A Hoeffding Inequality for Finite State Markov Chains and its Applications to Markovian Bandits

Vrettos Moulos
Department of Electrical Engineering and Computer Sciences
University of California Berkeley
vrettos@berkeley.edu

***Abstract*—This paper develops a Hoeffding inequality for the partial sums $\sum_{k=1}^{n} f(X_k)$, where $\{X_k\}_{k\in\mathbb{Z}_{>0}}$ is an irreducible Markov chain on a finite state space S, and $f : S \to [a, b]$ is a real-valued function. Our bound is simple, general, since it only assumes irreducibility and finiteness of the state space, and powerful. In order to demonstrate its usefulness we provide two applications in multi-armed bandit problems. The first is about identifying an approximately best Markovian arm, while the second is concerned with regret minimization in the context of Markovian bandits.**

I. Introduction

Let $\{X_k\}_{k\in\mathbb{Z}_{>0}}$ be a Markov chain on a finite state space S, with initial distribution q, and irreducible transition probability matrix P, governed by the probability law $\mathbb{P}_q$. Let π be its stationary distribution, and $f : S \to [a, b]$ be a real-valued function on the state space. Then the strong law of large numbers for Markov chains asserts that,

$$\frac{1}{n}\sum_{k=1}^{n} f(X_k) \overset{\mathbb{P}_q\,-a.s.}{\to} \mathbb{E}_\pi[f(X_1)], \text{ as } n \to \infty.$$

Moreover, the central limit theorem for Markov chains provides a rate for this convergence,

$$\sqrt{n}\left(\frac{1}{n}\sum_{k=1}^{n} f(X_k) - \mathbb{E}_\pi[f(X_1)]\right) \overset{d}{\to} N(0, \sigma^2), \text{ as } n \to \infty,$$

where $\sigma^2 = \lim_{n\to\infty} \frac{1}{n} \operatorname{var}_q\left(\sum_{k=1}^{n} f(X_k)\right)$ is the limiting variance.

Those asymptotic results are insufficient in many applications which require finite-sample estimates. One of the most central such application is the convergence of Markov chain Monte Carlo (MCMC) approximation techniques [1], where a finite-sample estimate is needed to bound the approximation error. Further applications include theoretical computer science and the approximation of the permanent [2], as well as statistical learning theory and multi-armed bandit problems [3].

Motivated by this discussion we provide a finite-sample Hoeffding inequality for finite Markov chains. In the special case that the random variables $\{X_k\}_{k\in\mathbb{Z}_{>0}}$ are independent and identically distributed according to π, Hoeffding's classic inequality [4] states that,

$$\mathbb{P}\left(\left|\sum_{k=1}^{n}\left(f(X_k) - \mathbb{E}[f(X_1)]\right)\right| \geq t\right) \leq 2\exp\left\{-\frac{t^2}{2\nu^2}\right\},$$

where $\nu^2 = \frac{1}{4}n(b-a)^2$. In Theorem 1 we develop a version of Hoeffding's inequality for finite state Markov chains. Our bound is very simple and easily computable, since it is based on martingale techniques and it only involves hitting times of Markov chains which are very well studied for many types of Markov chains [5]. It is worth mentioning that our bound is based solely on irreducibility, and it does not make any extra assumptions like aperiodicity or reversibility which prior works require.

There is a rich literature on finite-sample bounds for Markov chains. One of the earliest works [6] uses counting and a generalization of the method of types, in order to derive a Chernoff bound for Markov chains which are irreducible and aperiodic. An alternative approach [7], [8], uses the theory of large deviations to derive sharper Chernoff bounds. When reversibility is assumed, the transition probability matrix is symmetric with respect to the space $L^2(\pi)$, which enables the use of matrix perturbation theory. This idea leads to Hoeffding inequalities that involve the spectral gap of the Markov chain and was initiated in [9]. Refinements of this bound were given in a series of works [10]–[14]. In [15]–[17] a generalized spectral gap is introduced in order to obtain bounds even for a certain class of irreversible Markov chains as long as they posses a strictly positive generalized spectral gap. Information-theoretic ideas are used in [18] in order to derive a Hoeffding inequality for Markov chains with general state spaces that satisfy Doeblin's minorization condition, which in the case of a finite state space can be written as,

$$\exists m \in \mathbb{Z}_{>0}\ \exists y \in S\ \forall x \in S :\ P^m(x, y) > 0. \tag{1}$$

Of course there are irreducible transition probability matrices P for which (1) fails, but if we further assume aperiodicity, then (1) is satisfied. Our approach uses Doob's martingale combined with Azuma's inequality, and is probably closest related to the work in [19], where a bound for Markov chains with general state spaces is established using Dynkin's martingale. But the result in [19] heavily relies on the Markov chains satisfying Doeblin's condition (1). A regeneration approach for uniformly ergodic Markov chains, where one splits the Markov chain in i.i.d. blocks, and reduces the problem to the concentration of an i.i.d. process, can be found in [20].

Another line of research is related to the concentration of a function of n random variables around its mean, under Markovian or other dependent structures. This was pioneered

978-1-7281-6432-8/20/$31.00 ©2020 IEEE

by the works of Marton [21]–[23] who used the transportation method, and further developed using coupling ideas in [15], [24]–[27].

To illustrate the applicability of our bound we use it to study two Markovian multi-armed bandit problems. The stochastic multi-armed bandits problem is a prototypical statistical learning problem that exhibits an exploration-exploitation trade-off. One is given multiple options, referred to as arms, and each of them associated with a probability distribution. The emphasis is put on focusing as quickly as possible on the best available option, rather than estimating with high confidence the statistics of each option. The cornerstone of this field is the pioneering work of Lai and Robbins [28]. Here we study two variants of the multi-armed bandits problem where the probability distributions of the arms form Markov chains. First we consider the task of identifying with some fixed confidence an approximately best arm, and we use our bound to analyze the median elimination algorithm, originally proposed in [29] for the case of i.i.d. bandits. Then we turn into the problem of regret minimization for Markovian bandits, where we analyze the UCB algorithm that was introduced in [30] for i.i.d. bandits. For a thorough introduction to multi-armed bandits we refer the interested reader to the survey [31], and the books [32], [33].

II. A Hoeffding Inequality for Irreducible Finite State Markov Chains

The central quantity that shows up in our Hoeffding inequality, and makes it differ from the classical i.i.d. Hoeffding inequality, is the maximum hitting time of a Markov chain with an irreducible transition probability matrix P. This is defined as,

$$\mathrm{HitT}(P) = \max_{x,y\in S} \mathbb{E}[T_y \mid X_1 = x],$$

where $T_y = \inf\{n \geq 0 : X_{n+1} = y\}$ is the number of transitions taken in order to visit state y for the first time. $\mathrm{HitT}(P)$ is ensured to be finite due to irreducibility and the finiteness of the state space.

Theorem 1: Let $\{X_k\}_{k\in\mathbb{Z}_{>0}}$ be a Markov chain on a finite state space S, driven by an initial distribution q, and an irreducible transition probability matrix P. Let $f : S \to [a, b]$ be a real-valued function. Then, for any $t > 0$,

$$\mathbb{P}_q\left(\left|\sum_{k=1}^{n}\left(f(X_k) - \mathbb{E}_q\left[f(X_k)\right]\right)\right| \geq t\right) \leq 2\exp\left\{-\frac{t^2}{2\nu^2}\right\},$$

where $\nu^2 = \frac{1}{4}n(b-a)^2\mathrm{HitT}(P)^2$.

Proof: We define the sums,

$$S_{l,m} = f(X_l) + \ldots + f(X_m), \text{ for } 1 \leq l \leq m \leq n,$$

and the filtration $\mathcal{F}_0 = \sigma(\emptyset)$, $\mathcal{F}_k = \sigma(X_1, \ldots, X_k)$ for $k = 1, \ldots, n$. Then $\{\mathbb{E}(S_{1,n} \mid \mathcal{F}_k) - \mathbb{E}(S_{1,n} \mid \mathcal{F}_0)\}_{k=0}^{n}$, is a zero mean martingale with respect to $\{\mathcal{F}_k\}_{k=0}^{n}$, and let $\Delta_k = \mathbb{E}(S_{1,n} \mid \mathcal{F}_k) - \mathbb{E}(S_{1,n} \mid \mathcal{F}_{k-1})$, for $k = 1, \ldots, n$, be the martingale differences.

We first note the following bounds on the martingale differences,

$$\min_{y\in S} \mathbb{E}(S_{1,n} \mid \mathcal{F}_{k-1}, X_k = y) - \mathbb{E}(S_{1,n} \mid \mathcal{F}_{k-1}) \leq \Delta_k,$$

and

$$\Delta_k \leq \max_{x\in S} \mathbb{E}(S_{1,n} \mid \mathcal{F}_{k-1}, X_k = x) - \mathbb{E}(S_{1,n} \mid \mathcal{F}_{k-1}).$$

Therefore, in order to bound the variation of Δ_k it suffices to control,

$$\begin{aligned}
&\max_{x\in S} \mathbb{E}(S_{1,n} \mid \mathcal{F}_{k-1}, X_k = x) - \min_{y\in S} \mathbb{E}(S_{1,n} \mid \mathcal{F}_{k-1}, X_k = y) \\
&\quad = \max_{x,y\in S} \left\{\mathbb{E}[S_{k,n} \mid X_k = x] - \mathbb{E}[S_{k,n} \mid X_k = y]\right\} \\
&\quad = \max_{x,y\in S} \left\{\mathbb{E}[S_{1,n-k+1} \mid X_1 = x] - \mathbb{E}[S_{1,n-k+1} \mid X_1 = y]\right\},
\end{aligned}$$

where in the first equality we used the Markov property, and in the second the time-homogeneity.

We now use a hitting time argument. Observe the following pointwise statements,

$$S_{1,n-k+1} \leq T_y b + S_{T_y+1,n-k+1},$$

and,

$$S_{T_y+1,n-k+1} + T_y a \leq S_{T_y+1,T_y+n-k+1},$$

from which we deduce that,

$$S_{1,n-k+1} \leq T_y(b-a) + S_{T_y+1,T_y+n-k+1}.$$

Taking $\mathbb{E}[\cdot \mid X_1 = x]$-expectations, and using the strong Markov property we obtain,

$$\begin{aligned}
&\mathbb{E}[S_{1,n-k+1} \mid X_1 = x] \\
&\quad \leq (b-a)\,\mathbb{E}[T_y \mid X_1 = x] + \mathbb{E}[S_{1,n-k+1} \mid X_1 = y].
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\max_{x,y\in S} \left\{\mathbb{E}[S_{1,n-k+1} \mid X_1 = x] - \mathbb{E}[S_{1,n-k+1} \mid X_1 = y]\right\} \\
&\quad \leq (b-a)\mathrm{HitT}(P).
\end{aligned}$$

With this in our possession we apply Hoeffding's lemma, see for instance Lemma 2.3 in [34], in order to get,

$$\begin{aligned}
\mathbb{E}\left(e^{\theta\Delta_k} \mid \mathcal{F}_{k-1}\right) &\leq \exp\left\{\frac{\theta^2(b-a)^2\mathrm{HitT}(P)^2}{8}\right\} \\
&= \exp\left\{\frac{\theta^2\nu^2}{2n}\right\}, \text{ for all } \theta \in \mathbb{R}.
\end{aligned}$$

Using Markov's inequality, and successive conditioning we obtain that for $\theta > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\sum_{k=1}^{n}\left(f(X_k) - \mathbb{E}_q\left[f(X_k)\right]\right) \geq t\right) \\
&\quad \leq e^{-\theta t}\,\mathbb{E}\left[e^{\theta\left(\sum_{k=1}^{n}\Delta_k\right)}\right] \\
&\quad = e^{-\theta t}\,\mathbb{E}\left[\mathbb{E}\left(e^{\theta\Delta_n}\middle|\mathcal{F}_{n-1}\right)e^{\theta\left(\sum_{k=1}^{n-1}\Delta_k\right)}\right] \\
&\quad \leq \exp\left\{-\theta t + \frac{\theta^2\nu^2}{2n}\right\}\mathbb{E}\left[e^{\theta\left(\sum_{k=1}^{n-1}\Delta_k\right)}\right] \\
&\quad \leq \ldots \leq \exp\left\{-\theta t + \frac{\theta^2\nu^2}{2}\right\}.
\end{aligned}$$

Plugging in $\theta = t/\nu^2$, we see that,

$$\mathbb{P}\left(\sum_{k=1}^{n}\left(f(X_k) - \mathbb{E}_q\left[f(X_k)\right]\right) \geq t\right) \leq \exp\left\{-\frac{t^2}{2\nu^2}\right\}.$$

The conclusion follows by combining the inequality above for f and $-f$. ∎

Example 1: Consider a two-state Markov chain with $S = \{0,1\}$ and $P(0,1) = p,\ P(1,0) = r$, with $p, r \in (0,1]$. Then,

$$\begin{aligned}\mathrm{HitT}(P) &= \max\{\mathbb{E}[\text{Geometric}(p)], \mathbb{E}[\text{Geometric}(r)]\} \\ &= 1/\min\{p,r\},\end{aligned}$$

and Theorem 1 takes the form,

$$\mathbb{P}_q\left(\left|\sum_{k=1}^{n}\left(f(X_k) - \mathbb{E}_q\left[f(X_k)\right]\right)\right| \geq t\right) \leq 2\exp\left\{-\frac{2\min\{p^2,r^2\}t^2}{n(b-a)^2}\right\}.$$

Example 2: Consider the random walk on the m-cycle with state space $S = \{0,1,\ldots,m-1\}$, and transition probability matrix $P(x,y) = (1\{y \equiv x+1 \pmod m\} + 1\{y \equiv x-1 \pmod m\})/2$. If m is odd, then the Markov chain is aperiodic, while if m is even, then the Markov chain has period 2. Then,

$$\mathrm{HitT}(P) = \max_{y\in S}\mathbb{E}[T_y \mid X_1 = 0] = \max_{y\in S} y(m-y) = \lfloor m^2/4 \rfloor,$$

and Theorem 1 takes the form,

$$\mathbb{P}_q\left(\left|\sum_{k=1}^{n}\left(f(X_k) - \mathbb{E}_q\left[f(X_k)\right]\right)\right| \geq t\right) \leq 2\exp\left\{-\frac{2t^2}{n(b-a)^2\lfloor m^2/4\rfloor^2}\right\}.$$

Remark 1: Observe that the technique used to establish Theorem 1 is limited to Markov chains with a finite state space S. Indeed, if $\{X_k\}_{k\in\mathbb{Z}_{>0}}$ is a Markov chain on a countably infinite state space S with an irreducible and positive recurrent transition probability matrix P and a stationary distribution π, then we claim that,

$$\frac{1}{\pi(y)} \leq 1 + \sup_{x\in S}\mathbb{E}[T_y \mid X_1 = x], \text{ for all } y \in S,$$

from which it follows that $\sup_{x,y\in S}\mathbb{E}[T_y \mid X_1 = x] = \infty$, due to the fact that $\sum_{y\in S}\pi(y) = 1$ and S is countably infinite. The aforementioned inequality can be established as follows.

$$\begin{aligned}\frac{1}{\pi(y)} &= \mathbb{E}[\inf\{n \geq 1 : X_{n+1} = y\} \mid X_1 = y] \\ &= \sum_{x\in S}\mathbb{E}[\inf\{n \geq 1 : X_{n+1} = y\} \mid X_2 = x]P(y,x) \\ &\leq \sup_{x\in S}\mathbb{E}[\inf\{n \geq 1 : X_{n+1} = y\} \mid X_2 = x] \\ &= 1 + \sup_{x\in S}\mathbb{E}[T_y \mid X_1 = x].\end{aligned}$$

Moreover, through Theorem 1 we can obtain a concentration inequality for sums of a function evaluated on the transitions of a Markov chain. In particular, let

$$S^{(2)} = \{(x,y) \in S \times S : P(x,y) > 0\}.$$

On the state space $S^{(2)}$ define the transition probability matrix,

$$P^{(2)}\left((x,y),(z,w)\right) = I\{y = z\}P(y,w).$$

It is straightforward to verify that the fact that P is irreducible, implies that $P^{(2)}$ is irreducible as well. This readily gives the following theorem.

Theorem 2: Let $\{X_k\}_{k\in\mathbb{Z}_{>0}}$ be a Markov chain on a finite state space S, driven by an initial distribution q, and an irreducible transition probability matrix P. Let $f^{(2)} : S^{(2)} \to [a,b]$ be a real-valued function evaluated on the transitions of the Markov chain. Then, for any $t > 0$,

$$\mathbb{P}_q\left(\left|\sum_{k=1}^{n}\left(f^{(2)}(X_k, X_{k+1}) - \mathbb{E}_q\left[f^{(2)}(X_k, X_{k+1})\right]\right)\right| \geq t\right) \leq 2\exp\left\{-\frac{t^2}{2\nu^2}\right\},$$

where $\nu^2 = \frac{1}{4}n(b-a)^2\mathrm{HitT}\left(P^{(2)}\right)^2$.

Corollary 1: When the Markov chain is initialized with its stationary distribution, π, Theorem 1 and Theorem 2 give the following nonasymptotic versions of the weak law of large numbers for irreducible Markov chains. For any $\epsilon > 0$,

$$\mathbb{P}_\pi\left(\left|\frac{1}{n}\sum_{k=1}^{n} f(X_k) - \mathbb{E}_\pi\left[f(X_1)\right]\right| \geq \epsilon\right) \leq 2\exp\left\{-\frac{2n\epsilon^2}{(b-a)^2\mathrm{HitT}(P)^2}\right\},$$

and,

$$\mathbb{P}_\pi\left(\left|\frac{1}{n}\sum_{k=1}^{n} f^{(2)}(X_k, X_{k+1}) - \mathbb{E}_\pi\left[f^{(2)}(X_1, X_2)\right]\right| \geq \epsilon\right) \leq 2\exp\left\{-\frac{2n\epsilon^2}{(b-a)^2\mathrm{HitT}\left(P^{(2)}\right)^2}\right\}.$$

III. Markovian Multi-Armed Bandits

A. Setup

There are $K \geq 2$ arms, and each arm $a \in [K] = \{1,\ldots,K\}$ is associated with a parameter $\theta_a \in \mathbb{R}$ which uniquely encodes[1] an irreducible transition probability matrix P_{θ_a}. We will denote the overall parameter configuration of all K arms with $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_K) \in \mathbb{R}^K$. Arm a evolves according to the stationary Markov chain, $\{X_n^a\}_{n\in\mathbb{Z}_{>0}}$, driven by the irreducible transition probability matrix P_{θ_a} which has a unique stationary distribution π_{θ_a}, so that $X_1^a \sim \pi_{\theta_a}$. There is a common reward function $f : S \to [c,d]$ which generates the reward process $\{Y_n^a\}_{n\in\mathbb{Z}_{>0}} = \{f(X_n^a)\}_{n\in\mathbb{Z}_{>0}}$. The reward process, in general, is not going to be a Markov chain, unless f is injective, and it will have more complicated dependencies than the underlying Markov chain. Each time that we select arm a, this arm evolves by one transition and we observe the corresponding sample from the reward process $\{Y_n^a\}_{n\in\mathbb{Z}_{>0}}$, while all the other arms stay rested.

[1] $\mathbb{R}$ and the set of $|S| \times |S|$ irreducible transition probability matrices have the same cardinality, and hence there is a bijection between them.

The stationary reward of arm a is $\mu(\theta_a) = \sum_{x\in S} f(x)\pi_{\theta_a}(x)$. Let $\mu^*(\boldsymbol{\theta}) = \max_{a\in[K]} \mu(\theta_a)$ be the maximum stationary mean, and for simplicity assume that there exists a unique arm, $a^*(\boldsymbol{\theta})$, attaining this maximum stationary mean, i.e. $\{a^*(\boldsymbol{\theta})\} = \arg\max_{a\in[K]} \mu(\theta_a)$. In the following sections we will consider two objectives: identifying an ϵ best arm with some fixed confidence level δ using as few samples as possible, and minimizing the expected regret given some fixed time horizon T.

B. Approximate Best Arm Identification

In the approximate best arm identification problem, we are given an approximation accuracy $\epsilon > 0$, and a confidence level $\delta \in (0,1)$. Our goal is to come up with an adaptive algorithm $\mathcal{A}$ which collects a total of N samples, and returns an arm $\hat{a}$ that is within ϵ from the best arm, $a^*(\boldsymbol{\theta})$, with probability at least $1-\delta$, i.e.

$$\mathbb{P}^{\mathcal{A}}_{\boldsymbol{\theta}}(\mu^*(\boldsymbol{\theta}) \geq \mu(\theta_{\hat{a}}) + \epsilon) \leq \delta.$$

Such an algorithm is called (ϵ,δ)-PAC (probably approximately correct).

In [35] a lower bound for the sample complexity of any (ϵ,δ)-PAC algorithm is derived. The lower bound states that no matter the (ϵ,δ)-PAC algorithm $\mathcal{A}$, there exists an instance $\boldsymbol{\theta}$ such that the sample complexity is at least,

$$\mathbb{E}^{\mathcal{A}}_{\boldsymbol{\theta}}[N] = \Omega\left(\frac{K}{\epsilon^2}\log\frac{1}{\delta}\right).$$

A matching upper bound is provided for IID bandits in [29] in the form of the median elimination algorithm. We demonstrate the usefulness of our Hoeffding inequality, by providing an analysis of the median elimination algorithm in the more general setting of Markovian bandits.

Algorithm 1: The β-Median-Elimination algorithm.

Parameters: *number of arms* $K \geq 2$*, approximation accuracy* $\epsilon > 0$*, confidence level* $\delta \in (0,1)$*, parameter* β;
$r = 1,\ A_r = [K],\ \epsilon_r = \epsilon/4,\ \delta_r = \delta/2$;
while $|A_r| \geq 2$ **do**
  $N_r = \left\lceil \frac{4\beta}{\epsilon_r^2}\log\frac{3}{\delta_r}\right\rceil$;
  Sample each arm in A_r for N_r times;
  For $a \in A_r$ calculate $\bar{Y}_a[r] = \frac{1}{N_r}\sum_{n=1}^{N_r} Y_n^a$;
  $m_r = \textbf{median}\left((\bar{Y}_a[r])_{a\in A_r}\right)$;
  Pick A_{r+1} such that:
    - $A_{r+1} \subseteq \{a \in A_r : \bar{Y}_a[r] \geq m_r\}$;
    - $|A_{r+1}| = \lfloor |A_r|/2 \rfloor$;
  $\epsilon_{r+1} = 3\epsilon_r/4,\ \delta_{r+1} = \delta_r/2,\ r = r+1$;
end
return $\hat{a}$*, where* $A_r = \{\hat{a}\}$;

Theorem 3: If $\beta \geq \frac{1}{2}(d-c)^2 \max_{a\in[K]} \mathrm{HitT}(P_{\theta_a})^2$ then, the β-Median-Elimination algorithm is (ϵ,δ)-PAC, and its sample complexity is upper bounded by $O\left(\frac{K}{\epsilon^2}\log\frac{1}{\delta}\right)$.

Proof: The total number of sampling rounds is at most $\lceil \log_2 K \rceil$, and we can set them equal to $\lceil \log_2 K \rceil$ by setting $A_r = \{\hat{a}\}$, for $r \geq R_0$, where $A_{R_0} = \{\hat{a}\}$. Fix $r \in \{1,\ldots,\lceil \log_2 K \rceil\}$. We claim that,

$$\mathbb{P}^{\beta-\mathrm{ME}}_{\boldsymbol{\theta}}\left(\max_{a\in A_r}\mu(\theta_a) \geq \max_{a\in A_{r+1}}\mu(\theta_a) + \epsilon_r\right) \leq \delta_r. \quad (2)$$

We condition on the value of A_r. If $|A_r| = 1$, then the claim is trivially true, so we only consider the case $|A_r| \geq 2$. Let $\mu_r^* = \max_{a\in A_r}\mu(\theta_a)$, and $a_r^* \in \arg\max_{a\in A_r:\mu(\theta_a)=\mu_r^*} \bar{Y}_a[r]$. We consider the following set of bad arms,

$$B_r = \{b \in A_r : \bar{Y}_b[r] \geq \bar{Y}_{a_r^*}[r],\ \mu_r^* \geq \mu(\theta_b) + \epsilon_r\},$$

and observe that,

$$\mathbb{P}^{\beta-\mathrm{ME}}_{\boldsymbol{\theta}}\left(\mu_r^* \geq \mu_{r+1}^* + \epsilon_r\right) \leq \mathbb{P}^{\beta-\mathrm{ME}}_{\boldsymbol{\theta}}(|B_r| \geq |A_r|/2). \quad (3)$$

In order to upper bound the latter fix $b \in A_r$ and write,

$$\begin{aligned}\mathbb{P}^{\beta-\mathrm{ME}}_{\boldsymbol{\theta}}&\left(\bar{Y}_b[r] \geq \bar{Y}_{a_r^*}[r], \mu_r^* \geq \mu(\theta_b) + \epsilon_r \middle| \bar{Y}_{a_r^*}[r] > \mu_r^* - \epsilon_r/2\right)\\ &\leq \mathbb{P}_{\theta_b}(\bar{Y}_b[r] \geq \mu(\theta_b) + \epsilon_r/2) \leq \delta_r/3,\end{aligned}$$

where in the last inequality we used Corollary 1. Now via Markov's inequality this yields,

$$\mathbb{P}^{\beta-\mathrm{ME}}_{\boldsymbol{\theta}}\left(|B_r| \geq |A_r|/2 \middle| \bar{Y}_{a_r^*}[r] > \mu_r^* - \epsilon_r/2\right) \leq 2\delta_r/3. \quad (4)$$

Furthermore, Corollary 1 gives that for any $a \in A_r$,

$$\mathbb{P}_{\theta_a}(\bar{Y}_a[r] \leq \mu(\theta_a) - \epsilon_r/2) \leq \delta_r/3. \quad (5)$$

We obtain (2) by using (4) and (5) in (3).

With (2) in our possession, the fact that median elimination is (ϵ,δ)-PAC follows through a union bound,

$$\begin{aligned}&\mathbb{P}^{\beta-\mathrm{ME}}_{\boldsymbol{\theta}}(\mu^*(\boldsymbol{\theta}) \geq \mu(\theta_{\hat{a}}) + \epsilon)\\ &\quad\leq \mathbb{P}^{\beta-\mathrm{ME}}_{\boldsymbol{\theta}}\left(\bigcup_{r=1}^{\lceil \log_2 K\rceil}\left\{\mu_r^* \geq \mu_{r+1}^* + \epsilon_r\right\}\right)\\ &\quad\leq \sum_{r=1}^{\infty}\delta_r \leq \delta.\end{aligned}$$

Regarding the sample complexity, we have that the total number of samples is at most,

$$\begin{aligned}K\sum_{r=1}^{\lceil \log_2 K\rceil} N_r/2^{r-1} &\leq 2K + \frac{64\beta K}{\epsilon^2}\sum_{r=1}^{\infty}\left(\frac{8}{9}\right)^{r-1}\log\frac{2^r 3}{\delta}\\ &= O\left(\frac{K}{\epsilon^2}\log\frac{1}{\delta}\right).\end{aligned}$$

■

C. Regret Minimization

Our device to solve the regret minimization problem is an *adaptive allocation rule*, $\boldsymbol{\phi} = \{\phi_t\}_{t\in\mathbb{Z}_{>0}}$, which is a sequence of random variables where $\phi_t \in [K]$ is the arm that we select at time t. Let $N_a(t) = \sum_{s=1}^{t} I_{\{\phi_s=a\}}$, be the number of times we selected arm a up to time t. Our decision, ϕ_t, at time t is based on the information that we have accumulated so far. More precisely, the event $\{\phi_t = a\}$ is measurable with respect

to the σ-field generated by the past decisions $\phi_1, \ldots, \phi_{t-1}$, and the past observations $\{X_n^1\}_{n=1}^{N_1(t-1)}, \ldots, \{X_n^K\}_{n=1}^{N_K(t-1)}$.

Given a time horizon T, and a parameter configuration $\boldsymbol{\theta}$, the expected regret incurred when the adaptive allocation rule $\boldsymbol{\phi}$ is used, is defined as,

$$R_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(T) = \sum_{b \neq a^*(\boldsymbol{\theta})} \mathbb{E}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}[N_b(T)]\Delta_b(\boldsymbol{\theta}),$$

where $\Delta_b(\boldsymbol{\theta}) = \mu^*(\boldsymbol{\theta}) - \mu(\theta_b)$. Our goal is to come up with an adaptive allocation rule that makes the expected regret as small as possible.

There is a known asymptotic lower bound on how much we can minimize the expected regret. Any adaptive allocation rule that is uniformly good across all parameter configurations should satisfy the following instance specific, asymptotic regret lower bound (see [36] for details),

$$\sum_{b \neq a^*(\boldsymbol{\theta})} \frac{\Delta_b(\boldsymbol{\theta})}{\overline{D}\left(\theta_b \,\middle\|\, \theta_{a^*(\boldsymbol{\theta})}\right)} \leq \liminf_{T\to\infty} \frac{R_{\boldsymbol{\theta}}^{\boldsymbol{\phi}}(T)}{\log T},$$

where $\overline{D}\left(\theta \,\|\, \lambda\right)$ is the Kullback-Leibler divergence rate between the Markov chains with transition probability matrices P_θ and P_λ, given by,

$$\overline{D}\left(\theta \,\|\, \lambda\right) = \sum_{x,y \in S} \log \frac{P_\theta(x,y)}{P_\lambda(x,y)} \pi_\theta(x) P_\theta(x,y).$$

Here we utilize our Theorem 1 to provide a finite-time analysis of the β-UCB adaptive allocation rule for Markovian bandits, which is order optimal. The β-UCB adaptive allocation rule, is a simple and computationally efficient index policy based on upper confidence bounds which was initially proposed in [30] for IID bandits. It has already been studied in the context of Markovian bandits in [37], but in a more restrictive setting under the further assumptions of aperiodicity and reversibility due the use of the bounds from [9], [12]. For adaptive allocation rules that asymptotically match the lower bound we refer the interested reader to [36], [38].

Algorithm 2: The β-UCB adaptive allocation rule.

Parameters: *number of arms* $K \geq 2$*, time horizon* $T \geq K$*, parameter* β;
Pull each arm in $[K]$ once;
for $t = K$ *to* $T-1$, **do**

$$\phi_{t+1} \in \arg\max_{a \in [K]} \left\{ \bar{Y}_a(t) + \sqrt{\frac{2\beta \log t}{N_a(t)}} \right\}$$

end

Theorem 4: If $\beta > \frac{1}{2}(d-c)^2 \max_{a\in[K]} \mathrm{HitT}(P_{\theta_a})^2$ then,

$$R_{\boldsymbol{\theta}}^{\boldsymbol{\phi}_{\beta-\mathrm{UCB}}}(T) \leq 8\beta \left(\sum_{b \neq a^*(\boldsymbol{\theta})} \frac{1}{\Delta_b(\boldsymbol{\theta})} \right) \log T + \frac{\gamma}{\gamma - 2} \sum_{b \neq a^*(\boldsymbol{\theta})} \Delta_b(\boldsymbol{\theta}),$$

where $\gamma = \frac{4\beta}{(d-c)^2 \max_{a\in[K]} \mathrm{HitT}(P_{\theta_a})^2} > 2$.

Proof: Fix $b \neq a^*(\boldsymbol{\theta})$, and observe that,

$$N_b(T) \leq 1 + \frac{8\beta}{\Delta_b(\boldsymbol{\theta})^2} \log T + \sum_{t=2}^{T-1} I_{\left\{\phi_{t+1}=b,\ N_b(t) \geq \frac{8\beta}{\Delta_b(\boldsymbol{\theta})^2} \log T\right\}}.$$

On the event $\left\{\phi_{t+1} = b,\ N_b(t) \geq \frac{8\beta}{\Delta_b(\boldsymbol{\theta})^2} \log T\right\}$, we have that, either $\bar{Y}_b(t) \geq \mu(\theta_b) + \sqrt{\frac{2\beta \log t}{N_b(t)}}$, or $\bar{Y}_{a^*(\boldsymbol{\theta})}(t) \leq \mu^*(\boldsymbol{\theta}) - \sqrt{\frac{2\beta \log t}{N_{a^*(\boldsymbol{\theta})}(t)}}$, since otherwise the β-UCB index of $a^*(\boldsymbol{\theta})$ is larger than the β-UCB index of b which contradicts the assumption that $\phi_{t+1} = b$.

In addition, using Corollary 1, we obtain,

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}_{\beta-\mathrm{UCB}}} \left(\bar{Y}_b(t) \geq \mu(\theta_b) + \sqrt{\frac{2\beta \log t}{N_b(t)}} \right) \\
&= \sum_{n=1}^{t} \mathbb{P}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}_{\beta-\mathrm{UCB}}} \left(\bar{Y}_b(t) \geq \mu(\theta_b) + \sqrt{\frac{2\beta \log t}{N_b(t)}}, N_b(t) = n \right) \\
&\leq \sum_{n=1}^{t} \mathbb{P}_{\theta_b} \left(\frac{1}{n} \sum_{k=1}^{n} Y_k^b \geq \mu(\theta_b) + \sqrt{\frac{2\beta \log t}{n}} \right) \\
&\leq \sum_{n=1}^{t} \frac{1}{t^\gamma} = \frac{1}{t^{\gamma-1}}.
\end{aligned}$$

Similarly we can see that,

$$\mathbb{P}_{\boldsymbol{\theta}}^{\boldsymbol{\phi}_{\beta-\mathrm{UCB}}} \left(\bar{Y}_{a^*(\boldsymbol{\theta})}(t) \leq \mu^*(\boldsymbol{\theta}) - \sqrt{\frac{2\beta \log t}{N_{a^*(\boldsymbol{\theta})}(t)}} \right) \leq \frac{1}{t^{\gamma-1}}.$$

The conclusion now follows by putting everything together and using the integral estimate,

$$\sum_{t=2}^{T-1} \frac{1}{t^{\gamma-1}} \leq \int_1^\infty \frac{1}{t^{\gamma-1}} dt = \frac{1}{\gamma - 2}.$$

■

ACKNOWLEDGMENT

This research was supported in part by the NSF grant CCF-1816861.

REFERENCES

[1] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *The Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[2] M. Jerrum, A. Sinclair, and E. Vigoda, "A polynomial-time approximation algorithm for the permanent of a matrix with non-negative entries," in *Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing*. ACM, New York, 2001, pp. 712–721.

[3] V. Moulos, "Optimal Best Markovian Arm Identification with Fixed Confidence," in *33rd Annual Conference on Neural Information Processing Systems*, 2019.

[4] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.*, vol. 58, pp. 13–30, 1963.

[5] D. Aldous and J. Fill, "Reversible markov chains and random walks on graphs," 2002.

[6] L. D. Davisson, G. Longo, and A. Sgarro, "The error exponent for the noiseless encoding of finite ergodic Markov sources," *IEEE Trans. Inform. Theory*, vol. 27, no. 4, pp. 431–438, 1981.

[7] S. Watanabe and M. Hayashi, "Finite-length analysis on tail probability for Markov chain and application to simple hypothesis testing," *Ann. Appl. Probab.*, vol. 27, no. 2, pp. 811–845, 2017.

[8] V. Moulos and V. Anantharam, "Optimal Chernoff and Hoeffding Bounds for Finite State Markov Chains," 2019.

[9] D. Gillman, "A Chernoff bound for random walks on expander graphs," in *34th Annual Symposium on Foundations of Computer Science (Palo Alto, CA, 1993)*. IEEE Comput. Soc. Press, Los Alamitos, CA, 1993, pp. 680–691.

[10] I. H. Dinwoodie, "A probability inequality for the occupation measure of a reversible Markov chain," *Ann. Appl. Probab.*, vol. 5, no. 1, pp. 37–43, 1995.

[11] N. Kahale, "Large deviation bounds for Markov chains," *Combin. Probab. Comput.*, vol. 6, no. 4, pp. 465–474, 1997.

[12] P. Lezaud, "Chernoff-type bound for finite Markov chains," *Ann. Appl. Probab.*, vol. 8, no. 3, pp. 849–867, 1998.

[13] C. A. León and F. Perron, "Optimal Hoeffding bounds for discrete reversible Markov chains," *Ann. Appl. Probab.*, vol. 14, no. 2, pp. 958–970, 2004.

[14] B. a. Miasojedow, "Hoeffding's inequalities for geometrically ergodic Markov chains on general state space," *Statist. Probab. Lett.*, vol. 87, pp. 115–120, 2014.

[15] D. Paulin, "Concentration inequalities for Markov chains by Marton couplings and spectral methods," *Electron. J. Probab.*, vol. 20, pp. no. 79, 32, 2015.

[16] S. Rao, "A Hoeffding inequality for Markov chains," *Electron. Commun. Probab.*, vol. 24, pp. Paper No. 14, 11, 2019.

[17] J. Fan, B. Jiang, and Q. Sun, "Hoeffding's lemma for Markov Chains and its applications to statistical learning," 2018.

[18] I. Kontoyiannis, L. A. Lastras-Montaño, and S. P. Meyn, "Exponential Bounds and Stopping Rules for MCMC and General Markov Chains," in *Proceedings of the 1st International Conference on Performance Evaluation Methodolgies and Tools*, ser. valuetools '06. New York, NY, USA: ACM, 2006.

[19] P. W. Glynn and D. Ormoneit, "Hoeffding's inequality for uniformly ergodic Markov chains," *Statist. Probab. Lett.*, vol. 56, no. 2, pp. 143–146, 2002.

[20] R. Douc, E. Moulines, J. Olsson, and R. van Handel, "Consistency of the maximum likelihood estimator for general hidden Markov models," *Ann. Statist.*, vol. 39, no. 1, pp. 474–513, 2011.

[21] K. Marton, "Bounding $\bar{d}$-distance by informational divergence: a method to prove measure concentration," *Ann. Probab.*, vol. 24, no. 2, pp. 857–866, 1996.

[22] ——, "A measure concentration inequality for contracting Markov chains," *Geom. Funct. Anal.*, vol. 6, no. 3, pp. 556–571, 1996.

[23] ——, "Measure concentration for a class of random processes," *Probab. Theory Related Fields*, vol. 110, no. 3, pp. 427–439, 1998.

[24] P.-M. Samson, "Concentration of measure inequalities for Markov chains and Φ-mixing processes," *Ann. Probab.*, vol. 28, no. 1, pp. 416–461, 2000.

[25] K. Marton, "Measure concentration and strong mixing," *Studia Sci. Math. Hungar.*, vol. 40, no. 1-2, pp. 95–113, 2003.

[26] J.-R. Chazottes, P. Collet, C. Külske, and F. Redig, "Concentration inequalities for random fields via coupling," *Probab. Theory Related Fields*, vol. 137, no. 1-2, pp. 201–225, 2007.

[27] L. Kontorovich and K. Ramanan, "Concentration inequalities for dependent random variables via the martingale method," *Ann. Probab.*, vol. 36, no. 6, pp. 2126–2158, 2008.

[28] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. in Appl. Math.*, vol. 6, no. 1, pp. 4–22, 1985.

[29] E. Even-Dar, S. Mannor, and Y. Mansour, "Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems," *J. Mach. Learn. Res.*, vol. 7, pp. 1079–1105, 2006.

[30] 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, May 2002.

[31] S. Bubeck and N. Cesa-Bianchi, "Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems," *Foundations and Trends in Machine Learning*, vol. 5, no. 1, pp. 1–122, 2012.

[32] T. Lattimore and C. Szepesvári, "Bandit Algorithms," 2019.

[33] A. Slivkins, "Introduction to Multi-Armed Bandits," *Foundations and Trends® in Machine Learning*, vol. 12, no. 1-2, pp. 1–286, 2019.

[34] L. Devroye and G. Lugosi, *Combinatorial methods in density estimation*, ser. Springer Series in Statistics. Springer-Verlag, New York, 2001.

[35] S. Mannor and J. N. Tsitsiklis, "The sample complexity of exploration in the multi-armed bandit problem," *J. Mach. Learn. Res.*, vol. 5, pp. 623–648, 2004.

[36] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays. II. Markovian rewards," *IEEE Trans. Automat. Control*, vol. 32, no. 11, pp. 977–982, 1987.

[37] C. Tekin and M. Liu, "Online algorithms for the multi-armed bandit problem with Markovian rewards," in *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sep. 2010, pp. 1675–1682.

[38] V. Moulos, "Finite-time Analysis of Kullback-Leibler Upper Confidence Bounds for Optimal Adaptive Allocation with Multiple Plays and Markovian Rewards," 2020.