)} (\mu^* - \mu^k) E_{\alpha}[T_k^{\alpha}(n)]| \\
&= \left| R^{\alpha}(n) - (n\mu^* - \sum_{k=1}^{P(N,M)} \mu^k E_{\alpha}[T_k^{\alpha}(n)]) \right| \\
&= \Bigg| (n\mu^* - E_{\alpha}[\sum_{t=1}^{n} Y_{\alpha(t)}(t)]) \\
&\quad -(n\mu^* - \sum_{k=1}^{P(N,M)} \mu^k E_{\alpha}[T_k^{\alpha}(n)]) \Bigg| \\
&= \left| E_{\alpha}[\sum_{t=1}^{n} Y_{\alpha(t)}(t)] - \sum_{k=1}^{P(N,M)} \mu^k E_{\alpha}[T_k^{\alpha}(n)] \right| \\
&= \Bigg| E_{\alpha}[\sum_{j=1}^{N}\sum_{i=1}^{M}\sum_{l=1}^{T_{i,j}^{\alpha}(n)}\sum_{z\in S_{i,j}} \theta_z^{i,j} I(X_l^{i,j} = z)] \\
&\quad - \sum_{j=1}^{N}\sum_{i=1}^{M}\sum_{z\in S_{i,j}} \theta_z^{i,j}\pi_z^{i,j} E_{\alpha}[T_{i,j}^{\alpha}(n)] \Bigg| \\
&\leq \sum_{j=1}^{N}\sum_{i=1}^{M}\sum_{z\in S_{i,j}} |E_{\alpha}[\sum_{l=1}^{T_{i,j}^{\alpha}(n)} \theta_z^{i,j} I(X_l^{i,j} = z)]. \\
&\quad - \theta_z^{i,j}\pi_z^{i,j} E_{\alpha}[T_{i,j}^{\alpha}(n)]| \\
&= \sum_{j=1}^{N}\sum_{i=1}^{M}\sum_{z\in S_{i,j}} \theta_z^{i,j} |E_{\alpha}[\sum_{l=1}^{T_{i,j}^{\alpha}(n)} I(X_l^{i,j} = z)] \\
&\quad - \pi_z^{i,j} E_{\alpha}[T_{i,j}^{\alpha}(n)]| \\
&= \sum_{j=1}^{N}\sum_{i=1}^{M}\sum_{z\in S_{i,j}} \theta_z^{i,j} \left| E_{\alpha}[N(z, T_{i,j}^{\alpha}(n))] - \pi_z^{i,j} E_{\pi}[T_{i,j}^{\alpha}(n)] \right| \\
&\leq \sum_{j=1}^{N}\sum_{i=1}^{M}\sum_{z\in S_{i,j}} \theta_z^{i,j} C_{P_{i,j}} = C_{\mathbf{S},\mathbf{P},\Theta} \quad (8)
\end{aligned}$$

The inequality in (8) follows from Lemma 1. ■

Lemma 4: (Theorem 2.1 from [13]) Let $\{X_n, n = 1, 2, \ldots\}$ be an irreducible aperiodic Markov chain with finite state space S, transition matrix $\mathbf{P}$, a stationary distribution π_z, $\forall z \in S$, and an an initial distribution $\mathbf{q}$. Let $N_{\mathbf{q}} = ||(\frac{q_z}{\pi_z}), z \in S||_2$. The eigenvalue gap ϵ is defined as $\epsilon = 1 - \lambda_2$, where λ_2 is the second largest eigenvalue of the matrix $\mathbf{P}$. $\forall A \subset S$, define $t_A(n)$ as the total number of times that all states in the set A are visited up to time n. Then $\forall \gamma \geq 0$,

$$P(t_A(n) - n\pi_A \geq \gamma) \leq (1 + \frac{\gamma\epsilon}{10n} N_{\mathbf{q}} e^{-\gamma^2\epsilon/20n}), \quad (9)$$

where $\pi_A = \sum_{z\in A} \pi_z$.

Our main results on the regret of MLMR policy are shown in Theorem 1. We show that with using a constant L which is bigger than a value determined by the minimum eigenvalue gap of the transition matrix, maximum value of the number of

states, and maximum value of the rewards, our MLMR policy is guaranteed to achieve a regret that is uniformly logarithmic in time, and polynomial in the number of users and resources.

***Theorem** 1:* When using any constant $L \geq \frac{(50+40M)\theta_{\max}^2 s_{\max}^2}{\epsilon_{min}}$, the expected regret under the MLMR policy specified in Algorithm 1 is at most

$$\left[\frac{4M^3NL\ln n}{(\Delta_{\min})^2} + MN + M^2N\frac{s_{\max}}{\pi_{\min}}\left(1+\frac{\epsilon_{\max}\sqrt{L}}{10s_{\min}\theta_{\min}}\right)\frac{\pi}{3}\right]\Delta_{\max} + C_{\mathbf{S},\mathbf{P},\Theta}, \tag{10}$$

where $\Delta_{\min}, \Delta_{\max}, \pi_{\min}, s_{\max}, s_{\min}, \theta_{\max}, \theta_{\min}, \epsilon_{\max}, \epsilon_{\min}$ follow the definition in Table I; $C_{\mathbf{S},\mathbf{P},\Theta}$ follows the definition in Lemma 3.

Proof:

Denote $C_{t,n}$ as $\sqrt{\frac{L\ln n}{n}}$. Denote $C_{t,\mathbf{n}_{A_k}} = \sum_{(i,j)\in\mathcal{A}_k}\sqrt{\frac{L\ln n}{n_{i,j}}} = \sum_{i=1}^{M}\sqrt{\frac{L\ln n}{n_i^k}} = \sum_{i=1}^{M} C_{t,n_i^k}$. It is also denoted as $C_{t,(n_1^k,\ldots,n_M^k)}$ sometimes for a clear explanation in this proof.

We introduce $\widetilde{T}_{i,j}(n)$ as a counter after the initialization period. It is updated in the following way:

At each time slot after the initialization period, one of the two cases must happen: (1) an optimal arm is played; (2) a non-optimal arm is played. In the first case, $(\widetilde{T}_{i,j}(n))_{M\times N}$ won't be updated. When an non-optimal arm $k(n)$ is picked at time n, there must be at least one $(i,j)\in\mathcal{A}_k$ such that $n_{i,j}(n) = \min_{(i,j)\in\mathcal{A}_k} n_{i,j}$. If there is only one such arm, $\widetilde{T}_{i,j}(n)$ is increased by 1. If there are multiple such arms, we arbitrarily pick one, say (i',j'), and increment $\widetilde{T}_{i'j'}$ by 1.

Each time when a non-optimal arm is picked, exactly one element in $(\widetilde{T}_{i,j}(n))_{M\times N}$ is incremented by 1. This implies that the total number that we have played the non-optimal arms is equal to the summation of all counters in $(\widetilde{T}_{i,j}(n))_{M\times N}$. Therefore, we have:

$$\sum_{k:\mu_k<\mu*}\mathbb{E}[T_k(n)] = \sum_{i=1}^{M}\sum_{j=1}^{N}\mathbb{E}[\widetilde{T}_{i,j}(n)] \tag{11}$$

Also note for $\widetilde{T}_{i,j}(n)$, the following inequality holds:

$$\widetilde{T}_{i,j}(n) \leq n_{i,j}(n), \forall 1\leq i\leq M, 1\leq j\leq N. \tag{12}$$

Denote by $\widetilde{I}_{i,j}(n)$ the indicator function which is equal to 1 if $\widetilde{T}_{i,j}(n)$ is added by one at time n. Let l be an arbitrary positive integer. Then:

$$\begin{aligned}\widetilde{T}_{i,j}(n) &= \sum_{t=MN+1}^{n}\{\widetilde{I}_{i,j}(t)\}\\ &\leq l + \sum_{t=MN+1}^{n}\{\widetilde{I}_{i,j}(t), \widetilde{T}_{i,j}(t-1)\geq l\}\end{aligned}$$

When $\widetilde{I}_{i,j}(t)=1$, there exists some arm such that a non-optimal arm is picked for which $n_{i,j}$ is the minimum in this arm. We denote this arm as $k(t)$ since at each time that $\widetilde{I}_{i,j}(t)=1$, we may get different arms. Then,

$$\begin{aligned}\widetilde{T}_{i,j}(n) &\leq l + \sum_{t=MN+1}^{n}\{\hat{\overline{\theta^*}}(t-1) + C_{t-1,\mathbf{n}^*(t-1)}\\ &\leq \hat{\overline{\theta}}_{k(t-1)}(t-1) + C_{t-1,\mathbf{n}_{\mathcal{A}_{k(t-1)}}(t-1)}, \widetilde{T}_{i,j}(t-1)\geq l\}\\ &= l + \sum_{t=MN}^{n}\{\hat{\overline{\theta^*}}(t) + C_{t,\mathbf{n}^*(t)}\\ &\leq \hat{\overline{\theta}}_{k(t)}(t) + C_{t,\mathbf{n}_{\mathcal{A}_{k(t)}}(t)}, \widetilde{T}_{i,j}(t)\geq l\}\end{aligned}$$

Based on (12), $l\leq\widetilde{T}_{i,j}(t)$ implies,

$$l\leq\widetilde{T}_{i,j}(t)\leq n_{i,j}(t) = n_i^{k(t)} = \min_j n_j^{k(t)}.$$

So,

$$\forall 1\leq i\leq M, n_i^{k(t)}\geq l.$$

Then we could bound $\widetilde{T}_{i,j}(n)$ as,

$$\begin{aligned}\widetilde{T}_{i,j}(n) &\leq l + \sum_{t=MN}^{n}\{\min_{0<n_1^*,\ldots,n_M^*\leq t}\hat{\overline{\theta^*}}_{n_1^*,\ldots,n_M^*}\\ &\quad + C_{t,(n_1^*,\ldots,n_M^*)} \leq \max_{l\leq n_1^{k(t)},\ldots,n_M^{k(t)}\leq t}\hat{\overline{\theta}}_{k(t),n_1^{k(t)},\ldots,n_M^{k(t)}}\\ &\quad + C_{t,(n_1^{k(t)},\ldots,n_M^{k(t)})}\}\\ &\leq l + \sum_{t=1}^{\infty}[\sum_{n_1^*=1}^{t}\cdots\sum_{n_M^*=1}^{t}\sum_{n_1^{k(t)}=l}^{t}\cdots\sum_{n_M^{k(t)}=l}^{t}(\hat{\overline{\theta^*}}_{n_1^*,\ldots,n_M^*}\\ &\quad + C_{t,(n_1^*,\ldots,n_M^*)} \leq \hat{\overline{\theta}}_{k(t),n_1^{k(t)},\ldots,n_M^{k(t)}}\\ &\quad + C_{t,(n_1^{k(t)},\ldots,n_M^{k(t)})})]\end{aligned}$$

$\hat{\overline{\theta^*}}_{n_1^*,\ldots,n_M^*} + C_{t,(n_1^*,\ldots,n_M^*)} \leq \hat{\overline{\theta}}_{k(t),n_1^{k(t)},\ldots,n_M^{k(t)}} + C_{t,(n_1^{k(t)},\ldots,n_M^{k(t)})}$ means that at least one of the following must be true:

$$\hat{\overline{\theta^*}}_{n_1^*,\ldots,n_M^*} \leq \mu^* - C_{t,(n_1^*,\ldots,n_M^*)} \tag{13}$$

$$\hat{\overline{\theta}}_{k(t),n_1^{k(t)},\ldots,n_M^{k(t)}} \geq \mu_{k(t)} + C_{t,(n_1^{k(t)},\ldots,n_M^{k(t)})} \tag{14}$$

$$\mu^* < \mu_{k(t)} + 2C_{t,(n_1^{k(t)},\ldots,n_M^{k(t)})} \tag{15}$$

Here we first find the upper bound for $Pr\{\hat{\overline{\theta^*}}_{n_1^*,\ldots,n_M^*} \leq \mu^* - C_{t,(n_1^*,\ldots,n_M^*)}\}$:

$$\begin{aligned}&Pr\{\hat{\overline{\theta^*}}_{n_1^*,\ldots,n_M^*} \leq \mu^* - C_{t,(n_1^*,\ldots,n_M^*)}\}\\ &= Pr\{\sum_{i=1}^{M}\hat{\theta}_{i,n_i^*}^* \leq \sum_{i=1}^{M}\mu_i^* - \sum_{i=1}^{M}C_{t,n_i^*}\}\\ &\leq Pr\{\text{At least one of the following must hold:}\\ &\qquad \hat{\theta}_{1,n_1^*}^* \leq \mu_1^* - C_{t,n_1^*},\\ &\qquad \hat{\theta}_{2,n_2^*}^* \leq \mu_2^* - C_{t,n_2^*},\\ &\qquad \vdots\\ &\qquad \hat{\theta}_{M,n_M^*}^* \leq \mu_M^* - C_{t,n_M^*}\}\\ &\leq \sum_{i=1}^{M}Pr\{\hat{\theta}_{i,n_i^*}^* \leq \mu_i^* - C_{t,n_i^*}\}\end{aligned}$$

$\forall 1 \le i \le M$,

$$
\begin{aligned}
&Pr\{\hat{\theta}_{i,n_i^*} \le \mu_i^* - C_{t,n_i^*}\} \\
&= Pr\{\sum_{z=1}^{|S_i^*|} \frac{\theta_i^*(z)(z) n_i^*(z)}{n_i^*} \le \sum_{z=1}^{|S_i^*|} \theta_i^*(z)\pi_i^*(z) - C_{t,n_i^*}\} \\
&= Pr\{\sum_{z=1}^{|S_i^*|} (\theta_i^*(z) n_i^*(z) - n_i^*\theta_i^*(z)\pi_i^*(z)) \le -n_i^* C_{t,n_i^*}\} \\
&\le Pr\{\text{At least one of the following must hold:} \\
&\theta_i^*(1) n_i^*(1) - n_i^*\theta_i^*(1)\pi_i^*(1) \le -\frac{n_i^*}{|S_i^*|} C_{t,n_i^*}, \\
&\qquad \vdots \\
&\theta_i^*(|S_i^*|) n_i^*(|S_i^*|) - n_i^*\theta_i^*(|S_i^*|)\pi_i^*(|S_i^*|) \le -\frac{n_i^*}{|S_i^*|} C_{t,n_i^*}\}, \\
&\le \sum_{z=1}^{|S_i^*|} Pr\{\theta_i^*(z) n_i^*(z) - n_i^*\theta_i^*(z)\pi_i^*(z) \le -\frac{n_i^*}{|S_i^*|} C_{t,n_i^*}\} \\
&= \sum_{z=1}^{|S_i^*|} Pr\{n_i^*(z) - n_i^*\pi_i^*(z) \le -\frac{n_i^*}{|S_i^*|\theta_i^*(z)} C_{t,n_i^*}\} \\
&= \sum_{z=1}^{|S_i^*|} Pr\{(n_i^* - \sum_{l\ne z} n_i^*(l)) - n_i^*(1 - \sum_{l\ne z}\pi_i^*(z)) \\
&\qquad \le -\frac{n_i^*}{|S_i^*|\theta_i^*(z)} C_{t,n_i^*}\} \\
&= \sum_{z=1}^{|S_i^*|} Pr\{\sum_{l\ne z} n_i^*(l) - n_i^* \sum_{l\ne z}\pi_i^*(z) \ge \frac{n_i^*}{|S_i^*|\theta_i^*(z)} C_{t,n_i^*}\}
\end{aligned} \tag{16}
$$

$\forall 1 \le z \le |S_i^*|$, applying Lemma 4, we could find the upper bound of each probablilty in (16) as,

$$
\begin{aligned}
&Pr\{\hat{\theta}_{i,n_i^*} \le \mu_i^* - C_{t,n_i^*}\} \\
&\le \sum_{z=1}^{|S_i^*|} \left(1 + \frac{\epsilon_{i,j}}{10|S_i^*|\theta_i^*(z)}\right) \sqrt{\frac{L\ln t}{n_i^*}} N_{\mathbf{q}_{i,j}} e^{-\frac{n_i^* L \ln t \epsilon_{i,j}}{20|S_i^*|^2\theta_i^*(z)^2 n_i^*}} \\
&\le \sum_{z=1}^{|S_i^*|} \left(1 + \frac{\epsilon_{\max}\sqrt{Lt}}{10 s_{\min}\theta_{\min}}\right) N_{\mathbf{q}_{i,j}} e^{-\frac{L\ln t \epsilon_{\min}}{20 s_{\max}^2\theta_{\max}^2}} \\
&\le \frac{s_{\max}}{\pi_{\min}}\sqrt{t}\left(1 + \frac{\epsilon_{\max}\sqrt{L}}{10 s_{\min}\theta_{\min}}\right) t^{-\frac{L\epsilon_{\min}}{20 s_{\max}^2\theta_{\max}^2}} \qquad (17) \\
&= \frac{s_{\max}}{\pi_{\min}}\left(1 + \frac{\epsilon_{\max}\sqrt{L}}{10 s_{\min}\theta_{\min}}\right) t^{-\frac{L\epsilon_{\min} - 10 s_{\max}^2\theta_{\max}^2}{20 s_{\max}^2\theta_{\max}^2}}
\end{aligned}
$$

where (17) holds since for any $\mathbf{q}_{i,j}$,

$$
\begin{aligned}
N_{\mathbf{q}_{i,j}} = \left\| \frac{q_z^{i,j}}{\pi_z^{i,j}}, z \in S_{i,j} \right\|_2 &\le \sum_{z=1}^{|S_{i,j}|} \left\| \frac{q_z^{i,j}}{\pi_z^{i,j}} \right\|_2 \\
&\le \sum_{z=1}^{|S_{i,j}|} \frac{\left\| q_z^{i,j} \right\|_2}{\pi_{\min}} = \frac{1}{\pi_{\min}}.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&Pr\{\hat{\overline{\theta^*}}_{n_1^*,\ldots,n_M^*} \le \theta^* - C_{t,(n_1^*,\ldots,n_M^*)}\} \\
&\le \frac{M s_{\max}}{\pi_{\min}}\left(1 + \frac{\epsilon_{\max}\sqrt{L}}{10 s_{\min}\theta_{\min}}\right) t^{-\frac{L\epsilon_{\min} - 10 s_{\max}^2\theta_{\max}^2}{20 s_{\max}^2\theta_{\max}^2}}.
\end{aligned} \tag{18}
$$

With the similar calculation, we can also get the upper bound of the probability for (14):

$$
\begin{aligned}
&Pr\{\hat{\overline{\theta}}_{k(t),n_1^{k(t)},\ldots,n_M^{k(t)}} \ge \mu_k + C_{t,(n_1^{k(t)},\ldots,n_M^{k(t)})}\} \\
&\le \sum_{i=1}^{M} Pr\{\hat{\theta}_{i,n_i^k}^k \ge \mu_i^k + C_{t,n_i^k}\} \\
&= \sum_{i=1}^{M} Pr\{\sum_{z=1}^{|S_i^k|} \frac{\theta_i^k(z) n_i^k(z)}{n_i^k} \ge \sum_{z=1}^{|S_i^k|} \theta_i^k(z)\pi_i^k(z) + C_{t,n_i^k}\} \\
&\le \sum_{i=1}^{M}\sum_{z=1}^{|S_i^k|} Pr\{\theta_i^k(z) n_i^k(z) - n_i^k\theta_i^k(z)\pi_i^k(z) \ge \frac{n_i^k}{|S_i^*|} C_{t,n_i^k}\} \\
&= \sum_{i=1}^{M}\sum_{z=1}^{|S_i^k|} Pr\{n_i^k(z) - n_i^k\pi_i^k(z) \ge \frac{n_i^k}{|S_i^k|\theta_i^k(z)} C_{t,n_i^k}\} \\
&\le \sum_{i=1}^{M} \frac{s_{\max}}{\pi_{\min}}\left(1 + \frac{\epsilon_{\max}\sqrt{L}}{10 s_{\min}\theta_{\min}}\right) t^{-\frac{L\epsilon_{\min} - 10 s_{\max}^2\theta_{\max}^2}{20 s_{\max}^2\theta_{\max}^2}} \\
&\le \frac{M s_{\max}}{\pi_{\min}}\left(1 + \frac{\epsilon_{\max}\sqrt{L}}{10 s_{\min}\theta_{\min}}\right) t^{-\frac{L\epsilon_{\min} - 10 s_{\max}^2\theta_{\max}^2}{20 s_{\max}^2\theta_{\max}^2}}.
\end{aligned} \tag{19}
$$

Note that for $l \ge \left\lceil \frac{4L\ln n}{\left(\frac{\Delta_{k(t)}}{M}\right)^2} \right\rceil$,

$$
\begin{aligned}
&\mu^* - \mu_{k(t)} - 2C_{t,(n_1^{k(t)},\ldots,n_M^{k(t)})} \\
&= \mu^* - \mu_{k(t)} - 2\sum_{i=1}^{M}\sqrt{\frac{L\ln t}{n_i^{k(t)}}} \\
&\ge \mu^* - \mu_{k(t)} - M\sqrt{\frac{4L\ln n}{4L\ln n}\left(\frac{\Delta_{k(t)}}{M}\right)^2} \\
&= \mu^* - \mu_{k(t)} - \Delta_{k(t)} = 0.
\end{aligned} \tag{20}
$$

(20) implies that condition (15) is false when $l = \left\lceil \frac{4L\ln n}{\left(\frac{\Delta_{k(t)}}{M}\right)^2} \right\rceil$. If we let $l = \left\lceil \frac{4L\ln n}{\left(\frac{\Delta_{\min}^{i,j}}{M}\right)^2} \right\rceil$, then (15) is false for all $k(t), 1 \le t \le \infty$ where

$$
\Delta_{\min}^{i,j} = \min_k \{\Delta_k : (i,j) \in \mathcal{A}_k\}. \tag{21}
$$

Therefore,

$$\begin{aligned}
&\mathbb{E}[\widetilde{T}_{i,j}(n)] \\
&\leq \left\lceil \frac{4L\ln n}{\left(\frac{\Delta_{\min}^{i,j}}{M}\right)^2} \right\rceil + \sum_{t=1}^{\infty}\left(\sum_{n_1^*=1}^{t}\cdots\sum_{n_1^*=M}^{t}\sum_{n_1^k=1}^{t}\cdots\sum_{n_1^k=M}^{t}\right. \\
&\left. 2M\frac{s_{\max}}{\pi_{\min}}\left(1+\frac{\epsilon_{\max}\sqrt{L}}{10s_{\min}\theta_{\min}}\right)t^{-\frac{L\epsilon_{\min}-10s_{\max}^2\theta_{\max}^2}{20s_{\max}^2\theta_{\max}^2}}\right) \\
&\leq \frac{4M^2L\ln n}{\left(\Delta_{\min}^{i,j}\right)^2}+1 \\
&+M\frac{s_{\max}}{\pi_{\min}}\left(1+\frac{\epsilon_{\max}\sqrt{L}}{10s_{\min}\theta_{\min}}\right)\sum_{t=1}^{\infty}2t^{-\frac{L\epsilon_{\min}-(40M+10)s_{\max}^2\theta_{\max}^2}{20s_{\max}^2\theta_{\max}^2}} \\
&\leq \frac{4M^2L\ln n}{\left(\Delta_{\min}^{i,j}\right)^2}+1+M\frac{s_{\max}}{\pi_{\min}}\left(1+\frac{\epsilon_{\max}\sqrt{L}}{10s_{\min}\theta_{\min}}\right)\sum_{t=1}^{\infty}2t^{-2} \qquad (22)\\
&= \frac{4M^2L\ln n}{\left(\Delta_{\min}^{i,j}\right)^2}+1+M\frac{s_{\max}}{\pi_{\min}}\left(1+\frac{\epsilon_{\max}\sqrt{L}}{10s_{\min}\theta_{\min}}\right)\frac{\pi}{3}
\end{aligned}$$

where (22) holds since $L \geq \frac{(50+40M)\theta_{\max}^2 s_{\max}^2}{\epsilon_{min}}$.

So under our MLMR policy,

$$\begin{aligned}
R_\pi(n) &\leq \sum_{k=1}^{P(N,M)}(\mu^*-\mu^k)E_\pi[T_\pi^k(n)]+C_{\mathbf{S},\mathbf{P},\Theta} \\
&= \sum_{k:\theta_k<\theta*}\Delta_k\mathbb{E}[T_k(n)]+C_{\mathbf{S},\mathbf{P},\Theta} \\
&\leq \Delta_{\max}\sum_{k:\theta_k<\theta*}\mathbb{E}[T_k(n)]+C_{\mathbf{S},\mathbf{P},\Theta} \\
&= \Delta_{\max}\sum_{i=1}^{M}\sum_{j=1}^{N}\mathbb{E}[\widetilde{T}_{i,j}(n)]+C_{\mathbf{S},\mathbf{P},\Theta} \\
&\leq \left\lceil \sum_{i=1}^{M}\sum_{j=1}^{N}\frac{4M^2L\ln n}{\left(\Delta_{\min}^{i,j}\right)^2}+1\right. \qquad (23)\\
&\left. +M\frac{s_{\max}}{\pi_{\min}}\left(1+\frac{\epsilon_{\max}\sqrt{L}}{10s_{\min}\theta_{\min}}\right)\frac{\pi}{3}\right\rceil\Delta_{\max}+C_{\mathbf{S},\mathbf{P},\Theta} \\
&\leq \left\lceil \frac{4M^3NL\ln n}{(\Delta_{\min})^2}+MN\right. \\
&\left. +M^2N\frac{s_{\max}}{\pi_{\min}}\left(1+\frac{\epsilon_{\max}\sqrt{L}}{10s_{\min}\theta_{\min}}\right)\frac{\pi}{3}\right\rceil\Delta_{\max}+C_{\mathbf{S},\mathbf{P},\Theta}
\end{aligned}$$

■

VI. Examples and Simulation Results

We consider a system that consists of $M = 2$ users and $N = 4$ resources. The state of each resource evolves as an irreducible, aperiodic Markov chain with two states "0" and "1". For all the tables in this section, the element in the i-th row and j-th column represents the value for the user-resource pair (i, j). The transition probabilities are shown in the tables below:

0.5	0.4	0.7	0.3
0.2	0.9	0.9	0.7

$\mathbf{p}_{01}$

0.6	0.7	0.8	0.9
0.9	0.5	0.4	0.4

$\mathbf{p}_{10}$

The rewards on each states are:

0.6	0.5	0.2	0.4
0.3	0.7	0.8	0.3

$\boldsymbol{\theta}_0$

0.8	0.2	0.7	0.5
0.5	0.3	0.6	0.6

$\boldsymbol{\theta}_1$

For $1 \leq i \leq M$, $1 \leq j \leq N$, the stationary distribution of user-resource pair (i, j) on state "0" is calculated as $\frac{p_{10}^{i,j}}{p_{01}^{i,j}+p_{10}^{i,j}}$; the stationary distribution on state "1" is calculated as $\frac{p_{01}^{i,j}}{p_{01}^{i,j}+p_{10}^{i,j}}$. The eigenvalue gap is $\epsilon_{i,j} = p_{01}^{i,j} + p_{10}^{i,j}$. The expected reward $\mu_{i,j}$ for all the pairs can be calculated as:

0.6909	0.3909	0.4333	0.425
0.3363	0.4429	0.6615	0.4909

$\boldsymbol{\mu}$

We can see that the arm $\{(1,1),(2,3)\}$ is the optimal arm with greatest expected reward $\mu^* = 0.6909+0.6615 = 1.3524$. $\Delta_{\min} = 0.1706$.

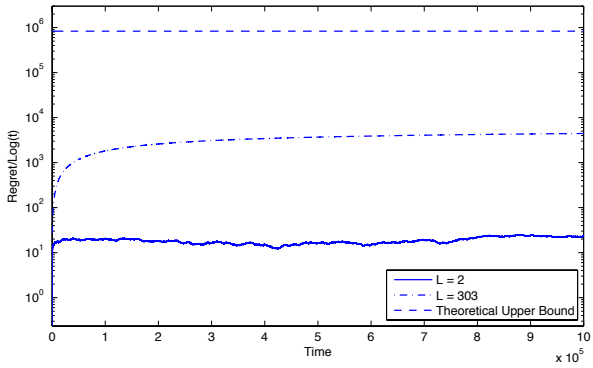

Fig. 1. Simulation Results of Example 1 with $\Delta_{\min} = 0.1706$

Figure 1 shows the simulation result of the regret (normalized with respect to the logarithm of time) for our MLMR policy for the above system with different choices of L. We also show the theoretical upper bound for comparison. The value of L to satisfy the condition in Theorem 1 is $L \geq \frac{(50+40M)R^2 s_{\max}^2}{\epsilon_{min}} = 303$, so we picked $L = 303$ in the simulation.

Note that in the proof of Theorem 1, when $L < \frac{(50+40M)R^2 s_{\max}^2}{\epsilon_{min}}$, $-\frac{L\epsilon_{\min}-(40M+10)s_{\max}^2\theta_{\max}^2}{20s_{\max}^2\theta_{\max}^2} > -2$. This implies $\sum_{t=1}^{\infty}2t^{-\frac{L\epsilon_{\min}-(40M+10)s_{\max}^2\theta_{\max}^2}{20s_{\max}^2\theta_{\max}^2}}$ does not converge anymore and thus we could not bound $\mathbb{E}[\widetilde{T}_{i,j}(n)]$ any more. Empirically, however, in 1 the case when $L = 2$ also seems to yield logarithmic regret over time and the performance is in fact better than $L = 303$, since the "bad" arms (arms

which are not optimal) are picked less when L is smaller. However, this may possibly be due to the fact that the cases when $\widetilde{T}_{i,j}(n)$ grows faster than $\log(t)$ only happens with very small probability when $L = 2$.

Table II shows the number of times that resource j has been matched with user i up to time $n = 10^7$.

999470	153	185	196
136	293	999155	420

$n_{i,j}(10^7)$, $L = 2$

892477	30685	39410	37432
26813	50341	850265	72585

$n_{i,j}(10^7)$, $L = 303$

TABLE II

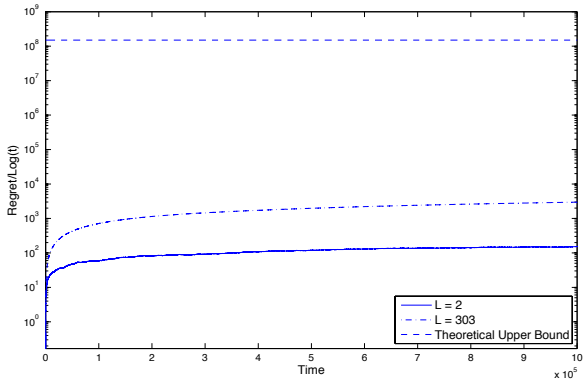

Fig. 2. Simulation Results of Example 2 with $\Delta_{\min} = 0.0091$

Figure 2 shows another example with the same transition probabilities as in the previous example and different rewards on states as below:

0.7	0.3	0.5	0.5
0.65	0.7	0.8	0.4

$\boldsymbol{\theta}_0$

0.4	0.6	0.7	0.45
0.5	0.5	0.6	0.55

$\boldsymbol{\theta}_1$

The expected reward $\mu_{i,j}$ for all the pairs can be calculated as:

0.5636	0.4091	0.5933	0.4875
0.6227	0.5714	0.6615	0.4954

$\boldsymbol{\mu}$

$\{(1,1),(2,3)\}$ is still the optimal arm. However, compared with the previous example, we can see that the expected reward of three other arms $\{(1,3),(2,1)\}$, $\{(1,3),(2,2)\}$, $\{(1,1),(2,2)\}$ are all very close to the expected reward of the optimal arm. For this example, $\Delta_{\min} = 0.0091$, which is much smaller compared with the previous example. In this case, we can see from Figure 2 that the non-optimal arms are picked much more compared with the previous example. This is because we have several arms of which the expected rewards are very close to μ^*, so the policy has to spend a lot more time to explore on those non-optimal arms to make sure those are non-optimal arms. This fact can be seen clearly in Table III, which presents the number of times that resource j has been matched with user i up to time $n = 10^7$ under both cases when $L = 2$ and $L = 303$.

817529	544	179832	2099
175583	3610	820097	714

$n_{i,j}(10^7)$, $L = 2$

346395	60031	472346	121232
301491	146317	482545	69651

$n_{i,j}(10^7)$, $L = 303$

TABLE III

VII. Conclusion

We have presented the MLMR policy for the problem of learning combinatorial matchings of users to resources when the reward process is Markovian. We showed that this policy requires only polynomial storage and computation per step, and yields a regret that grows uniformly logarithmically over time and only polynomially with the number of users and resources.

In future work, we would like to consider the case when the rewards evolve not just when a user-resource pair is selected, but rather at each discrete time. Further, we would like to investigate if it is possible to analyze regret with respect to the best non-static policy. Finally, exploring distributed schemes is also of interest, though likely to be highly challenging in case of limited information exchange between users.

References

[1] S. Pandey, D. Chakrabarti, and D. Agarwal, "Multi-armed bandit problems with dependent arms," In *Proc. of the 24th International Conference on Machine Learning* , Corvalis, June 2007.

[2] Y. Gai, B. Krishnamachari, and R. Jain, "Learning multiuser channel allocations in cognitive radio networks: a combinatorial multi-armed bandit formulation," *IEEE Symp. on Dynamic Spectrum Access Networks (DySPAN)*, Singapore, April 2010.

[3] K. Liu and Q. Zhao, "Decentralized multi-armed bandit with multiple distributed players," *Information Theory and Applications Workshop (ITA)*, San Diego, January 2010.

[4] A. Anandkumar, N. Michael, and A. K. Tang. "Opportunistic spectrum access with multiple users: learning under competition," *Proc. of IEEE INFOCOM*, San Deigo, March 2010.

[5] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4-22, 1985.

[6] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part I: IID rewards," *IEEE Tran. on Auto. Control*, vol. 32, no. 11, pp. 968-976, 1987.

[7] R. Agrawal, "Sample mean based index policies with O(log n) regret for the multi-armed bandit problem," *Advances in Applied Probability*, vol. 27, no. 4, pp. 1054-1078, 1995.

[8] 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.

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

[10] C. Tekin, M. Liu, "Online algorithms for the multi-armed bandit problem with markovian rewards", http://arxiv.org/abs/1007.2238.

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

[12] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1, pp. 83-97, 1955.

[13] D. Gillman, "A chernoff bound for random walks on expander graphs", *SIAM Journal on Computing*, vol. 27, no. 4, pp. 1203-1220, 1998