Learning Multiple Markov Chains via Adaptive Allocation

Mohammad Sadegh Talebi, Odalric-Ambrym Maillard

Learning Multiple Markov Chains via Adaptive Allocation

Mohammad Sadegh Talebi and Odalric-Ambrym Maillard
Inria Lille – Nord Europe
{sadegh.talebi,odalric.maillard}@inria.fr

November 29, 2019

Abstract

We study the problem of learning the transition matrices of a set of Markov chains from a single stream of observations on each chain. We assume that the Markov chains are ergodic but otherwise unknown. The learner can sample Markov chains sequentially to observe their states. The goal of the learner is to sequentially select various chains to learn transition matrices uniformly well with respect to some loss function. We introduce a notion of loss that naturally extends the squared loss for learning distributions to the case of Markov chains, and further characterize the notion of being *uniformly good* in all problem instances. We present a novel learning algorithm that efficiently balances *exploration* and *exploitation* intrinsic to this problem, without any prior knowledge of the chains. We provide finite-sample PAC-type guarantees on the performance of the algorithm. Further, we show that our algorithm asymptotically attains an optimal loss.

1 Introduction

We study a *variant* of the following sequential adaptive allocation problem: A learner is given a set of K arms, where to each arm $k \in [K]$, an unknown real-valued distributions ν_k with mean μ_k and variance $\sigma_k^2 > 0$ is associated. At each round $t \in \mathbb{N}$, the learner must select an arm $k_t \in [K]$, and receives a sample drawn from ν_k. Given a total budget of n pulls, the objective is to estimate the expected values $(\mu_k)_{k\in[K]}$ of all distributions uniformly well. The quality of estimation in this problem is classically measured through *expected* quadratic estimation error, $\mathbb{E}[(\mu_k - \hat{\mu}_{k,n})^2]$ for the empirical mean estimate $\hat{\mu}_{k,n}$ built with the $T_{k,n} = \sum_{t=1}^{n} \mathbb{I}\{k = k_t\}$ many samples received from ν_k at time n, and the performance of an allocation strategy is the maximal error, $\max_{k\in[K]} \mathbb{E}[(\mu_k - \hat{\mu}_{k,n})^2]$. Using ideas from the Multi-Armed Bandit (MAB) literature, previous works (e.g., [1, 2]) have provided optimistic sampling strategies with near-optimal performance guarantees for this setup.

This generic adaptive allocation problem is related to several application problems arising in optimal experiment design [3, 4], active learning [5], or Monte-Carlo methods [6]; we refer to [1, 7, 2, 8] and references therein for further motivation. We extend this line of work to the case where each process is a *discrete Markov chain*, hence introducing the problem of *active bandit learning of Markov chains*. More precisely, we no longer assume that $(\nu_k)_k$ are real-valued distributions, but we study the case where each ν_k is a discrete Markov process over a state space $\mathcal{S} \subset \mathbb{N}$. The law of the observations $(X_{k,i})_{i\in\mathbb{N}}$ on arm (or chain) k is given by $\nu_k(X_{k,1}, \dots X_{k,n}) = p_k(X_{k,1}) \prod_{i=2}^{n} P_k(X_{k,i-1}, X_{k,i})$, where p_k denotes the initial distribution

of states, and P_k is the transition function of the Markov chain. The goal of the learner is to learn the *transition matrices* $(P_k)_{k\in[K]}$ uniformly well on the chains. Note that the chains are *not* controlled (we only decide which chain to advance, not the states it transits to).

Before discussing the challenges of the extension to Markov chains, let us give further comments on the performance measure considered in bandit allocation for real-valued distributions: Using the expected quadratic estimation error on each arm k makes sense since when $T_{k,n}, k \in [K]$ are deterministic, it coincides with $\sigma_k^2/T_{k,n}$, thus suggesting to pull the distributions proportionally to σ_k^2. However, for a learning strategy, $T_{k,n}$ typically depends on all past observations. The presented analyses in these series of works rely on Wald's second identity as the technical device, heavily relying on the use of a quadratic loss criterion, which prevents one from extending the approach therein to other distances. Another peculiarity arising in working with expectations is the order of "max" and "expectation" operators. While it makes more sense to control the *expected value of the maximum*, the works cited above look at *maximum of the expected value*, which is more in line with a pseudo-loss definition rather than the loss; actually in extensions of these works a pseudo-loss is considered instead of this performance measure. As we show, all of these difficulties can be avoided by resorting to a *high probability* setup. Hence, in this paper, we deviate from using an *expected* loss criterion, and rather use a high-probability control. We formally define our performance criterion in Section 2.3.

1.1 Related Work

On the one hand, our setup can be framed into the line of works on active bandit allocation, considered for the estimation of reward distributions in MABs as introduced in [1, 7], and further studied in [2, 9]. This has been extended to stratified sampling for Monte-Carlo methods in [10, 8], or to continuous mean functions in, e.g., [11]. On the other hand, our extension from real-valued distributions to Markov chains can be framed into the rich literature on Markov chain estimation; see, e.g., [12, 13, 14, 15, 16, 17]. This stream of works extends a wide range of results from the i.i.d. case to the Markov case. These include, for instance, the law of large numbers for (functions of) state values [17], the central limit theorem for Markov sequences [13] (see also [17, 18]), and Chernoff-type or Bernstein-type concentration inequalities for Markov sequences [19, 20]. Note that the majority of these results are available for *ergodic* Markov chains.

Another stream of research on Markov chains, which is more relevant to our work, investigates learning and estimation of the *transition matrix* (as opposed to its full law); see, e.g., [16, 15, 21, 22]. Among the recent studies falling in this category, [22] investigates learning of the transition matrix with respect to a loss function induced by f-divergences in a *minimax setup*, thus extending [23] to the case of Markov chains. [21] derives a PAC-type bound for learning the transition matrix of an ergodic Markov chain with respect to the *total variation* loss. It further provides a matching lower bound. Among the existing literature on learning Markov chains, to the best of our knowledge, [21] is the closest to ours. There are however two aspects distinguishing our work: Firstly, the challenge in our problem resides in dealing with *multiple* Markov chains, which is present neither in [21] nor in the other studies cited above. Secondly, our notion of loss does not coincide with that considered in [21], and hence, the lower bound of [21] does not apply to our case.

Among the results dealing with multiple chains, we may refer to learning in the Markovian bandits setup [24, 25, 26]. Most of these studies address the problem of reward maximization over a finite time horizon. We also mention that in a recent study, [27] introduces the so-called active exploration in Markov decision processes, where the transition kernel is *known*, and the goal is rather to learn the mean reward associated to various states. To the best of our knowledge, none of these works address the problem of learning the *transition matrix*. Last, as we target

high-probability performance bounds (as opposed to those holding in expectation), our approach is naturally linked to the Probably Approximately Correct (PAC) analysis. [28] provides one of the first PAC bounds for learning discrete distributions. Since then, the problem of learning discrete distributions has been well studied; see, e.g., [29, 30, 23] and references therein. We refer to [23] for a rather complete characterization of learning distribution in a *minimax* setting under a big class of smooth loss functions. We remark that except for very few studies (e.g., [29]), most of these results are provided for discrete distributions.

1.2 Overview and Contributions

Our contributions are the following: (i) For the problem of learning Markov chains, we consider a notion of loss function, which appropriately extends the loss function for learning distributions to the case of Markov chains. Our notion of loss is similar to that considered in [22] (we refer to Section 2.3 for a comparison between our notion and the one in [22]). In contrast to existing works on similar bandit allocation problems, our loss function avoids technical difficulties faced when extending the squared loss function to this setup. We further characterize the notion of a "uniformly good algorithm" under the considered loss function for ergodic chains; (ii) We present an *optimistic* algorithm, called **BA-MC**, for active learning of Markov chains, which is simple to implement and does not require any prior knowledge of the chains. To the best of our knowledge, this constitutes the first algorithm for active bandit allocation for learning Markov chains; (iii) We provide non-asymptotic PAC-type, and asymptotic bounds, on the loss incurred by **BA-MC**, indicating three regimes. In the first regime, which holds for any budget $n \geq 4K$, we present (in Theorem 1) a high-probability bound on the loss scaling as $\widetilde{\mathcal{O}}(\frac{KS^2}{n})$, where $\widetilde{\mathcal{O}}(\cdot)$ hides $\log(\log(n))$ factors. Here, K and S respectively denote the number of chains and the number of states in a given chain. This result holds for homogenous Markov chains. We then characterize a cut-off budget n_{cutoff} (in Theorem 2) so that when $n \geq n_{\text{cutoff}}$, the loss behaves as $\widetilde{\mathcal{O}}(\frac{\Lambda}{n} + \frac{C_0}{n^{3/2}})$, where $\Lambda = \sum_k \sum_{x,y} P_k(x,y)(1 - P_k(x,y))$ denotes the sum of variances of all states and all chains, and where P_k denotes the transition probability of chain k. This latter bound constitutes the second regime, in view of the fact that $\frac{\Lambda}{n}$ equals the asymptotically optimal loss (see Section 2.4 for more details). Thus, this bound indicates that the *pseudo-excess loss* incurred by the algorithm vanishes at a rate $C_0 n^{-3/2}$ (we refer to Section 4 for a more precise definition). Furthermore, we carefully characterize the constant C_0. In particular, we discuss that C_0 does not deteriorate with mixing times of the chains, which, we believe, is a strong feature of our algorithm. We also discuss how various properties of the chains, e.g., discrepancies between stationary distribution of various states a given chain, may impact the learning performance. Finally, we demonstrate a third regime, the asymptotic one, when the budget n grows large, in which we show (in Theorem 3) that the loss of **BA-MC** matches the asymptotically optimal loss $\frac{\Lambda}{n}$. All proofs are provided in the supplementary material.

Markov chains have been successfully used for modeling a broad range of practical problems, and their success makes the studied problem in this paper relevant in practice. There are practical applications in reinforcement learning (e.g., active exploration in MDPs [27]) and in rested Markov bandits (e.g., channel allocation in wireless communication systems where a given channel's state follows a Markov chain[1]), for which we believe our contributions could serve as a technical tool.

[1] For example, in the Gilbert-Elliott channels [31].

2 Preliminaries and Problem Statement

2.1 Preliminaries

Before describing our model, we recall some preliminaries on Markov chains; these are standard definitions and results, and can be found in, e.g., [32, 33]. Consider a Markov chain defined on a finite state space $\mathcal{S}$ with cardinality S. Let $\mathcal{P}_{\mathcal{S}}$ denote the collection of all row-stochastic matrices over $\mathcal{S}$. The Markov chain is specified by its transition matrix $P \in \mathcal{P}_{\mathcal{S}}$ and its initial distribution p: For all $x, y \in \mathcal{S}$, $P(x, y)$ denotes the probability of transition to y if the current state is x. In what follows, we may refer to a chain by just referring to its transition matrix.

We recall that a Markov chain P is *ergodic* if $P^m > 0$ (entry-wise) for some $m \in \mathbb{N}$. If P is ergodic, then it has a unique *stationary distribution* π satisfying $\pi = \pi P$. Moreover $\underline{\pi} := \min_{x \in \mathcal{S}} \pi(x) > 0$. A chain P is said to be *reversible* if its stationary distribution π satisfies *detailed balance equations*: For all $x, y \in \mathcal{S}$, $\pi(x)P(x, y) = \pi(y)P(y, x)$. Otherwise, P is called *non-reversible*. For a Markov chain P, the largest eigenvalue is $\lambda_1(P) = 1$ (with multiplicity one). In a *reversible* chain P, all eigenvalues belong to $(-1, 1]$. We define the *absolute spectral gap* of a reversible chain P as $\gamma(P) = 1 - \lambda_\star(P)$, where $\lambda_\star(P)$ denotes the second largest (in absolute value) eigenvalue of P. If P is reversible, the absolute spectral gap $\gamma(P)$ controls the convergence rate of the state distributions of the chain towards the stationary distribution π. If P is *non-reversible*, the convergence rate is determined by the *pseudo-spectral gap* as introduced in [20] as follows. Define $P^\star$ as: $P^\star(x, y) := \pi(y)P(y, x)/\pi(x)$ for all $x, y \in \mathcal{S}$. Then, the pseudo-spectral gap $\gamma_{\mathsf{ps}}(P)$ is defined as: $\gamma_{\mathsf{ps}}(P) := \max_{\ell \geq 1} \frac{\gamma((P^\star)^\ell P^\ell)}{\ell}$.

2.2 Model and Problem Statement

We are now ready to describe our model. We consider a learner interacting with a finite set of Markov chains indexed by $k \in [K] := \{1, 2, \ldots, K\}$. For ease of presentation, we assume that all Markov chains are defined on the same state space[2] $\mathcal{S}$ with cardinality S. The Markov chain k, or for short chain k, is specified by its transition matrix $P_k \in \mathcal{P}_{\mathcal{S}}$. In this work, we assume that all Markov chains are ergodic, which implies that any chain k admits a unique stationary distribution, which we denote by π_k. Moreover, the minimal element of π_k is bounded away from zero: $\underline{\pi}_k := \min_{x \in \mathcal{S}} \pi_k(x) > 0$. The initial distributions of the chains are assumed to be arbitrary. Further, we let $\gamma_k := \gamma(P_k)$ to denote the absolute spectral gap of chain k if k is reversible; otherwise, we define the pseudo-spectral gap of k by $\gamma_{\mathsf{ps},k} := \gamma_{\mathsf{ps}}(P_k)$.

A related quantity in our results is the *Gini index* of the various states. For a chain k, the *Gini index* for state $x \in \mathcal{S}$ is defined as

$$G_k(x) := \sum_{y \in \mathcal{S}} P_k(x, y)(1 - P_k(x, y)).$$

Note that $G_k(x) \leq 1 - \frac{1}{S}$. This upper bound is verified by the fact that the maximal value of $G_k(x)$ is achieved when $P_k(x, y) = \frac{1}{S}$ for all $y \in \mathcal{S}$ (in view of the concavity of $z \mapsto \sum_{x \in \mathcal{S}} z(x)(1 - z(x))$). In this work, we assume that for all k, $\sum_{x \in \mathcal{S}} G_k(x) > 0$.[3] Another related quantity in our results is the sum (over states) of inverse stationary distributions: For a

[2] Our algorithm and results are straightforwardly extended to the case where the Markov chains are defined on different state spaces.

[3] We remark that there exist chains with $\sum_x G_k(x) = 0$. In view of the definition of the Gini index, such chains are necessarily deterministic (or degenerate), namely their transition matrices belong to $\{0, 1\}^{S \times S}$. One example is a deterministic cycle with S nodes. We note that such chains may fail to satisfy irreducibility or aperiodicity.

chain k, we define $H_k := \sum_{x \in \mathcal{S}} \pi_k(x)^{-1}$. Note that $S^2 \leq H_k \leq S\underline{\pi}_k^{-1}$. The quantity H_k reflects the discrepancy between individual elements of π_k.

The online learning problem. The learner wishes to design a sequential allocation strategy to adaptively sample various Markov chains so that all transition matrices are learnt uniformly well. The game proceeds as follows: Initially all chains are assumed to be non-stationary with arbitrary initial distributions chosen by the environment. At each step $t \geq 1$, the learner samples a chain k_t, based on the past decisions and the observed states, and observes the state $X_{k_t,t}$. The state of k_t evolves according to P_{k_t}. The state of chains $k \neq k_t$ does not change: $X_{k,t} = X_{k,t-1}$ for all $k \neq k_t$.

We introduce the following notations: Let $T_{k,t}$ denote the number of times chain k is selected by the learner up to time t: $T_{k,t} := \sum_{t'=1}^{t} \mathbb{I}\{k_{t'} = k\}$, where $\mathbb{I}\{\cdot\}$ denotes the indicator function. Likewise, we let $T_{k,x,t}$ represent the number of observations of chain k, up to time t, when the chain was in state x: $T_{k,x,t} := \sum_{t'=1}^{t} \mathbb{I}\{k_{t'} = k, X_{k,t'} = x\}$. Further, we note that the learner only controls $T_{k,t}$ (or equivalently, $\sum_x T_{k,x,t}$), but not the number of visits to individual states. At each step t, the learner maintains empirical estimates of the stationary distributions, and estimates transition probabilities of various chains based on the observations gathered up to t. We define the empirical stationary distribution of chain k at time t as $\hat{\pi}_{k,t}(x) := T_{k,x,t}/T_{k,t}$ for all $x \in \mathcal{S}$. For chain k, we maintain the following *smoothed estimation* of transition probabilities:

$$\widehat{P}_{k,t}(x,y) := \frac{\alpha + \sum_{t'=2}^{t} \mathbb{I}\{X_{k,t'-1} = x, X_{k,t'} = y\}}{\alpha S + T_{k,x,t}}, \quad \forall x, y \in \mathcal{S}, \tag{1}$$

where α is a positive constant. In the literature, the case of $\alpha = \frac{1}{S}$ is usually referred to as the *Laplace-smoothed* estimator. The learner is given a budget of n samples, and her goal is to obtain an accurate estimation of transition matrices of the Markov chains. The accuracy of the estimation is determined by some notion of loss, which will be discussed later. The learner adaptively selects various chains so that the minimal loss is achieved.

2.3 Performance Measures

We are now ready to provide a precise definition of our notion of loss, which would serve as the performance measure of a given algorithm. Given $n \in \mathbb{N}$, we define the loss of an adaptive algorithm $\mathcal{A}$ as:

$$L_n(\mathcal{A}) := \max_{k \in [K]} L_{k,n}, \quad \text{with} \quad L_{k,n} := \sum_{x \in \mathcal{S}} \hat{\pi}_{k,n}(x) \|P_k(x,\cdot) - \widehat{P}_{k,n}(x,\cdot)\|_2^2 .$$

The use of the L_2-norm in the definition of loss is quite natural in the context of learning and estimation of distributions, as it is directly inspired by the quadratic estimation error used in active bandit allocation (e.g., [2]). Given a budget n, the loss $L_n(\mathcal{A})$ of an adaptive algorithm $\mathcal{A}$ is a random variable, due to the evolution of the various chains as well as the possible randomization in the algorithm. Here, we aim at controlling this random quantity in a *high probability* setup as follows: Let $\delta \in (0,1)$. For a given algorithm $\mathcal{A}$, we wish to find $\varepsilon := \varepsilon(n,\delta)$ such that

$$\mathbb{P}\left(L_n(\mathcal{A}) \geq \varepsilon\right) \leq \delta . \tag{2}$$

Remark 1 *We remark that the empirical stationary distribution $\hat{\pi}_{k,t}$ may differ from the stationary distribution associated to the smoothed estimator $\widehat{P}_{k,t}$ of the transition matrix. Our algorithm and results, however, do not rely on possible relations between $\hat{\pi}_{k,t}$ and $\widehat{P}_{k,t}$, though one could have*

used smoothed estimators for π_k. The motivation behind using empirical estimate $\hat{\pi}_{k,t}$ of π_k in L_n is that it naturally corresponds to the occupancy of various states according to a given sample path.

Comparison with other losses. We now turn our attention to the comparison between our loss function and some other possible notions. First, we compare ours to the loss function $L'_n(\mathcal{A}) = \max_k \sum_{x\in\mathcal{S}} \|P_k(x,\cdot) - \widehat{P}_{k,n}(x,\cdot)\|_2^2$. Such a notion of loss might look more natural or simpler, since the weights $\hat{\pi}_{k,n}(x)$ are replaced simply with 1 (equivalently, uniform weights). However, this means a strategy may incur a high loss for a part of the state space that is rarely visited, even though we have absolutely no control on the chain. For instance, in the extreme case when some states x are reachable with a very small probability, $T_{k,x,n}$ may be arbitrarily small thus resulting in a large loss L'_n for all algorithms, while it makes little sense to penalize an allocation strategy for these "virtual" states. Weighting the loss according to the empirical frequency $\hat{\pi}_{k,n}$ of visits avoids such a phenomenon, and is thus more meaningful.

In view of the above discussion, it is also tempting to replace the empirical state distribution $\hat{\pi}_{k,n}$ with its expectation π_k, namely to define a *pseudo-loss* function of the form $L''_n(\mathcal{A}) = \max_k \sum_x \pi_k(x)\|P_k(x,\cdot) - \widehat{P}_{k,n}(x,\cdot)\|_2^2$ (as studied in, e.g., [22] in a different setup). We recall that our aim is to derive performance guarantees on the algorithm's loss that hold with high probability (for $1-\delta$ portions of the sample paths of the algorithm for a given δ). To this end, L_n (which uses $\hat{\pi}_{k,n}$) is more natural and meaningful than L''_n as L_n penalizes the algorithm's performance by the relative visit counts of various states in a given sample path (through $\hat{\pi}_{k,n}$), and not by the expected value of these. This matters a lot in the small-budget regime, where $\hat{\pi}_{k,n}$ could differ significantly from π_k — Otherwise when n is large enough, $\hat{\pi}_{k,n}$ becomes well-concentrated around π_k with high probability. To clarify further, let us consider the small-budget regime, and some state x where $\pi_k(x)$ is not small. In the case of L_n, using $\hat{\pi}_{k,n}$ we penalize the performance by the mismatch between $\widehat{P}_{k,n}(x,\cdot)$ and $P_k(x,\cdot)$, weighted proportionally to the number of rounds the algorithm has actually visited x. In contrast, in the case of L''_n, weighting the mismatch proportionally to $\pi_k(x)$ does not seem reasonable since in a given sample path, the algorithm might not have visited x enough even though $\pi_k(x)$ is not small. We remark that our results in subsequent sections easily apply to the pseudo-loss L''_n, at the expense of an additive second-order term, which might depend on the mixing times.

Finally, we position the high-probability guarantee on L_n, in the sense of Eq. (2), against those holding in expectation. Prior studies on bandit allocation, such as [7, 2], whose objectives involve a max operator, consider expected squared distance. The presented analyses in these series of works rely on Wald's second identity as the technical device. This prevents one to extend the approach therein to other distances. Another peculiarity arising in working with expectations is the order of "max" and "expectation" operators. While it makes more sense to control the *expected value of the maximum*, the works cited above look at *maximum of the expected value*, which is more in line with a pseudo-loss definition rather than the loss. All of these difficulties can be avoided by resorting to a high probability setup (in the sense of Eq. (2).

Further intuition and example. We now provide an illustrative example to further clarify some of the above comments. Let us consider the following two-state Markov chain: $P = \begin{bmatrix} 1/2 & 1/2 \\ \varepsilon & 1-\varepsilon \end{bmatrix}$, where $\varepsilon \in (0,1)$. The stationary distribution of this Markov chain is $\pi = [\frac{\varepsilon}{2+\varepsilon}, \frac{2}{2+\varepsilon}]$. Let s_1 (resp. s_2) denote the state corresponding to the first (resp. second) row of the transition matrix. In view of π, when $\varepsilon \ll 1$, the chain tends to stay in s_2 (the lazy state) most of the time: Out of n observations, one gets on average only $n\pi(s_1) = n\varepsilon/(2+\varepsilon)$ observations of state s_1, which

means, for $\varepsilon \ll 1/n$, essentially no observation of state s_1. Hence, no algorithm can estimate the transitions from s_1 in such a setup, and all strategies would suffer a huge loss according to L'_n, no matter how samples are allocated to this chain. Thus, L'_n has limited interest in order to distinguish between good and base sampling strategies. On the other hand, using L_n enables to better distinguish between allocation strategies, since the weight given to s_1 would be essentially 0 in this case, thus focusing on the good estimation of s_2 (and other chains) only.

2.4 Static Allocation

In this subsection, we investigate the optimal loss asymptotically achievable by an oracle policy that is aware of some properties of the chains. To this aim, let us consider a non-adaptive strategy where sampling of various chains is deterministic. Therefore, $T_{k,n}, k = 1, \ldots, K$ are not random. The following lemma is a consequence of the central limit theorem:

Lemma 1 *We have for any chain k: $T_{k,n} L_{k,n} \to_{T_{k,n}\to\infty} \sum_x G_k(x)$.*

The proof of this lemma consists in two steps: First, we provide lower and upper bounds on $L_{k,n}$ in terms of the loss $\widetilde{L}_{k,n}$ incurred by the learner had she used an empirical estimator (corresponding to $\alpha = 0$ in (1)). Second, we show that by the central limit theorem, $T_{k,n}\widetilde{L}_{k,n} \to_{T_{k,n}\to\infty} \sum_x G_k(x)$.

Now, consider an oracle policy $\mathcal{A}_{\text{oracle}}$, who is aware of $\sum_{x\in\mathcal{S}} G_k(x)$ for various chains. In view of the above discussion, and taking into account the constraint $\sum_{k\in[K]} T_{k,n} = n$, it would be asymptotically optimal to allocate $T_{k,n} = \eta_k n$ samples to chain k, where

$$\eta_k := \frac{1}{\Lambda} \sum_{x\in\mathcal{S}} G_k(x)\,, \quad \text{with} \quad \Lambda := \sum_{k\in[K]} \sum_{x\in\mathcal{S}} G_k(x)\,.$$

The corresponding loss would satisfy: $nL_n(\mathcal{A}_{\text{oracle}}) \to_{n\to\infty} \Lambda$. We shall refer to the quantity $\frac{\Lambda}{n}$ as *the asymptotically optimal loss*, which is a *problem-dependent* quantity. The coefficients $\eta_k, k \in [K]$ characterize the discrepancy between the transition matrices of the various chains, and indicate that an algorithm needs to account for such discrepancy in order to achieve the asymptotically optimal loss. Having characterized the notion of asymptotically optimal loss, we are now ready to define the notion of *uniformly good algorithm*:

Definition 1 (Uniformly Good Algorithm) *An algorithm $\mathcal{A}$ is said to be* uniformly good *if, for any problem instance, it achieves the asymptotically optimal loss when n grows large; that is, $\lim_{n\to\infty} nL_n(\mathcal{A}) = \Lambda$ for all problem instances.*

3 The BA-MC Algorithm

In this section, we introduce an algorithm designed for adaptive bandit allocation of a set of Markov chains. It is designed based on the *optimistic principle*, as in MAB problems (e.g., [34, 35]), and relies on an index function. More precisely, at each time t, the algorithm maintains an index function $b_{k,t+1}$ for each chain k, which provides an upper confidence bound (UCB) on the loss incurred by k at t; more precisely, with high probability, $b_{k,t+1} \geq L_{k,t} := \sum_{x\in\mathcal{S}} \hat{\pi}_{k,t}(x) \|P_k(x,\cdot) - \widehat{P}_{k,t}(x,\cdot)\|_2^2$, where $\widehat{P}_{k,t}$ denotes the smoothed estimate of P_k with some $\alpha > 0$ (see Eq. (1)). Now, by sampling a chain $k_t \in \operatorname{argmax}_{k\in[K]} b_{k,t+1}$ at time t, we can balance exploration and exploitation by selecting more the chains with higher estimated losses or those with higher uncertainty in these estimates.

In order to specify the index function $b_{k,\cdot}$, let us choose $\alpha = \frac{1}{3S}$ (we motivate this choice of α later on), and for each state $x \in \mathcal{S}$, define the estimate of Gini coefficient at time t as $\widehat{G}_{k,t}(x) := \sum_{y \in \mathcal{S}} \widehat{P}_{k,t}(x,y)(1 - \widehat{P}_{k,t}(x,y))$. The index $b_{k,t+1}$ is then defined as

$$b_{k,t+1} = \frac{2\beta}{T_{k,t}} \sum_{x \in \mathcal{S}} \mathbb{I}\{T_{k,x,t} > 0\} \widehat{G}_{k,t}(x) + \frac{6.6\beta^{3/2}}{T_{k,t}} \sum_{x \in \mathcal{S}} \frac{T_{k,x,t}^{3/2}}{(T_{k,x,t} + \alpha S)^2} \sum_{y \in \mathcal{S}} \sqrt{\widehat{P}_{k,t}(I - \widehat{P}_{k,t})(x,y)} + \frac{28\beta^2 S}{T_{k,t}} \sum_{x \in \mathcal{S}} \frac{\mathbb{I}\{T_{k,x,t} > 0\}}{T_{k,x,t} + \alpha S},$$

where $\beta := \beta(n,\delta) := c \log\left(\left\lceil \frac{\log(n)}{\log(c)} \right\rceil \frac{6KS^2}{\delta} \right)$, with $c > 1$ being an arbitrary choice. In this paper, we choose $c = 1.1$.

We remark that the design of the index $b_{k,\cdot}$ above comes from the application of empirical Bernstein concentration for α-smoothed estimators (see Lemma 4 in the supplementary) to the loss function $L_{k,t}$. In other words, Lemma 4 guarantees that with high probability, $b_{k,t+1} \geq L_{k,t}$. Our concentration inequality (Lemma 4) is new, to our knowledge, and could be of independent interest.

Having defined the index function $b_{k,\cdot}$, we are now ready to describe our algorithm, which we call **BA-MC** (Bandit Allocation for Markov Chains). **BA-MC** receives as input a confidence parameter δ, a budget n, as well as the state space $\mathcal{S}$. It initially samples each chain twice (hence, this phase lasts for $2K$ rounds). Then, **BA-MC** simply consists in sampling the chain with the largest index $b_{k,t+1}$ at each round t. Finally, it returns, after n pulls, an estimate $\widehat{P}_{k,n}$ for each chain k. We provide the pseudo-code of **BA-MC** in Algorithm 1. Note that **BA-MC** does not require any prior knowledge of the chains (neither the initial distribution nor the mixing time).

Algorithm 1 **BA-MC** – Bandit Allocation for Markov Chains

Input: Confidence parameter δ, budget n, state space $\mathcal{S}$;
Initialize: Sample each chain twice;
for $t = 2K + 1, \ldots, n$ **do**
  Sample chain $k_t \in \operatorname{argmax}_k b_{k,t+1}$;
  Observe $X_{k,t}$, and update $T_{k,x,t}$ and $T_{k,t}$;
end for

In order to provide more insights into the design of **BA-MC**, let us remark that (as shown in Lemma 8 in the supplementary) $b_{k,t+1}$ provides a high-probability UCB on the quantity $\frac{1}{T_{k,t}} \sum_x G_k(x)$ as well. Now by sampling the chain $k_t \in \operatorname{argmax}_{k \in [K]} b_{k,t+1}$ at time t, in view of discussions in Section 2.4, **BA-MC** would try to mimic an oracle algorithm being aware of $\sum_x G_k(x)$ for various chains.

We remark that our concentration inequality in Lemma 4 (of the supplementary) parallels the one presented in Lemma 8.3 in [36]. In contrast, our concentration lemma makes appear the terms $T_{k,x,t} + \alpha S$ in the denominator, whereas Lemma 8.3 in [36] makes appear terms $T_{k,x,t}$ in the denominator. This feature plays an important role to deal with situations where some states are not sampled up to time t, that is for when $T_{k,x,t} = 0$ for some x.

4 Performance Bounds

We are now ready to study the performance bounds on the loss $L_n(\textbf{BA-MC})$ in both asymptotic and non-asymptotic regimes. We begin with a generic non-asymptotic bound as follows:

Theorem 1 (BA-MC, Generic Performance) *Let $\delta \in (0,1)$. Then, for any budget $n \geq 4K$, with probability at least $1-\delta$, the loss under $\mathcal{A} =$ **BA-MC** satisfies*

$$L_n(\mathcal{A}) \leq \frac{304KS^2\beta^2}{n} + \widetilde{\mathcal{O}}\Big(\frac{K^2S^2}{n^2}\Big)\,.$$

The proof of this theorem, provided in Section C in the supplementary, reveals the motivation to choose $\alpha = \frac{1}{3S}$: It verifies that to minimize the dependency of the loss on S, on must choose $\alpha \propto S^{-1}$. In particular, the proof does not rely on the ergodicity assumption:

Remark 2 *Theorem 1 is valid even if the Markov chains $P_k, k \in [K]$ are reducible or periodic.*

In the following theorem, we state another non-asymptotic bound on the performance of **BA-MC**, which refines Theorem 1. To present this result, we recall the notation $\Lambda := \sum_k \sum_x G_k(x)$, and that for any chain k, $\eta_k = \frac{1}{\Lambda}\sum_{x\in\mathcal{S}} G_k(x)$, $H_k := \sum_{x\in\mathcal{S}} \pi_k(x)^{-1}$, and $\underline{\pi}_k := \min_{x\in\mathcal{S}} \pi_k(x) > 0$.

Theorem 2 *Let $\delta \in (0,1)$, and assume that $n \geq n_{cutoff} := K \max_k \left(\frac{300}{\gamma_{\mathsf{ps},k}\underline{\pi}_k} \log\left(\frac{2K}{\delta}\sqrt{\underline{\pi}_k^{-1}}\right)\right)^2$. Then, with probability at least $1-2\delta$,*

$$L_n(\mathcal{A}) \leq \frac{2\beta\Lambda}{n} + \frac{C_0\beta^{3/2}}{n^{3/2}} + \widetilde{\mathcal{O}}(n^{-2})\,,$$

where $C_0 := 150K\sqrt{S\Lambda \max_k H_k} + 3\sqrt{S\Lambda}\max_k \frac{H_k}{\eta_k}$.

Recalling the asymptotic loss of the oracle algorithm discussed in Section 2.4 being equal to Λ/n, in view of the Bernstein concentration, the oracle would incur a loss at most $\frac{2\beta\Lambda}{n}$ for when the budget n is finite. In this regard, we may look at the quantity $L_n(\mathcal{A}) - \frac{2\beta\Lambda}{n}$ as the *pseudo-excess loss* of $\mathcal{A}$ (we refrain from calling this quantity the *excess loss*, as $\frac{2\beta\Lambda}{n}$ is not equal to the high-probability loss of the oracle). Theorem 2 implies that when n is greater than the cut-off budget n_{cutoff}, the pseudo-excess loss under **BA-MC** vanishes at a rate $\widetilde{\mathcal{O}}(n^{-3/2})$. In particular, Theorem 2 characterizes the constant C_0 controlling the main term of the pseudo-excess loss: $C_0 = \mathcal{O}(K\sqrt{S\Lambda\max_k H_k} + \sqrt{S\Lambda}\max_k \frac{H_k}{\eta_k})$. This further indicates that the pseudo-excess loss is controlled by the quantity $\frac{H_k}{\eta_k}$, which captures (i) the discrepancy among the $\sum_x G_k(x)$ values of various chains k, and (ii) the discrepancy between various stationary probabilities $\pi_k(x), x \in \mathcal{S}$. We emphasize that the dependency of the learning performance (through C_0) on H_k is in alignment with the result obtained by [21] for the estimation of a single ergodic Markov chain.

The proof of this theorem, provided in Section D in the supplementary, shows that to determine the cut-off budget n_{cutoff}, one needs to determine the value of n such that with high probability, for any chain k and state x, the term $T_{k,n}(T_{k,x,n} + \alpha S)^{-1}$ approaches $\pi_k(x)^{-1}$, which is further controlled by $\gamma_{\mathsf{ps},k}$ (or γ_k if k is reversible) as well as the minimal stationary distribution $\underline{\pi}_k$. This in turn allows us to show that, under **BA-MC**, the number $T_{k,n}$ of samples for any chain k comes close to the quantity $\eta_k n$. Finally, we remark that the proof of Theorem 2 also reveals that the result in the theorem is indeed valid for any constant $\alpha > 0$.

In the following theorem, we characterize the asymptotic performance of **BA-MC**:

Theorem 3 (BA-MC, Asymptotic Regime) *Under $\mathcal{A} =$ **BA-MC**, $\limsup_{n\to\infty} nL_n(\mathcal{A}) = \Lambda$.*

The above theorem asserts that, asymptotically, the loss under **BA-MC** matches the asymptotically optimal loss Λ/n characterized in Section 2.4. We may thus conclude that **BA-MC** is uniformly good (in the sense of Definition 1). The proof of Theorem 3 (provided in Section E of the supplementary) proceeds as follows: It divides the estimation problem into two consecutive sub-problems, the one with the budget $n_0 = \sqrt{n}$ and the other with the rest $n - \sqrt{n}$ of pulls. We then show when $n_0 = \sqrt{n} \geq n_{\text{cutoff}}$, the number of samples on each chain k at the end of the first sub-problem is lower bounded by $\Omega(n^{1/4})$, and as a consequence, the index b_k would be accurate enough: $b_{k,n_0} \in \frac{1}{T_{k,n_0}}(\sum_x G_k(x), \sum_x G_k(x) + \widetilde{\mathcal{O}}(n^{-1/8}))$ with high probability. This allows us to relate the allocation under **BA-MC** in the course of the second sub-problem to that of the oracle, and further to show that the difference vanishes as $n \to \infty$.

Below, we provide some further comments about the presented bounds in Theorems 1–3:

Various regimes. Theorem 1 provides a non-asymptotic bound on the loss valid for any n, while Theorem 3 establishes the optimality of **BA-MC** in the asymptotic regime. In view of the inequality $\Lambda \leq K(S-1)$, the bound in Theorem 1 is at least off by a factor of S from the asymptotic loss Λ/n. Theorem 2 bridges between the two results thereby establishing a third regime, in which the algorithm enjoys the asymptotically optimal loss up to an additive pseudo-excess loss scaling as $\widetilde{\mathcal{O}}(n^{-3/2})$.

The effect of mixing. It is worth emphasizing that the mixing times of the chains *do not* appear explicitly in the bounds, and only control (through the pseudo-spectral gap $\gamma_{\mathsf{ps},k}$) the cut-off budget n_{cutoff} that ensures when the pseudo-excess loss vanishes at a rate $n^{-3/2}$. This is indeed a strong aspect of our results due to our meaningful definition of loss, which could be attributed to the fact that our loss function employs empirical estimates $\hat{\pi}_{k,n}$ in lieu of π_k. Specifically speaking, as argued in [36], given the number of samples of various states (akin to using $\hat{\pi}_{k,t}(x)$ in the loss definition), the convergence of frequency estimates towards the true values is independent of the mixing time of the chain. We note that despite the dependence of n_{cutoff} on the mixing times, **BA-MC** does not need to estimate them as when $n \leq n_{\text{cutoff}}$, it still enjoys the loss guarantee of Theorem 1. We also mention that to define an index function for the loss function $\max_k \sum_x \pi_k(x) \|P_k(x,\cdot) - \widehat{P}_{k,n}(x,\cdot)\|_2^2$, one may have to derive confidence bounds on the mixing time and/or stationary distribution π_k as well.

More on the pseudo-excess loss. We stress that the notion of the pseudo-excess loss bears some similarity with the definition of regret for active bandit learning of distributions as introduced in [7, 2] (see Section 1). In the latter case, the regret typically decays as $n^{-3/2}$ similarly to the pseudo-excess loss in our case. An interesting question is whether the decay rate of the pseudo-excess loss, as a function of n, can be improved. And more importantly, if a (problem-dependent) lower bound on the pseudo-excess loss can be established. These questions are open even for the simpler case of active learning of distributions in the i.i.d. setup; see, e.g., [37, 8, 2]. We plan to address these as a future work.

5 Conclusion

In this paper, we addressed the problem of active bandit allocation in the case of discrete and ergodic Markov chains. We considered a notion of loss function appropriately extending the loss function for learning distributions to the case of Markov chains. We further characterized the notion of a "uniformly good algorithm" under the considered loss function. We presented

an algorithm for learning Markov chains, which we called **BA-MC**. Our algorithm is simple to implement and does not require any prior knowledge of the Markov chains. We provided non-asymptotic PAC-type bounds on the loss incurred by **BA-MC**, and showed that asymptotically, it incurs an optimal loss. We further discussed that the (pseudo-excess) loss incurred by **BA-MC** in our bounds does not deteriorate with mixing times of the chains. As a future work, we plan to derive a (problem-dependent) lower bound on the pseudo-excess loss. Another interesting, yet very challenging, future direction is to devise adaptive learning algorithms for restless Markov chains, where the state of various chains evolve at each round independently of the learner's decision.

Acknowledgements

This work has been supported by CPER Nord-Pas-de-Calais/FEDER DATA Advanced data science and technologies 2015-2020, the French Ministry of Higher Education and Research, Inria, and the French Agence Nationale de la Recherche (ANR), under grant ANR-16-CE40-0002 (project BADASS).

References

[1] András Antos, Varun Grover, and Csaba Szepesvári. Active learning in multi-armed bandits. In *International Conference on Algorithmic Learning Theory*, pages 287–302. Springer, 2008.

[2] Alexandra Carpentier, Alessandro Lazaric, Mohammad Ghavamzadeh, Rémi Munos, and Peter Auer. Upper-confidence-bound algorithms for active learning in multi-armed bandits. In *International Conference on Algorithmic Learning Theory*, pages 189–203. Springer, 2011.

[3] Valerii Vadimovich Fedorov. *Theory of optimal experiments*. Elsevier, 1972.

[4] Hovav A. Dror and David M. Steinberg. Sequential experimental designs for generalized linear models. *Journal of the American Statistical Association*, 103(481):288–298, 2008.

[5] David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.

[6] Pierre Etoré and Benjamin Jourdain. Adaptive optimal allocation in stratified sampling methods. *Methodology and Computing in Applied Probability*, 12(3):335–360, 2010.

[7] András Antos, Varun Grover, and Csaba Szepesvári. Active learning in heteroscedastic noise. *Theoretical Computer Science*, 411(29-30):2712–2728, 2010.

[8] Alexandra Carpentier, Remi Munos, and András Antos. Adaptive strategy for stratified Monte Carlo sampling. *Journal of Machine Learning Research*, 16:2231–2271, 2015.

[9] James Neufeld, András György, Dale Schuurmans, and Csaba Szepesvári. Adaptive Monte Carlo via bandit allocation. In *Proceedings of the 31st International Conference on International Conference on Machine Learning*, pages 1944–1952, 2014.

[10] Alexandra Carpentier and Rémi Munos. Adaptive stratified sampling for Monte-Carlo integration of differentiable functions. In *Advances in Neural Information Processing Systems*, pages 251–259, 2012.

[11] Alexandra Carpentier and Odalric-Ambrym Maillard. Online allocation and homogeneous partitioning for piecewise constant mean-approximation. In *Advances in Neural Information Processing Systems*, pages 1961–1969, 2012.

[12] Patrick Billingsley. Statistical methods in Markov chains. *The Annals of Mathematical Statistics*, pages 12–40, 1961.

[13] Claude Kipnis and S. R. Srinivasa Varadhan. Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Communications in Mathematical Physics*, 104(1):1–19, 1986.

[14] Moshe Haviv and Ludo Van der Heyden. Perturbation bounds for the stationary probabilities of a finite Markov chain. *Advances in Applied Probability*, 16(4):804–818, 1984.

[15] Nicky J. Welton and A. E. Ades. Estimation of Markov chain transition probabilities and rates from fully and partially observed data: uncertainty propagation, evidence synthesis, and model calibration. *Medical Decision Making*, 25(6):633–645, 2005.

[16] Bruce A. Craig and Peter P. Sendi. Estimation of the transition matrix of a discrete-time Markov chain. *Health Economics*, 11(1):33–42, 2002.

[17] Sean P. Meyn and Richard L. Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

[18] Emmanuel Rio. *Asymptotic theory of weakly dependent random processes*. Springer, 2017.

[19] Pascal Lezaud. Chernoff-type bound for finite Markov chains. *Annals of Applied Probability*, pages 849–867, 1998.

[20] Daniel Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*, 20, 2015.

[21] Geoffrey Wolfer and Aryeh Kontorovich. Minimax learning of ergodic Markov chains. In *Algorithmic Learning Theory*, pages 903–929, 2019.

[22] Yi HAO, Alon Orlitsky, and Venkatadheeraj Pichapati. On learning Markov chains. In *Advances in Neural Information Processing Systems*, pages 648–657, 2018.

[23] Sudeep Kamath, Alon Orlitsky, Dheeraj Pichapati, and Ananda Theertha Suresh. On learning distributions from their samples. In *Conference on Learning Theory*, pages 1066–1100, 2015.

[24] Ronald Ortner, Daniil Ryabko, Peter Auer, and Rémi Munos. Regret bounds for restless Markov bandits. *Theoretical Computer Science*, 558:62–76, 2014.

[25] Cem Tekin and Mingyan Liu. Online learning of rested and restless bandits. *IEEE Transactions on Information Theory*, 58(8):5588–5611, 2012.

[26] Christopher R. Dance and Tomi Silander. Optimal policies for observing time series and related restless bandit problems. *Journal of Machine Learning Research*, 20(35):1–93, 2019.

[27] Jean Tarbouriech and Alessandro Lazaric. Active exploration in Markov decision processes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 974–982, 2019.

[28] Michael Kearns, Yishay Mansour, Dana Ron, Ronitt Rubinfeld, Robert E. Schapire, and Linda Sellie. On the learnability of discrete distributions. In *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing*, pages 273–282. ACM, 1994.

[29] David Gamarnik. Extension of the PAC framework to finite and countable Markov chains. *IEEE Transactions on Information Theory*, 49(1):338–345, 2003.

[30] Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885, 2015.

[31] Mordechai Mushkin and Israel Bar-David. Capacity and coding for the Gilbert-Elliott channels. *IEEE Transactions on Information Theory*, 35(6):1277–1290, 1989.

[32] James R. Norris. *Markov chains*. Number 2. Cambridge University Press, 1998.

[33] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. Markov chains and mixing times. *American Mathematical Society, Providence, RI*, 2009.

[34] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

[35] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

[36] Daniel J. Hsu, Aryeh Kontorovich, and Csaba Szepesvári. Mixing time estimation in reversible Markov chains from a single sample path. In *Advances in Neural Information Processing Systems*, pages 1459–1467, 2015.

[37] Alexandra Carpentier and Rémi Munos. Minimax number of strata for online stratified sampling: The case of noisy samples. *Theoretical Computer Science*, 558:77–106, 2014.

[38] Odalric-Ambrym Maillard. Mathematics of statistical sequential decision making. *Habilitation à Diriger des Recherches*, 2019.

A Concentration Inequalities

Lemma 2 ([38, Lemma 2.4]) *Let $Z = (Z_t)_{t\in\mathbb{N}}$ be a sequence of random variables generated by a predictable process, and $\mathcal{F} = (\mathcal{F}_t)_t$ be its natural filtration. Let $\varphi : \mathbb{R} \to \mathbb{R}_+$ be a convex upper-envelope of the cumulant generating function of the conditional distributions with $\varphi(0) = 0$, and let $\varphi_\star$ denote its Legendre-Fenchel transform, that is:*

$$\begin{aligned} \forall \lambda \in \mathcal{D}, \forall t, \qquad & \log \mathbb{E}\left[\exp\left(\lambda Z_t\right) | \mathcal{F}_{t-1}\right] \leq \varphi(\lambda)\,, \\ \forall x \in \mathbb{R}, \qquad & \varphi_\star(x) = \sup_{\lambda\in\mathbb{R}} (\lambda x - \varphi(\lambda))\,, \end{aligned}$$

where $\mathcal{D} = \{\lambda \in \mathbb{R} : \forall t, \log \mathbb{E}\left[\exp(\lambda Z_t)|\mathcal{F}_{t-1}\right] \leq \varphi(\lambda) < \infty\}$. Assume that $\mathcal{D}$ contains an open neighborhood of 0. Let $\varphi_{\star,+}^{-1} : \mathbb{R} \to \mathbb{R}_+$ (resp. $\varphi_{\star,-}^{-1}$) be its reverse map on $\mathbb{R}_+$ (resp. $\mathbb{R}_-$), that is

$$\varphi_{\star,-}^{-1}(z) := \sup\{x \leq 0 : \varphi_\star(x) > z\} \quad \textit{and} \quad \varphi_{\star,+}^{-1}(z) := \inf\{x \geq 0 : \varphi_\star(x) > z\}\,.$$

Let N_n be an integer-valued random variable that is $\mathcal{F}$-measurable and almost surely bounded by n. Then, for all $c \in (1, n]$ and $\delta \in (0, 1)$,

$$\mathbb{P}\left[\frac{1}{N_n}\sum_{t=1}^{N_n} Z_t \geq \varphi_{\star,+}^{-1}\left(\frac{c}{N_n}\log\left(\left\lceil\frac{\log(n)}{\log(c)}\right\rceil\frac{1}{\delta}\right)\right)\right] \leq \delta\,,$$
$$\mathbb{P}\left[\frac{1}{N_n}\sum_{t=1}^{N_n} Z_t \leq \varphi_{\star,-}^{-1}\left(\frac{c}{N_n}\log\left(\left\lceil\frac{\log(n)}{\log(c)}\right\rceil\frac{1}{\delta}\right)\right)\right] \leq \delta\,.$$

Moreover, if N is a possibly unbounded $\mathbb{N}$-valued random variable that is $\mathcal{F}$-measurable, then for all $c > 1$ and $\delta \in (0, 1)$,

$$\mathbb{P}\left[\frac{1}{N}\sum_{t=1}^{N} Z_t \geq \varphi_{\star,+}^{-1}\left(\frac{c}{N}\log\left[\frac{\log(N)\log(cN)}{\delta\log^2(c)}\right]\right)\right] \leq \delta\,,$$
$$\mathbb{P}\left[\frac{1}{N}\sum_{t=1}^{N} Z_t \leq \varphi_{\star,-}^{-1}\left(\frac{c}{N}\log\left[\frac{\log(N)\log(cN)}{\delta\log^2(c)}\right]\right)\right] \leq \delta\,.$$

We provide an immediate consequence of this lemma for the case of *sub-Gamma* random variables:

Corollary 1 *Let $Z = (Z_t)_{t\in\mathbb{N}}$ be a sequence of random variables generated by a predictable process, and $\mathcal{F} = (\mathcal{F}_t)_t$ be its natural filtration. Assume for all $t \in \mathbb{N}$, $|Z_t| \leq b$ and $\mathbb{E}[Z_s^2|\mathcal{F}_{s-1}] \leq v$ for some positive numbers v and b. Let N_n be an integer-valued random variable that is $\mathcal{F}$-measurable and almost surely bounded by n. Then, for all $c \in (1, n]$ and $\delta \in (0, 1)$,*

$$\mathbb{P}\left[\frac{1}{N_n}\sum_{t=1}^{N_n} Z_t \geq \sqrt{\frac{2\zeta(n,\delta)v}{N_n}} + \frac{\zeta(n,\delta)b}{3N_n}\right] \leq \delta\,,$$
$$\mathbb{P}\left[\frac{1}{N_n}\sum_{t=1}^{N_n} Z_t \leq -\sqrt{\frac{2\zeta(n,\delta)v}{N_n}} - \frac{\zeta(n,\delta)b}{3N_n}\right] \leq \delta\,,$$

where $\zeta(n, \delta) := c\log\left(\left\lceil\frac{\log(n)}{\log(c)}\right\rceil\frac{1}{\delta}\right)$, with $c > 1$ being an arbitrary parameter.

Proof. The proof follows by an application of Lemma 2 for sub-Gamma random variables with parameters (v, b). Note that sub-Gamma random variables satisfy $\varphi(\lambda) \leq \frac{\lambda^2 v}{2(1-b\lambda)}$, for all $\lambda \in (0, 1/b)$, so that

$$\varphi_{\star,+}^{-1}(z) = \sqrt{2vz} + bz \quad \text{and} \quad \varphi_{\star,-}^{-1}(z) = -\sqrt{2vz} - bz\,.$$

Plugging these into the first statements of Lemma 2 completes the proof. □

As a consequence of this corollary, we present the following lemma:

Lemma 3 (Bernstein-Markov Concentration) *Let $(X_t)_{1\leq t\leq n}$ be generated from an ergodic Markov chain defined on a finite state-space $\mathcal{S}$ with transition matrix P. Consider the smoothed estimator $\widehat{P}_n$ of P defined as follows: For all $(x, y) \in \mathcal{S}^2$,*

$$\widehat{P}_n(x, y) := \frac{\alpha + \sum_{t=2}^{n}\mathbb{I}\{X_{t-1} = x, X_t = y\}}{\alpha S + T_{x,n}}\,,$$

with $\alpha > 0$. Then, for any $\delta \in (0,1)$, it holds that with probability at least $1-\delta$, for all $(x,y) \in \mathcal{S}^2$,

$$|\widehat{P}_n(x,y) - P(x,y)| \leq \sqrt{\left(\frac{T_{x,n}}{T_{x,n}+\alpha S}\right) \frac{2P(I-P)(x,y)\zeta(n,\delta)}{T_{x,n}+\alpha S}} + \frac{\frac{1}{3}\zeta(n,\delta) + \alpha|1-SP(x,y)|}{T_{x,n}+\alpha S},$$

where $\zeta(n,\delta) := c\log\left(\left\lceil \frac{\log(n)}{\log(c)} \right\rceil \frac{2S^2}{\delta}\right)$, with $c > 1$ being an arbitrary parameter.

Proof. The proof uses similar steps as in the one of Lemma 8.3 in [36]. Consider a pair $(x,y) \in \mathcal{S}^2$. We have

$$\begin{aligned}\widehat{P}_n(x,y) - P(x,y) &= \frac{\alpha + \sum_{t=2}^{n} \mathbb{I}\{X_{t-1}=x, X_t=y\}}{\alpha S + T_{x,n}} - P(x,y) \\ &= \frac{T_{x,n}}{T_{x,n}+\alpha S} Y_n + \frac{\alpha(1-SP(x,y))}{T_{x,n}+\alpha S},\end{aligned}$$

where $Y_n := \frac{1}{T_{x,n}} \left(\sum_{t=2}^{n} \mathbb{I}\{X_{t-1}=x, X_t=y\} - T_{x,n}P(x,y)\right)$. Hence,

$$|\widehat{P}_n(x,y) - P(x,y)| \leq \frac{T_{x,n}}{T_{x,n}+\alpha S}|Y_n| + \frac{\alpha|1-SP(x,y)|}{T_{x,n}+\alpha S}. \tag{3}$$

To control Y_n, we define the sequence $(Z_t)_{1\leq t\leq n}$, with $Z_1 := 0$, and

$$Z_t := \mathbb{I}\{X_{t-1}=x\}(\mathbb{I}\{X_t=y\} - P(x,y)), \quad \forall t \geq 2.$$

Note that for all t, $Z_t \in [-P(x,y), 1-P(x,y)]$ almost surely. Moreover, denoting by $(\mathcal{F}_t)_t$ the filtration generated by $(X_t)_{1\leq t\leq n}$, we observe that $(Z_t)_{1\leq t\leq n}$ is $\mathcal{F}_{t-1}$-measurable and $\mathbb{E}[Z_t|\mathcal{F}_{t-1}] = 0$. Hence, it is a martingale difference sequence with respect to $(\mathcal{F}_t)_t$, and it satisfies $Z_t \in [-P(x,y), 1-P(x,y)]$ for all t, and

$$\mathbb{E}[Z_t^2|\mathcal{F}_{t-1}] = P(x,y)(1-P(x,y))\mathbb{I}\{X_{t-1}=x\}, \quad \forall t \geq 2.$$

Applying Corollary 1 yields

$$|Y_n| \leq \sqrt{\frac{2P(I-P)(x,y)\zeta(n,\delta)}{T_{x,n}}} + \frac{\zeta(n,\delta)}{3T_{x,n}},$$

with probability at least $1-\delta$. Plugging the above bound into (3) gives the announced result. □

Lemma 4 (Empirical Bernstein-Markov Concentration) *Let $(X_t)_{1\leq t\leq n}$ be generated from an ergodic Markov chain defined on $\mathcal{S}$ with transition matrix P. Consider the smoothed estimator $\widehat{P}_n$ of P as defined in Lemma 3. Then, with probability at least $1-\delta$, for all $(x,y) \in \mathcal{S}^2$,*

$$|\widehat{P}_n(x,y) - P(x,y)| \leq \left(\frac{2T_{x,n}\widehat{P}_n(I-\widehat{P}_n)(x,y)\zeta}{(T_{x,n}+\alpha S)^2} + c_1\frac{\sqrt{T_{x,n}\widehat{P}_n(I-\widehat{P}_n)(x,y)}}{(T_{x,n}+\alpha S)^2} + \frac{c_2}{(T_{x,n}+\alpha S)^2}\right)^{1/2},$$

where $\zeta := \zeta(n,\delta) := c\log\left(\left\lceil \frac{\log(n)}{\log(c)} \right\rceil \frac{2S^2}{\delta}\right)$, where $c > 1$ is an arbitrary parameter, $\zeta' := \frac{1}{3}\zeta + \alpha(S-1)$, and

$$c_1 = \sqrt{8\zeta}(2\zeta+\zeta') \quad \textit{and} \quad c_2 := \zeta'^2 + 4\zeta(4\zeta+\zeta'+2\sqrt{\zeta\zeta'}) + \zeta'\sqrt{8\zeta}(5.3\sqrt{\zeta}+\sqrt{2\zeta'}).$$

Proof. Fix a pair $(x, y) \in \mathcal{S}^2$. Recall from Lemma 3 that with probability at least $1 - \delta$,

$$|\widehat{P}_n(x,y) - P(x,y)| \le \sqrt{\frac{2\zeta T_{x,n} P(I-P)(x,y)}{(T_{x,n}+\alpha S)^2}} + \frac{\zeta'}{T_{x,n}+\alpha S}\,,$$

so that

$$(\widehat{P}_n(x,y) - P(x,y))^2 \le \frac{2\zeta T_{x,n} P(I-P)(x,y)}{(T_{x,n}+\alpha S)^2} + \frac{\zeta'^2}{(T_{x,n}+\alpha S)^2} + \sqrt{\frac{8\zeta T_{x,n} P(I-P)(x,y)}{(T_{x,n}+\alpha S)^2}}\frac{\zeta'}{T_{x,n}+\alpha S}\,. \tag{4}$$

Next we derive an upper bound on $P(I-P)(x,y)$. By Taylor's expansion, we have

$$\begin{aligned}
P(I-P)(x,y) &= \widehat{P}_n(I-\widehat{P}_n)(x,y) + (I-2\widehat{P}_n)(P-\widehat{P}_n)(x,y) - (P-\widehat{P}_n)(x,y)^2 \\
&= \widehat{P}_n(I-\widehat{P}_n)(x,y) + (I-\widehat{P}_n-P)(P-\widehat{P}_n)(x,y) \\
&\le \widehat{P}_n(I-\widehat{P}_n)(x,y) + |(I-\widehat{P}_n-P)(x,y)|\left(\sqrt{\frac{2\zeta T_{x,n} P(I-P)(x,y)}{(T_{x,n}+\alpha S)^2}} + \frac{\zeta'}{T_{x,n}+\alpha S}\right) \\
&\le \widehat{P}_n(I-\widehat{P}_n)(x,y) + \sqrt{\frac{8\zeta T_{x,n} P(I-P)(x,y)}{(T_{x,n}+\alpha S)^2}} + \frac{2\zeta'}{T_{x,n}+\alpha S}\,.
\end{aligned}$$

Using the fact that $a \le b\sqrt{a} + c$ implies $a \le b^2 + b\sqrt{c} + c$ for nonnegative numbers a, b, c, we get

$$\begin{aligned}
P(I-P)(x,y) &\le \widehat{P}_n(I-\widehat{P}_n)(x,y) + \frac{2\zeta'}{T_{x,n}+\alpha S} + \sqrt{\frac{8\zeta T_{x,n}}{(T_{x,n}+\alpha S)^2}\left(\widehat{P}_n(I-\widehat{P}_n)(x,y) + \frac{2\zeta'}{T_{x,n}+\alpha S}\right)} \\
&\quad + \frac{8\zeta T_{x,n}}{(T_{x,n}+\alpha S)^2} \\
&\le \widehat{P}_n(I-\widehat{P}_n)(x,y) + \sqrt{\frac{8\zeta T_{x,n}}{(T_{x,n}+\alpha S)^2}\widehat{P}_n(I-\widehat{P}_n)(x,y)} + \frac{8\zeta + 2\zeta' + 4\sqrt{\zeta\zeta'}}{T_{x,n}+\alpha S}\,,
\end{aligned} \tag{5}$$

where we have used $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ valid for all $a, b \ge 0$. Taking square-root from both sides and using the fact $\sqrt{a+b} \le \sqrt{a} + \frac{b}{2\sqrt{a}}$ valid for all $a, b > 0$ give

$$\begin{aligned}
\sqrt{P(I-P)(x,y)} &\le \sqrt{\widehat{P}_n(I-\widehat{P}_n)(x,y)} + \frac{1}{\sqrt{T_{x,n}+\alpha S}}\left(\sqrt{2\zeta} + \sqrt{8\zeta + 2\zeta' + 4\sqrt{\zeta\zeta'}}\right) \\
&\le \sqrt{\widehat{P}_n(I-\widehat{P}_n)(x,y)} + \frac{5.3\sqrt{\zeta} + \sqrt{2\zeta'}}{\sqrt{T_{x,n}+\alpha S}}\,,
\end{aligned} \tag{6}$$

where we have used

$$\sqrt{2\zeta} + \sqrt{8\zeta + 2\zeta' + 4\sqrt{\zeta\zeta'}} \le \sqrt{2\zeta} + \sqrt{6\zeta + 2(\sqrt{\zeta} + \sqrt{\zeta'})^2} \le 5.3\sqrt{\zeta} + \sqrt{2\zeta'}\,.$$

Plugging (5) and (6) into (4), we obtain

$$
\begin{aligned}
&(\widehat{P}_n(x,y) - P(x,y))^2 \\
&\quad \le \frac{2\zeta T_{x,n}}{(T_{x,n}+\alpha S)^2}\left(\widehat{P}_n(I-\widehat{P}_n)(x,y) + \sqrt{\frac{8\zeta T_{x,n}}{(T_{x,n}+\alpha S)^2}\widehat{P}_n(I-\widehat{P}_n)(x,y)} + \frac{8\zeta + 2\zeta' + 4\sqrt{\zeta\zeta'}}{T_{x,n}+\alpha S}\right) \\
&\quad + \frac{\zeta'}{T_{x,n}+\alpha S}\sqrt{\frac{8\zeta T_{x,n}}{(T_{x,n}+\alpha S)^2}}\left(\sqrt{\widehat{P}_n(I-\widehat{P}_n)(x,y)} + \frac{5.3\sqrt{\zeta}+\sqrt{2\zeta'}}{\sqrt{T_{x,n}+\alpha S}}\right) + \frac{\zeta'^2}{(T_{x,n}+\alpha S)^2} \\
&\quad \le \frac{2\zeta T_{x,n}\widehat{P}_n(I-\widehat{P}_n)(x,y)}{(T_{x,n}+\alpha S)^2} + c_1\frac{\sqrt{T_{x,n}\widehat{P}_n(I-\widehat{P}_n)(x,y)}}{(T_{x,n}+\alpha S)^2} + \frac{c_2}{(T_{x,n}+\alpha S)^2}\,,
\end{aligned}
$$

with

$$
c_1 := \sqrt{8\zeta}(2\zeta+\zeta') \quad \text{and} \quad c_2 := \zeta'^2 + 4\zeta(4\zeta+\zeta'+2\sqrt{\zeta\zeta'}) + \zeta'\sqrt{8\zeta}(5.3\sqrt{\zeta}+\sqrt{2\zeta'})\,,
$$

which after taking the square-root from both sides yields the announced result. □

Next we recall a result for the convergence of empirical stationary distributions in a Markov chain to its stationary distribution:

Lemma 5 ([20]) *Let $(X_t)_{1\le t\le n}$ be an ergodic and reversible Markov chain defined on $\mathcal{S}$ with stationary distribution π and spectral gap γ. Let $\hat{\pi}_n$ denote the corresponding empirical stationary distribution of the Markov chain. For any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$
|\hat{\pi}_n(x) - \pi(x)| \le \sqrt{\frac{8\pi(x)(1-\pi(x))}{\gamma n}\log\left(\frac{1}{\delta}\sqrt{\frac{2}{\min_x \pi(x)}}\right)} + \frac{20}{\gamma n}\log\left(\frac{1}{\delta}\sqrt{\frac{2}{\min_x \pi(x)}}\right), \qquad \forall x \in \mathcal{S}.
$$

B Technical Lemmas

Before providing the proofs of the main theorems, we provide some technical lemmas. We begin with the following definition:

Definition 2 (Definition of the Event C) *Let $n \ge 1$ and $\delta > 0$. For any $(x,y) \in \mathcal{S}^2$ and $k \in [K]$ define*

$$
C_{x,y,k}(n,\delta) := \left\{\forall t \le n : |(\widehat{P}_{k,t} - P_k)(x,y)| \le \sqrt{\frac{2T_{k,x,t}P_k(I-P_k)(x,y)\beta(n,\delta)}{(T_{k,x,t}+\alpha S)^2}} + \frac{\beta(n,\delta)}{3(T_{k,x,t}+\alpha S)}\right\},
$$

where $\beta(n,\delta) := c\log\left(\left\lceil\frac{\log(n)}{\log(c)}\right\rceil\frac{6KS^2}{\delta}\right)$. Define

$$
C := C(n,\delta) := \cap_{k\in[K]}\cap_{x,y\in\mathcal{S}} C_{x,y,k}(n,\delta)\,.
$$

Lemma 6 *For any $n \ge 1$ and $\delta \in (0,1)$, it holds that $\mathbb{P}(C(n,\delta)) \ge 1-\delta$.*

Proof. Let $n \ge 1$ and $\delta > 0$. Define $\zeta(n,\delta) = c\log\left(\left\lceil\frac{\log(n)}{\log(c)}\right\rceil\frac{2KS^2}{\delta}\right)$, and note that $\beta(n,\delta) = \zeta(n,\delta) + c\log(3)$. Applying Lemma 3, we obtain

$$
|(\widehat{P}_{k,t} - P_k)(x,y)| \le \sqrt{\frac{2T_{k,x,t}P_k(I-P_k)(x,y)\zeta(n,\delta)}{(T_{k,x,t}+\alpha S)^2}} + \frac{\frac{1}{3}\zeta(n,\delta) + \alpha(S-1)}{T_{k,t,x}+\alpha S}\,,
$$

for all (x, y), and uniformly for all $t \leq n$, with probability at least $1 - \frac{\delta}{K}$. With the choice of $\alpha = \frac{1}{3S}$, and noting that $\beta(n, \delta) \geq \zeta(n, \delta)$ and

$$\frac{\zeta(n,\delta)}{3} + \frac{S-1}{3S} \leq \frac{\beta(n,\delta)}{3},$$

we obtain $\mathbb{P}(\cap_{x,y\in\mathcal{S}} C_{x,y,k}(n, \delta)) \geq 1 - \delta/K$ for all k. Finally, using a union bound gives $\mathbb{P}(C(n, \delta)) \geq 1 - \delta$. □

In the following lemma we provide an upper bound on the loss $L_{k,n}$, which is valid for all n.

Lemma 7 (Upper Bound on the Loss) *Assume that the event C holds. Then, for any budget $n > 1$ and chain k,*

$$L_{k,n} \leq \frac{2\beta}{T_{k,n}} \sum_x G_k(x)\mathbb{I}\{T_{k,x,n} > 0\} + \frac{2\sqrt{2}}{3}\frac{\beta^{3/2}\sqrt{S}}{T_{k,n}} \sum_x \frac{T_{k,x,n}^{3/2}\sqrt{G_k(x)}}{(T_{k,x,n} + \alpha S)^2} + \frac{S\beta^2}{9T_{k,n}} \sum_x \frac{T_{k,x,n}}{(T_{k,x,n} + \alpha S)^2}\,.$$

Proof. Let $n > 1$ and consider a chain k. To simplify the notation, we omit the dependence of various quantities on k (hence $T_{x,n} := T_{k,x,n}$, $T_n := T_{k,n}$, and so on). On the event C, we have

$$|\widehat{P}_n(x, y) - P(x, y)| \leq \sqrt{\left(\frac{T_{x,n}}{T_{x,n} + \alpha S}\right) \frac{2P(I - P)(x, y)\beta}{T_{x,n} + \alpha S}} + \frac{\beta}{3(T_{x,n} + \alpha S)}\,,$$

so that

$$(\widehat{P}_n(x, y) - P(x, y))^2 \leq \frac{2\beta P(I - P)(x, y)}{T_{x,n} + \alpha S} + \frac{\frac{1}{9}\beta^2}{(T_{x,n} + \alpha S)^2} + \frac{\frac{2\sqrt{2}}{3}\beta^{3/2}}{(T_{x,n} + \alpha S)^2}\sqrt{T_{x,n}P(I - P)(x, y)}\,.$$

Hence, we obtain the announced upper bound on the loss:

$$\begin{aligned}
L_n &= \sum_x \hat{\pi}_n(x) \sum_y (\widehat{P}_n(x, y) - P(x, y))^2 \\
&\leq \frac{2\beta}{T_n} \sum_x \frac{T_{x,n}G(x)}{T_{x,n} + \alpha S} + \frac{2\sqrt{2}}{3}\frac{\beta^{3/2}}{T_n} \sum_x \frac{T_{x,n}^{3/2}}{(T_{x,n} + \alpha S)^2} \sum_y \sqrt{P(I - P)(x, y)} + \frac{S\beta^2}{9T_n} \sum_x \frac{T_{x,n}}{(T_{x,n} + \alpha S)^2} \\
&\leq \frac{2\beta}{T_n} \sum_x G(x)\mathbb{I}\{T_{x,n} > 0\} + \frac{2\sqrt{2}}{3}\frac{\beta^{3/2}\sqrt{S}}{T_n} \sum_x \frac{T_{x,n}^{3/2}\sqrt{G(x)}}{(T_{x,n} + \alpha S)^2} + \frac{S\beta^2}{9T_n} \sum_x \frac{T_{x,n}}{(T_{x,n} + \alpha S)^2}\,,
\end{aligned}$$

where the last step follows from the Cauchy-Schwarz inequality. □

The following lemma presents bounds on the index $b_{k,\cdot}$ on the event C (defined in Definition 2):

Lemma 8 (Bounds on the Index) *Consider a chain k, and assume that the event C holds. Then, for any time t,*

$$b_{k,t+1} \leq \frac{2\beta}{T_{k,t}} \sum_x G_k(x)\mathbb{I}\{T_{k,x,t} > 0\} + \frac{13\beta^{3/2}\sqrt{S}}{T_{k,t}} \sum_x \sqrt{\frac{G_k(x)\mathbb{I}\{T_{k,x,t} > 0\}}{T_{k,x,t} + \alpha S}} + \frac{39\beta^2 S}{T_{k,t}} \sum_x \frac{\mathbb{I}\{T_{k,x,t} > 0\}}{T_{k,x,t} + \alpha S}\,,$$

$$b_{k,t+1} \geq \frac{2\beta}{T_{k,t}} \sum_x G_k(x)\mathbb{I}\{T_{k,x,t} > 0\}\,.$$

Proof. Fix a chain k and time t. To ease notation, let us omit the dependence of various quantities on k throughout. We first recall the definition of the index b_{t+1}:

$$b_{t+1} = \frac{2\beta}{T_t} \sum_x \widehat{G}_t(x) \mathbb{I}\{T_{x,t} > 0\} + \frac{c_1}{T_t} \sum_x \frac{T_{x,t}^{3/2}}{(T_{x,t} + \alpha S)^2} \sum_y \sqrt{\widehat{P}_t(I - \widehat{P}_t)(x,y)} + \frac{c_2 S}{T_t} \sum_x \frac{\mathbb{I}\{T_{x,t} > 0\}}{T_{x,t} + \alpha S},$$

where $c_1 = 6.6\beta^{3/2}$ and $c_2 = 28\beta^2$.

To derive an upper bound on b_t, we first find an upper bound on $\widehat{P}_t(I - \widehat{P}_t)(x,y)$ as follows. First, using Taylor's expansion, we have

$$\begin{aligned}
\widehat{P}_t(I - \widehat{P}_t)(x,y) &= P(I-P)(x,y) + (I - 2P)(\widehat{P}_t - P)(x,y) - (\widehat{P}_t - P)(x,y)^2 \\
&= P(I-P)(x,y) + (I - P - \widehat{P}_t)(\widehat{P}_t - P)(x,y) \\
&\le P(I-P)(x,y) + \sqrt{\frac{8\beta P(I-P)(x,y)}{T_{x,t} + \alpha S}} + \frac{2\beta}{3(T_{x,t} + \alpha S)} \qquad (7) \\
&\le \left(\sqrt{P(I-P)(x,y)} + \sqrt{\frac{2\beta}{T_{x,t} + \alpha S}} \right)^2,
\end{aligned}$$

where (7) follows from the definition of C. Hence,

$$\sqrt{\widehat{P}_t(I - \widehat{P}_t)(x,y)} \le \sqrt{P(I-P)(x,y)} + \sqrt{\frac{2\beta}{T_{x,t} + \alpha S}}. \qquad (8)$$

Using (7) and (8), we obtain the following upper bound on b_t, on the event C:

$$\begin{aligned}
b_{t+1} &\le \frac{2\beta}{T_t} \sum_x \sum_y \left(P(I-P)(x,y) + \sqrt{\frac{8\beta P(I-P)(x,y)}{T_{x,t} + \alpha S}} + \frac{2\beta}{3(T_{x,t} + \alpha S)} \right) \mathbb{I}\{T_{x,t} > 0\} \\
&\quad + \frac{6.6\beta^{3/2}}{T_t} \sum_x \frac{T_{x,t}^{3/2}}{(T_{x,t} + \alpha S)^2} \sum_y \left(\sqrt{P(I-P)(x,y)} + \frac{\sqrt{2\beta}}{\sqrt{T_{x,t} + \alpha S}} \right) + \frac{28\beta^2 S}{T_t} \sum_x \frac{\mathbb{I}\{T_{x,t} > 0\}}{T_{x,t} + \alpha S} \\
&\le \frac{2\beta}{T_t} \sum_x G(x) \mathbb{I}\{T_{x,t} > 0\} + \frac{13\beta^{3/2}\sqrt{S}}{T_t} \sum_x \sqrt{\frac{G(x)\mathbb{I}\{T_{x,t} > 0\}}{T_{x,t} + \alpha S}} + \frac{39\beta^2 S}{T_t} \sum_x \frac{\mathbb{I}\{T_{x,t} > 0\}}{T_{x,t} + \alpha S}.
\end{aligned}$$

To prove the lower bound on the index, we recall from the proof of Lemma 4 (see (5) with the choices $\zeta = \beta$ and $\zeta' = \frac{\beta}{3}$) that

$$\widehat{P}_t(I - \widehat{P}_t)(x,y) \ge P(I-P)(x,y) - \sqrt{\frac{8\beta T_{x,t}}{(T_{x,t} + \alpha S)^2} \widehat{P}_t(I - \widehat{P}_t)(x,y)} - \frac{12}{T_{x,t} + \alpha S}.$$

Putting this together with the definition of b_{t+1} leads to $b_{t+1} \ge \frac{2\beta}{T_t} \sum_x G(x) \mathbb{I}\{T_{x,t} > 0\}$, and thus completes the proof. □

C Proof of Theorem 1

Consider a chain k and assume that the event C (defined in Definition 2) holds. Applying Lemma 7, we obtain

$$\begin{aligned}
L_{k,n} &\leq \frac{2\beta}{T_{k,n}} \sum_{x:T_{k,x,n}>0} G_k(x) + \frac{2\sqrt{2}}{3}\frac{\beta^{3/2}\sqrt{S}}{T_{k,n}} \sum_x \frac{T_{k,x,n}^{3/2}\sqrt{G_k(x)}}{(T_{k,x,n}+\alpha S)^2} + \frac{S\beta^2}{9T_{k,n}} \sum_x \frac{T_{k,x,n}}{(T_{k,x,n}+\alpha S)^2} \\
&\leq \frac{2\beta}{T_{k,n}} \sum_{x:T_{k,x,n}>0} G_k(x) + \frac{2\sqrt{2}}{3}\frac{\beta^{3/2}}{T_{k,n}} \sqrt{S \sum_{x:T_{k,x,n}>0} G_k(x)} \sqrt{\sum_{x:T_{k,x,n}>0} \frac{1}{T_{k,x,n}+\alpha S}} \\
&\quad + \frac{S\beta^2}{9T_{k,n}} \sum_{x:T_{k,x,n}>0} \frac{1}{T_{k,x,n}+\alpha S} \\
&= \left(\sqrt{\frac{2\beta}{T_{k,n}} \sum_{x:T_{k,x,n}>0} G_k(x)} + \sqrt{\frac{S\beta^2}{9T_{k,n}} \sum_{x:T_{k,x,n}>0} \frac{1}{T_{k,x,n}+\alpha S}} \right)^2 ,
\end{aligned}$$

where we have used Cauchy-Schwarz in the second inequality. Introducing

$$A_{1,k} := \frac{S\beta^2}{9T_{k,n}} \sum_{x:T_{k,x,n}>0} \frac{1}{T_{k,x,n}+\alpha S} \quad \text{and} \quad A_{2,k} := \frac{2\beta}{T_{k,n}} \sum_{x:T_{k,x,n}>0} G_k(x)\,,$$

we provide upper bounds on $A_{1,k}$ and $A_{2,k}$ in the following lemmas:

Lemma 9 *On the event C, it holds for any chain k and any n:*

$$A_{1,k} \leq \frac{0.175KS^2\beta^2}{n-K}\,.$$

Lemma 10 *Assume that the event C holds. Then for any chain k and n:*

$$A_{2,k} \leq \frac{288KS^2\beta^2}{n-2K} + \frac{550K^2S^2\beta^2}{(n-2K)^2}\,.$$

Applying Lemmas 9 and 10 gives

$$\begin{aligned}
L_{k,n} &\leq (\sqrt{A_{1,k}} + \sqrt{A_{2,k}})^2 \leq \frac{KS^2\beta^2}{n-2K} \left(\sqrt{288 + \frac{550K}{n-2K}} + \sqrt{0.175} \right)^2 \\
&\leq KS^2\beta^2 \left(\frac{304}{n-2K} + \frac{564K}{(n-2K)^2} \right) ,
\end{aligned}$$

where we have used

$$\left(\sqrt{288 + \frac{550K}{n-2K}} + \sqrt{0.175} \right)^2 \leq 289 + \frac{550K}{n-2K} + 2\sqrt{0.175}\sqrt{288 + \frac{550K}{n-2K}} \leq 304 + \frac{564K}{n-2K}\,.$$

Finally, using the inequality $(n-2K)^{-1} \leq n^{-1} + 4Kn^{-2}$ valid for $n \geq 4K$, and noting that the event C holds with a probability higher than $1-\delta$, we get the desired bound on the loss. $\square$

C.1 Proof of Lemma 9

Assume that C holds. We claim that there exists a chain j such that $T_{j,n} \geq \frac{n}{K}$. We show this claim by contradiction: If for all j, $T_{j,n} < \frac{n}{K}$, then $\sum_{j=1}^{K} T_{j,n} < n$, which is a contradiction.

Now, consider a chain j such that $T_{j,n} \geq \frac{n}{K}$. Let $t+1 \leq n$ be the last time that it has been sampled. Hence, $T_{j,t+1} = T_{j,n}$ and $T_{j,t} = T_{j,n} - 1 \geq \frac{n}{K} - 1$. Applying Lemma 8 for chain j, it follows that on the event C,

$$
\begin{aligned}
b_{j,t+1} &\leq \frac{2\beta}{T_{j,t}} \sum_{x:T_{j,x,t}>0} G_j(x) + \frac{13\beta^{3/2}\sqrt{S}}{T_{j,t}} \sum_{x:T_{j,x,t}>0} \sqrt{\frac{G_j(x)}{T_{j,x,t}+\alpha S}} + \frac{39\beta^2 S}{T_{j,t}} \sum_{x:T_{j,x,t}>0} \frac{1}{T_{j,x,t}+\alpha S} \\
&\leq \frac{2\beta}{T_{j,t}} \sum_{x:T_{j,x,t}>0} G_j(x) + \frac{13\beta^{3/2}\sqrt{S}}{T_{j,t}} \sqrt{\sum_{x:T_{j,x,t}>0} G_j(x)} \sqrt{\sum_{x:T_{j,x,t}>0} \frac{1}{T_{j,x,t}+\alpha S}} \\
&\quad + \frac{39\beta^2 S}{T_{j,t}} \sum_{x:T_{j,x,t}>0} \frac{1}{T_{j,x,t}+\alpha S} \\
&\leq \frac{K}{n-K} \left(2\beta \sum_x G_j(x) + 13\beta^{3/2} \sqrt{S \sum_x G_j(x)} \sqrt{\sum_x \frac{1}{1+\alpha S}} + 39\beta^2 S \sum_x \frac{1}{1+\alpha S} \right) \\
&\leq \frac{K}{n-K} \left(2\beta \sum_x G_j(x) + 12 S\beta^{3/2} \sqrt{\sum_x G_j(x)} + 30\beta^2 S^2 \right) .
\end{aligned}
$$

Noting that $\sum_x G_j(x) \leq S-1$, we get

$$
b_{j,t+1} \leq \frac{K}{n-K} \left(2\beta(S-1) + 12 S\beta^{3/2}\sqrt{S-1} + 30\beta^2 S^2 \right) \leq \frac{44 K S^2 \beta^2}{n-K} ,
$$

where we have used that $S \geq 2$. Note that for any chain i, by the definition of $b_{i,t+1}$,

$$
b_{i,t+1} \geq \frac{28\beta^2 S}{T_{i,t}} \sum_x \frac{\mathbb{I}\{T_{i,x,t}>0\}}{T_{i,x,t}+\alpha S} \geq \frac{28\beta^2 S}{T_{i,n}} \sum_x \frac{\mathbb{I}\{T_{i,x,n}>0\}}{T_{i,x,n}+\alpha S} .
$$

Furthermore, since j is played at time t, it holds that for any chain $i \neq j$, $b_{i,t+1} \leq b_{j,t+1}$, so that for any chain i,

$$
b_{j,t+1} \geq b_{i,t+1} \geq \frac{28\beta^2 S}{T_{i,n}} \sum_x \frac{\mathbb{I}\{T_{i,x,n}>0\}}{T_{i,x,n}+\alpha S} .
$$

Thus, combining this with the upper bound on $b_{j,t+1}$ leads to the desired results. □

C.2 Proof of Lemma 10

The proof borrows some ideas from the proof of Lemma 1 in [2]. Consider a chain j that is sampled at least once after initialization, and let $t+1 (> 2K)$ be the last time it was sampled. Hence, $T_{j,t} = T_{j,n} - 1$ and $T_{j,t+1} = T_{j,n}$. Moreover, let X_{t+1} be the observed state of j at $t+1$. Then, $T_{j,X_{t+1},t} = T_{j,X_{t+1},n} - 1$ and $T_{j,X_{t+1},t+1} = T_{j,X_{t+1},n}$, whereas for all $x \neq X_{t+1}$, $T_{j,x,t} = T_{j,x,t+1} = T_{j,x,n}$. We thus have, $T_{j,x,t} \geq T_{j,x,n} - 1$ for all $x \in \mathcal{S}$.

By the design of the algorithm, for any chain k, $b_{k,t+1} \le b_{j,t+1}$. Using Lemma 8 yields

$$\begin{aligned}\frac{2\beta}{T_{k,n}}\sum_x G_k(x) &\le \frac{2\beta}{T_{j,n}-1}\sum_x G_j(x) + \frac{13\beta^{3/2}\sqrt{S}}{T_{j,n}-1}\sum_x\sqrt{\frac{G_j(x)}{T_{j,x,t}+\alpha S}} + \frac{39\beta^2 S}{T_{j,n}-1}\sum_x\frac{1}{T_{j,x,t}+\alpha S}\\ &\le \frac{2\beta}{T_{j,n}-1}\sum_x G_j(x) + \frac{26\beta^{3/2}\sqrt{S}}{T_{j,n}-1}\sum_x\sqrt{\frac{G_j(x)}{T_{j,x,n}+\alpha S}}\\ &\quad + \frac{156\beta^2 S}{T_{j,n}-1}\sum_x\frac{1}{T_{j,x,n}+\alpha S}\,,\end{aligned}$$

where in the second line we have used that for $\alpha = \frac{1}{3S}$ and $T_{j,x,n} \ge 1$,

$$T_{j,x,t} + \alpha S \ge T_{j,x,n} - 1 + \alpha S \ge \frac{T_{j,x,n}+\alpha S}{4}\,.$$

The above holds for any chain k, and any chain j that is sampled at least once after the initialization (hence, $T_{j,n} > 2$). Summing over such choices of j gives

$$\begin{aligned}&\frac{2\beta}{T_{k,n}}\sum_x G_k(x)\sum_{j:T_{j,n}>2}(T_{j,n}-1)\\ &\quad\le 2\beta\sum_j\sum_x G_j(x) + 26\beta^{3/2}\sqrt{S}\sum_j\sum_x\sqrt{\frac{G_j(x)}{T_{j,x,n}+\alpha S}} + 156\beta^2 S\sum_j\sum_x\frac{1}{T_{j,x,n}+\alpha S}\\ &\quad\le 2\beta\Lambda + 26\beta^{3/2}\sqrt{S\Lambda}\sqrt{\sum_j\sum_x\frac{1}{T_{j,x,n}+\alpha S}} + 156\beta^2 S\sum_j\sum_x\frac{1}{T_{j,x,n}+\alpha S}\,,\end{aligned}$$

where we have used Cauchy-Schwarz in the last inequality, and that $\sum_j\sum_x G_j(x) = \Lambda$. Noting that $\sum_{j:T_{j,n}>2}(T_{j,n}-1) \ge n - 2K$ yields

$$\frac{2\beta}{T_{k,n}}\sum_x G_k(x) \le \frac{2\beta\Lambda}{n-2K} + \frac{26\beta^{3/2}\sqrt{\Lambda}}{n-2K}\sqrt{\sum_j\sum_x\frac{S}{T_{j,x,n}+\alpha S}} + \frac{156\beta^2}{n-2K}\sum_j\sum_x\frac{S}{T_{j,x,n}+\alpha S}\,.$$

By Lemma 9, $\sum_x \frac{S}{T_{j,x,n}+\alpha S} \le \frac{1.58KS^2}{n-K}T_{j,n}$ for any chain j, which gives

$$\begin{aligned}\frac{2\beta}{T_{k,n}}\sum_x G_k(x) &\le \frac{2\beta\Lambda}{n-2K} + \frac{33\beta^{3/2}\sqrt{\Lambda}}{n-2K}\sqrt{\frac{KS^2}{n-K}\sum_j T_{j,n}} + \frac{263KS^2\beta^2}{(n-2K)(n-K)}\sum_j T_{j,n}\\ &\le \frac{2\beta KS}{n-2K} + \frac{33\beta^{3/2}KS^{3/2}}{n-2K}\sqrt{\frac{n}{n-2K}} + \frac{263KS^2\beta^2 n}{(n-2K)^2}\\ &\le \frac{288KS^2\beta^2}{n-2K} + \frac{550K^2S^2\beta^2}{(n-2K)^2}\,,\end{aligned}$$

where we have used $\sum_j T_{j,n} = n$, $\Lambda \le KS$, $S \ge 2$, and $\sqrt{\frac{n}{n-2K}} \le 1 + \frac{K}{n-2K}$. □

D Proof of Theorem 2

Let $\delta \in (0,1)$ and $n \ge n_{\text{cutoff}}$. To control the loss in this case, we first state the following result for the concentration of empirical state distribution $\hat{\pi}_{k,n}$.

Lemma 11 (Concentration of Empirical State Distributions) *Assume that event C holds and $n \geq n_{\textit{cutoff}}$. Then, for any chain k and state x, $\hat{\pi}_{k,n}(x)^{-1} \leq 2\pi_k(x)^{-1}$ with probability at least $1-\delta$.*

Recalling that $\hat{\pi}_{k,n}(x) = \frac{T_{k,x,n}}{T_{k,n}}$ for all $x \in \mathcal{S}$, on the event C (defined in Definition 2), we have by Lemma 7 and Lemma 11,

$$\begin{aligned}
L_{k,n} &\leq \frac{2\beta}{T_{k,n}} \sum_x G_k(x) + \frac{2\sqrt{2}}{3} \frac{\beta^{3/2}\sqrt{S}}{T_{k,n}^{3/2}} \sum_x \sqrt{\frac{G_k(x)}{\hat{\pi}_{k,n}(x)}} + \frac{S\beta^2}{9T_{k,n}^2} \sum_x \frac{1}{\hat{\pi}_{k,n}(x)} \\
&\leq \frac{2\beta}{T_{k,n}} \sum_x G_k(x) + \frac{4\beta^{3/2}\sqrt{S}}{3T_{k,n}^{3/2}} \sum_x \sqrt{\frac{G_k(x)}{\pi_k(x)}} + \frac{2SH_k\beta^2}{9T_{k,n}^2} \\
&\leq \frac{2\beta}{T_{k,n}} \sum_x G_k(x) + \frac{4\beta^{3/2}}{3T_{k,n}^{3/2}} \sqrt{SH_k \sum_x G_k(x)} + \frac{2SH_k\beta^2}{9T_{k,n}^2}\,,
\end{aligned}$$

with probability at least $1-\delta$, where the last step follows from the Cauchy-Schwarz inequality.

To control the right-hand side of the above, we first provide an upper bound on $\frac{2\beta}{T_{k,n}} \sum_x G_k(x)$ assuming that the event C holds:

Lemma 12 *Assume that the event C holds. Then, for any chain k and $n \geq n_{\textit{cutoff}}$, it holds that*

$$\frac{2\beta}{T_{k,n}} \sum_x G_k(x) \leq \frac{A_1}{n} + \frac{A_2}{n^{3/2}} + \frac{A_3}{n^2} + \widetilde{\mathcal{O}}(n^{-5/2})\,,$$

with probability at least $1-\delta$, where

$$A_1 = 2\beta\Lambda, \quad A_2 = 150\beta^{3/2}K\sqrt{S\Lambda H_{\max}}, \quad A_3 = \frac{3912KSH_{\max}\beta^2}{\eta_{\min}}\,.$$

Applying Lemma 12, and noting $\mathbb{P}(C) \geq 1-\delta$ (see Lemma 6), we obtain the following bound on $L_{k,n}$, which holds with probability greater than $1-2\delta$:

$$\begin{aligned}
L_{k,n} &\leq \frac{2\beta}{T_{k,n}} \sum_x G_k(x) + \frac{4\beta^{3/2}}{3T_{k,n}^{3/2}} \sqrt{SH_k \sum_x G_k(x)} + \frac{2SH_k\beta^2}{9T_{k,n}^2} \\
&\leq \frac{2\beta}{T_{k,n}} \sum_x G_k(x) + \left(\frac{2\beta}{T_{k,n}} \sum_x G_k(x)\right)^{3/2} \frac{0.48\sqrt{SH_k}}{\sum_x G_k(x)} + \frac{2SH_k}{9(\sum_x G_k(x))^2} \left(\frac{\beta}{T_{k,n}} \sum_x G_k(x)\right)^2 \\
&\leq \frac{A_1}{n} + \frac{A_2}{n^{3/2}} + \frac{A_3}{n^2} + \left(\frac{A_1}{n} + \frac{A_2}{n^{3/2}} + \frac{A_3}{n^2}\right)^{3/2} \frac{0.48\sqrt{SH_k}}{\sum_x G_k(x)} \\
&\quad + \frac{2SH_k}{9(\sum_x G_k(x))^2} \left(\frac{A_1}{n} + \frac{A_2}{n^{3/2}} + \frac{A_3}{n^2}\right)^2 + \widetilde{\mathcal{O}}(n^{-5/2}) \\
&\overset{(a)}{\leq} \frac{A_1}{n} + \frac{A_2}{n^{3/2}} + \frac{A_3}{n^2} + \frac{0.84\sqrt{SH_k}}{\sum_x G_k(x)} \left(\frac{A_1^{3/2}}{n^{3/2}} + \frac{A_2^{3/2}}{n^{9/4}} + \frac{A_3^{3/2}}{n^3}\right) \\
&\quad + \frac{4SH_k}{9(\sum_x G_k(x))^2} \left(\frac{A_1^2}{n^2} + \frac{A_2^2}{n^3} + \frac{A_3^2}{n^4}\right) + \widetilde{\mathcal{O}}(n^{-5/2}) \\
&\leq \frac{A_1}{n} + \frac{1}{n^{3/2}} \left(A_2 + \frac{0.84\sqrt{SH_k}}{\sum_x G_k(x)} A_1^{3/2}\right) + \widetilde{\mathcal{O}}(n^{-2}) \\
&\leq \frac{2\beta\Lambda}{n} + \frac{\beta^{3/2}}{n^{3/2}} \left(150K\sqrt{S\Lambda H_{\max}} + \frac{2.4\sqrt{SH_k\Lambda}}{\eta_k}\right) + \widetilde{\mathcal{O}}(n^{-2})\,,
\end{aligned}$$

where (a) follows from the fact that for positive numbers $a_i, i = 1, \dots, m$, we have by Jensen's inequality,

$$\Big(\frac{1}{m}\sum_{i=1}^{m} a_i\Big)^{3/2} \le \frac{1}{m}\sum_{i=1}^{m} a_i^{3/2}\,.$$

so that $(\sum_{i=1}^m a_i)^{3/2} \le \sqrt{m}\sum_{i=1}^m a_i^{3/2}$. Finally, taking the maximum over k completes the proof. $\square$

D.1 Proof of Lemma 11

By Lemma 5, we have for all chains k and all $x \in \mathcal{S}$,

$$|\hat{\pi}_{k,n}(x) - \pi_k(x)| \le \xi_{k,x,n} := \sqrt{\frac{8\pi_k(x)\varepsilon_k}{T_{k,n}}} + \frac{20\varepsilon_k}{T_{k,n}}\,,$$

with probability at least $1-\delta$, where $\varepsilon_k := \frac{1}{\gamma_k}\log\Big(\frac{K}{\delta}\sqrt{\frac{2}{\underline{\pi}_k}}\Big)$. It is easy to verify that if $T_{k,n} \ge \frac{96\varepsilon_k}{\underline{\pi}_k}$, then $\xi_{k,x,n} \le \pi_k(x)/2$, so that for all k and all x,

$$\begin{aligned}\frac{1}{\hat{\pi}_{k,n}(x)} &= \frac{1}{\pi_k(x)} + \frac{\pi_k(x) - \hat{\pi}_{k,n}(x)}{\hat{\pi}_{k,n}(x)\pi_k(x)} \\ &\le \frac{1}{\pi_k(x)} + \frac{\xi_{k,x,n}}{\pi_k(x)(\pi_k(x) - \xi_{k,x,n})} \\ &\le \frac{2}{\pi_k(x)}\,,\end{aligned}$$

with probability at least $1-\delta$.

It remains to show that if $n \ge n_{\text{cutoff}}$, we have $T_{k,n} \ge \frac{96\varepsilon_k}{\underline{\pi}_k}$. Indeed, when C occurs, as a consequence of Lemma 9, one has

$$\frac{S\beta^2}{9T_{k,n}}\sum_{x:T_{k,x,n}>0}\frac{1}{T_{k,x,n}+\alpha S} \le \frac{0.175KS^2\beta^2}{n-K}\,,$$

with probability at least $1-\delta$. Using the trivial bound $T_{k,x,n} \le T_{k,n}$, it follows that

$$\frac{S^2\beta^2}{9T_{k,n}(T_{k,n}+\alpha S)} \le \frac{0.175KS^2\beta^2}{n-K}\,,$$

so that

$$T_{k,n} \ge \sqrt{\frac{n-K}{1.575K}} - \frac{1}{3} \ge 0.56\sqrt{\frac{n}{K}} - \frac{1}{3} \ge 0.327\sqrt{\frac{n}{K}}\,.$$

with probability greater than $1-\delta$. Putting together, we deduce that if n satisfies $0.327\sqrt{\frac{n}{K}} \ge \frac{96\varepsilon_k}{\underline{\pi}_k}$, we have $\xi_{k,x,n} \le \pi_k(x)/2$, and the lemma follows.

Moreover, when the chain k is non-reversible, we may use [20, Theorem 3.4] (instead of Lemma 9), and follow the exact same lines as above to deduce that if $n \ge K\max_k\Big(\frac{300}{\gamma_{\mathsf{ps},k}\underline{\pi}_k}\log\Big(\frac{2K}{\delta}\sqrt{\underline{\pi}_k^{-1}}\Big)\Big)^2$, the assertion of the lemma follows. $\square$

D.2 Proof of Lemma 12

The proof borrows some ideas from the proof of Lemma 1 in [2]. Consider a chain j that is sampled at least once after initialization, and let $t+1 (> 2K)$ be the last time it was sampled. Hence, $T_{j,t} = T_{j,n} - 1$ and $T_{j,t+1} = T_{j,n}$. Moreover, let X_{t+1} be the observed state of j at $t+1$. Then, $T_{j,X_{t+1},t} = T_{j,X_{t+1},n} - 1$ and $T_{j,X_{t+1},t+1} = T_{j,X_{t+1},n}$, whereas for all $x \neq X_{t+1}$, $T_{j,x,t} = T_{j,x,t+1} = T_{j,x,n}$. We thus have, $T_{j,x,t} \geq T_{j,x,n} - 1$ for all $x \in \mathcal{S}$.

By the design of the algorithm, for any chain k, $b_{k,t+1} \leq b_{j,t+1}$. Applying Lemma 8 gives

$$\begin{aligned}\frac{2\beta}{T_{k,t}} \sum_x G_k(x) &\leq \frac{2\beta}{T_{j,t}} \sum_x G_j(x) + \frac{c_3\sqrt{S}}{T_{j,t}} \sum_x \sqrt{\frac{G_j(x)}{T_{j,x,t} + \alpha S}} + \frac{c_4 S}{T_{j,t}} \sum_x \frac{1}{T_{j,x,t} + \alpha S} \\ &\leq \frac{2\beta}{T_{j,n} - 1} \sum_x G_j(x) + \frac{2c_3\sqrt{S}}{T_{j,n} - 1} \sum_x \sqrt{\frac{G_j(x)}{T_{j,x,n} + \alpha S}} + \frac{4c_4 S}{T_{j,n} - 1} \sum_x \frac{1}{T_{j,x,n} + \alpha S}\,,\end{aligned}$$

where $c_3 = 13\beta^{3/2}$ and $c_4 = 39\beta^2$, and where where in the second line we have used that for $\alpha = \frac{1}{3S}$ and $T_{j,x,n} \geq 1$

$$T_{j,x,t} + \alpha S \geq T_{j,x,n} - 1 + \alpha S \geq \frac{T_{j,x,n} + \alpha S}{4}\,.$$

Now, applying Lemma 11 and using $T_{k,t} \leq T_{k,n}$ yield

$$\begin{aligned}\frac{2\beta}{T_{k,n}} \sum_x G_k(x) &\leq \frac{2\beta}{T_{j,n} - 1} \sum_x G_j(x) + \frac{c_3\sqrt{8S}}{(T_{j,n} - 1)T_{j,n}^{1/2}} \sum_x \sqrt{\frac{G_j(x)}{\pi_j(x)}} + \frac{8c_4 S}{(T_{j,n} - 1)T_{j,n}} \sum_x \frac{1}{\pi_j(x)} \\ &\leq \frac{1}{T_{j,n} - 1} \left(2\beta \sum_x G_j(x) + \frac{\sqrt{8}c_3}{T_{j,n}^{1/2}} \sqrt{SH_j \sum_x G_j(x)} + \frac{8c_4 S H_j}{T_{j,n}} \right)\,.\end{aligned}$$

Note that the above relation is valid for any k, and any j that is sampled after initialization (i.e., $T_{j,n} > 2$). Summing over such choices of j gives

$$\sum_{j:T_{j,n}>2} \frac{2\beta \sum_x G_k(x)}{T_{k,n}} (T_{j,n} - 1) \leq \sum_{j:T_{j,n}>2} \left(2\beta \sum_x G_j(x) + \frac{c_3\sqrt{8SH_j \sum_x G_j(x)}}{T_{j,n}^{1/2}} + \frac{8c_4 S H_j}{T_{j,n}} \right)\,.$$

Noting that $\sum_{j:T_{j,n}>2} (T_{j,n} - 1) \geq n - 2K$, we have

$$\begin{aligned}\frac{2\beta}{T_{k,n}} \sum_x G_k(x) &\leq \frac{1}{n - 2K} \sum_j \left(2\beta \sum_x G_j(x) + \frac{c_3\sqrt{8SH_j \sum_x G_j(x)}}{T_{j,n}^{1/2}} + \frac{8c_4 S H_j}{T_{j,n}} \right) \\ &\leq \frac{2\beta\Lambda}{n - 2K} + \frac{c_3\sqrt{8S}}{n - 2K} \sum_j \sqrt{H_j \frac{\sum_x G_j(x)}{T_{j,n}}} + \frac{8c_4 S}{n - 2K} \sum_j \frac{H_j}{T_{j,n}}\,. \qquad (9)\end{aligned}$$

To carefully control the right-hand side of the last inequality, we use the following lemma:

Lemma 13 *Under the same assumption of Lemma 12, we have for any chain j,*

$$\begin{aligned}\frac{\beta}{T_{j,n}} &\leq \frac{\beta}{\eta_{\min}(n - 2K)} + \frac{c_3}{\eta_{\min}^2 (n - 2K)^{3/2}} \sqrt{2SH_{\max}/\Lambda} + \frac{4c_4 S H_{\max}}{\Lambda \eta_{\min}^3 (n - 2K)^2}\,, \\ \sqrt{\frac{2\beta}{T_{j,n}} \sum_x G_j(x)} &\leq \sqrt{\frac{2\beta\Lambda}{n - 2K}} + \frac{18\beta\sqrt{SH_{\max}}}{\eta_{\min}(n - 2K)}\,.\end{aligned}$$

Now, applying Lemma 13 yields

$$\begin{aligned}\frac{2\beta}{T_{k,n}}\sum_x G_k(x) &\le \frac{2\beta\Lambda}{n-2K} + \frac{37\beta\sqrt{SH_{\max}}}{n-2K}\sum_j\left(\sqrt{\frac{2\beta\Lambda}{n-2K}} + \frac{18\beta\sqrt{SH_{\max}}}{\eta_{\min}(n-2K)}\right)\\ &\quad + \frac{312\beta SH_{\max}}{n-2K}\sum_j\left(\frac{\beta}{\eta_{\min}(n-2K)} + \frac{c_3}{\eta_{\min}^2(n-2K)^{3/2}}\sqrt{2SH_j/\Lambda} + \frac{4c_4SH_j}{\Lambda\eta_{\min}^3(n-2K)^2}\right)\\ &\le \frac{2\beta\Lambda}{n-2K} + \frac{53\beta^{3/2}K\sqrt{SH_{\max}\Lambda}}{(n-2K)^{3/2}} + \frac{978KSH_{\max}\beta^2}{\eta_{\min}(n-2K)^2} + \widetilde{\mathcal{O}}((n-2K)^{-5/2})\,.\end{aligned}$$

Finally, using the inequality $(n-2K)^{-1} \le n^{-1} + 4Kn^{-2}$ and $n - 2K \ge n/2$ valid for all $n \ge 4K$, we get the desired result:

$$\frac{2\beta}{T_{k,n}}\sum_x G_k(x) \le \frac{2\beta\Lambda}{n} + \frac{150\beta^{3/2}K\sqrt{S\Lambda H_{\max}}}{n^{3/2}} + \frac{3912KSH_{\max}\beta^2}{\eta_{\min}n^2} + \widetilde{\mathcal{O}}(n^{-5/2})\,.$$

□

D.3 Proof of Lemma 13

The proof borrows some ideas from the proof of Lemma 1 in [2]. Consider a chain j that is sampled at least once after initialization, and let $t+1 (> 2K)$ be the last time it was sampled. Hence, $T_{j,t} = T_{j,n} - 1$. Using the same arguments as in the beginning of the proof of Lemma 12, we have on the event C,

$$\frac{2\beta}{T_{k,n}}\sum_x G_k(x) \le \frac{1}{T_{j,n}-1}\left(2\beta\sum_x G_j(x) + c_3\sqrt{\frac{8SH_j\sum_x G_j(x)}{T_{j,n}}} + \frac{8c_4SH_j}{T_{j,n}}\right). \tag{10}$$

Note that (10) is valid for any k, and any j that is sampled after initialization.

Now consider a chain j such that $T_{j,n} - 2 \ge \eta_j(n-2K)$. In other words, j is over-sampled (w.r.t. budget $n-2K$). In particular, j is sampled at least once after initialization. Hence, using (10) and noting that $T_{j,n} \ge \eta_j(n-2K) + 2$, we obtain

$$\begin{aligned}\frac{2\beta}{T_{k,n}}\sum_x G_k(x) &\le \frac{1}{\eta_j(n-2K)}\left(2\beta\sum_x G_j(x) + c_3\sqrt{\frac{8SH_j\sum_x G_j(x)}{\eta_j(n-2K)}} + \frac{8c_4SH_j}{\eta_j(n-2K)}\right)\\ &\overset{\text{(a)}}{\le} \frac{2\beta\Lambda}{n-2K} + \frac{c_3\sqrt{8S\Lambda H_j}}{\eta_j(n-2K)^{3/2}} + \frac{8c_4SH_j}{\eta_j^2(n-2K)^2}\,,\end{aligned} \tag{11}$$

where (a) follows from the definition of η_j. Multiplying both sides on $\frac{\eta_k}{2\Lambda}$ gives:

$$\begin{aligned}\frac{\beta}{T_{k,n}} &\le \frac{\beta}{\eta_k(n-2K)} + \frac{c_3\sqrt{2S\Lambda H_j}}{\Lambda\eta_k\eta_j(n-2K)^{3/2}} + \frac{4c_4SH_j}{\Lambda\eta_k\eta_j^2(n-2K)^2}\\ &\le \frac{\beta}{\eta_{\min}(n-2K)} + \frac{c_3}{\eta_{\min}^2(n-2K)^{3/2}}\sqrt{2SH_j/\Lambda} + \frac{4c_4SH_j}{\Lambda\eta_{\min}^3(n-2K)^2}\,,\end{aligned} \tag{12}$$

thus verifying the first statement of the lemma. To derive the second statement, we take square-root from both sides of (11):

$$\begin{aligned}\sqrt{\frac{2\beta}{T_{k,n}}\sum_x G_k(x)} &\le \sqrt{\frac{2\beta\Lambda}{n-2K}+\frac{c_3\sqrt{8S\Lambda H_j}}{\eta_j(n-2K)^{3/2}}+\frac{8c_4SH_j}{\eta_j^2(n-2K)^2}}\\ &\le \sqrt{\frac{2\beta\Lambda}{n-2K}+\frac{c_3\sqrt{8S\Lambda H_j}}{\eta_j(n-2K)^{3/2}}}+\frac{\sqrt{8c_4SH_j}}{\eta_j(n-2K)}\\ &\le \sqrt{\frac{2\beta\Lambda}{n-2K}}+\frac{\sqrt{c_3/\beta}\sqrt{SH_{\max}}}{\eta_{\min}(n-2K)}+\frac{\sqrt{8c_4SH_{\max}}}{\eta_{\min}(n-2K)}\,,\end{aligned}$$

where the second and third inequalities respectively follow from $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$ and $\sqrt{a+b}\le\sqrt{a}+\frac{b}{2\sqrt{a}}$ valid for all $a,b>0$. Plugging $c_3=13\beta^{3/2}$ and $c_4=39\beta^2$ into the last inequality verifies the second statement and concludes the proof. □

E Asymptotic Analyses — Proofs of Lemma 1 and Theorem 3

E.1 Proof of Lemma 1

Consider a chain k, and let us denote by $\widetilde{P}_{k,n}$ the corresponding empirical estimator of P_k (corresponding to $\alpha=0$). That is, for all $(x,y)\in\mathcal{S}^2$, $\widetilde{P}_{k,n}(x,y)=\frac{1}{T_{k,x,n}}\sum_{t=2}^n\mathbb{I}\{X_{t-1}=x,X_t=y\}$. Further, let $\widetilde{L}_{k,n}$ denote the corresponding loss for chain k. To prove the lemma, we first show that $\lim_{T_{k,n}\to\infty}T_{k,n}L_{k,n}=\lim_{T_{k,n}\to\infty}T_{k,n}\widetilde{L}_{k,n}$.

To this end, we derive upper and lower bounds on $L_{k,n}$ in terms of $\widetilde{L}_{k,n}$. We have for all $(x,y)\in\mathcal{S}^2$:

$$\begin{aligned}|(\widehat{P}_{k,n}-\widetilde{P}_{k,n})(x,y)| &= \left|\frac{\sum_{t=2}^n\mathbb{I}\{X_{t-1}=x,X_t=y\}+\alpha}{T_{k,x,n}+\alpha S}-\frac{\sum_{t=2}^n\mathbb{I}\{X_{t-1}=x,X_t=y\}}{T_{k,x,n}}\right|\\ &= \frac{\alpha}{T_{k,x,n}(T_{k,x,n}+\alpha S)}\left|T_{k,x,n}-S\sum_{t=2}^n\mathbb{I}\{X_{t-1}=x,X_t=y\}\right|\\ &\le \frac{\alpha S T_{k,x,n}}{T_{k,x,n}(T_{k,x,n}+\alpha S)}\le\frac{\alpha S}{T_{k,x,n}}\,.\end{aligned}$$

We therefore get, on the one hand,

$$L_{k,n}\ge\sum_x\hat{\pi}_{k,n}(x)\|\widehat{P}_{k,n}(x,\cdot)-\widetilde{P}_{k,n}(x,\cdot)\|_2^2+\widetilde{L}_{k,n}\ge\widetilde{L}_{k,n}\,,$$

and on the other hand,

$$\begin{aligned}L_{k,n} &\le \sum_x\hat{\pi}_{k,n}(x)\|\widehat{P}_{k,n}(x,\cdot)-\widetilde{P}_{k,n}(x,\cdot)\|_2^2+\sum_x\hat{\pi}_{k,n}(x)\|P_k(x,\cdot)-\widetilde{P}_{k,n}(x,\cdot)\|_2^2\\ &\quad+\underbrace{2\sum_x\hat{\pi}_{k,n}(x)\sum_y\left|\widehat{P}_{k,n}(x,y)-\widetilde{P}_{k,n}(x,y)\right|\left|\widetilde{P}_{k,n}(x,y)-P_k(x,y)\right|}_{A}\\ &\le \sum_x\hat{\pi}_{k,n}(x)\|\widehat{P}_{k,n}(x,\cdot)-\widetilde{P}_{k,n}(x,\cdot)\|_2^2+\widetilde{L}_{k,n}+2A\\ &\le \sum_x\frac{T_{k,x,n}}{T_{k,n}}\sum_y\Big(\frac{\alpha S}{T_{k,x,n}}\Big)^2+\widetilde{L}_{k,n}+2A=\frac{\alpha^2S^3}{T_{k,n}}\sum_x\frac{1}{T_{k,x,n}}+\widetilde{L}_{k,n}+2A\,.\end{aligned}$$

Furthermore, A is upper bounded as follows:

$$\begin{aligned} A &\le \sum_x \hat{\pi}_{k,n}(x) \frac{\alpha S}{T_{k,x,n}} \sum_y \left| \widetilde{P}_{k,n}(x,y) - P_k(x,y) \right| \\ &\le \sqrt{\sum_x \sum_y \hat{\pi}_{k,n}(x) \frac{\alpha^2 S^2}{T_{k,x,n}^2}} \sqrt{\sum_x \hat{\pi}_{k,n}(x) \sum_y (\widetilde{P}_{k,n}(x,y) - P_k(x,y))^2} \\ &= \sqrt{\sum_x \frac{\alpha^2 S^3}{T_{k,n} T_{k,x,n}}} \sqrt{\widetilde{L}_{k,n}}\,, \end{aligned}$$

where we have used Cauchy-Schwarz in the second line. In summary, we have shown that

$$\widetilde{L}_{k,n} \le L_{k,n} \le \widetilde{L}_{k,n} + \frac{\alpha^2 S^3}{T_{k,n}} \sum_x \frac{1}{T_{k,x,n}} + 2\sqrt{\sum_x \frac{\alpha^2 S^3}{T_{k,n} T_{k,x,n}}} \sqrt{\widetilde{L}_{k,n}}\,,$$

so that

$$T_{k,n}\widetilde{L}_{k,n} \le T_{k,n} L_{k,n} \le T_{k,n}\widetilde{L}_{k,n} + \alpha^2 S^3 \sum_x \frac{1}{T_{k,x,n}} + 2\sqrt{\sum_x \frac{\alpha^2 S^3}{T_{k,x,n}}} \sqrt{T_{k,n}\widetilde{L}_{k,n}}\,.$$

Taking the limit when $T_{k,n} \to \infty$ and noting the fact that when $T_{k,n} \to \infty$, by ergodicity, $T_{k,x,n} \to \infty$ for all $x \in \mathcal{S}$, we obtain $\lim_{T_{k,n}\to\infty} T_{k,n} L_{k,n} = \lim_{T_{k,n}\to\infty} T_{k,n}\widetilde{L}_{k,n}$.

It remains to compute $\lim_{T_{k,n}\to\infty} T_{k,n}\widetilde{L}_{k,n}$. We have

$$\begin{aligned} \widetilde{L}_{k,n} &= \frac{1}{T_{k,n}} \sum_x T_{x,n} \sum_y (\widetilde{P}_{k,n}(x,y) - P_k(x,y))^2 \\ &= \frac{1}{T_{k,n}} \sum_x \sum_y \underbrace{[\sqrt{T_{k,x,n}}(\widetilde{P}_{k,n}(x,y) - P_k(x,y))]^2}_{Z(x,y)}\,. \end{aligned}$$

When $T_{k,n} \to \infty$, by ergodicity, we have $T_{k,x,n} \to \infty$ for all $x \in \mathcal{S}$. Therefore, by the central limit theorem, $\sqrt{T_{k,x,n}}(\widetilde{P}_{k,n}(x,y) - P_k(x,y))$ converges (in distribution) to a Normal distribution with variance $P_k(I - P_k)(x,y)$, and $Z(x,y)$ converges to a Gamma distribution with mean $P_k(I - P_k)(x,y)$. Hence, the mean of $\widetilde{L}_{k,n}$ would be $\frac{1}{T_{k,n}} \sum_x \sum_y P_k(I - P_k)(x,y) = \frac{1}{T_{k,n}} \sum_x G_k(x)$, thus completing the proof. □

E.2 Proof of Theorem 3

Let n be a budget such that $\sqrt{n} \ge n_{\text{cutoff}}$, and let $n_0 := \sqrt{n}$. The proof proceeds in two steps. We first consider the case with budget n_0, and show that at the end of this sub-problem, the index of each chain is well estimated. Then, we consider allocation in the second sub-problem.

Step 1: Bounds on the index at $t = n_0$. Fix a chain k. To derive upper and lower bounds on the index $b_{k,t}$ at $t = n_0$, we derive a lower bound on T_{k,n_0}. Recall that by Lemma 9, we have with the choice $n = n_0$,

$$\frac{S\beta^2}{9T_{k,n_0}} \sum_{x:T_{k,x,n_0}>0} \frac{1}{T_{k,x,n_0} + \alpha S} \le \frac{0.187 K S^2 \beta^2}{n_0 - K}\,,$$

with probability at least $1-\delta$. Using the trivial bound $T_{k,x,n_0} \leq T_{k,n_0}$, it follows that

$$\frac{S^2\beta^2}{9T_{k,n_0}(T_{k,n_0}+\alpha S)} \leq \frac{0.175KS^2\beta^2}{n_0-K}\,,$$

so that

$$T_{k,n_0} \geq \sqrt{\frac{n_0-K}{1.575K}} - \frac{1}{3} \geq \frac{1}{4}\sqrt{\frac{n_0}{K}} \geq \frac{n^{1/4}}{4\sqrt{K}}\,,$$

with probability greater than $1-\delta$.

Noting that $n_0 \geq n_{\text{cutoff}}$, we apply Lemma 8 and 11 to obtain

$$\frac{2\beta}{T_{k,n_0}}\sum_x G_k(x) \leq b_{k,n_0+1} \leq \frac{2\beta}{T_{k,n_0}}\sum_x G_k(x) + \frac{19\beta^{3/2}\sqrt{S}}{T_{k,n_0}^{3/2}}\sum_x\sqrt{\frac{G(x)}{\pi_k(x)}} + \frac{90\beta^2 S}{T_{k,n_0}^2}\sum_x\frac{1}{\pi_k(x)}\,,$$

with probability at least $1-2\delta$. Using Cauchy-Schwarz and recalling $H_k := \sum_k \pi_k(x)^{-1}$, we obtain give

$$\frac{2\beta}{T_{k,n_0}}\sum_x G_k(x) \leq b_{k,n_0+1} \leq \frac{2\beta}{T_{k,n_0}}\sum_x G_k(x) + \frac{19\beta^{3/2}}{T_{k,n_0}^{3/2}}\sqrt{SH_k\sum_x G_k(x)} + \frac{78\beta^2 SH_k}{T_{k,n_0}^2}\,,$$

so that

$$\sum_x G_k(x) \leq \frac{T_{k,n_0}}{2\beta}b_{k,n_0+1} \leq \sum_x G_k(x) + \frac{10\beta^{1/2}}{T_{k,n_0}^{1/2}}\sqrt{SH_k\sum_x G_k(x)} + \frac{39\beta SH_k}{T_{k,n_0}}\,,$$

with probability at least $1-2\delta$. Using the lower bound $T_{k,n_0} \geq \frac{n^{1/4}}{4\sqrt{K}}$ yields

$$\sum_x G_k(x) \leq \frac{T_{k,n_0}}{2\beta}b_{k,n_0+1} \leq \sum_x G_k(x) + \frac{20\beta^{1/2}}{n^{1/8}}\sqrt{SKH_k\sum_x G_k(x)} + \frac{156\beta SH_k\sqrt{K}}{n^{1/4}}\,,$$

with probability at least $1-2\delta$. Let us write the last inequality as

$$\sum_x G_k(x) \leq \frac{T_{k,n_0}}{2\beta}b_{k,n_0+1} \leq \sum_x G_k(x) + \varepsilon_n\,,$$

where $\varepsilon_n = \widetilde{\mathcal{O}}(n^{-1/8})$, thus giving

$$\frac{2\beta}{T_{k,n_0}}\sum_x G_k(x) \leq b_{k,n_0+1} \leq \frac{2\beta}{T_{k,n_0}}\Big(\sum_x G_k(x) + \varepsilon_n\Big)\,.$$

Step 2: The second sub-problem. Now let us consider $n_0+1 \leq t \leq n$. It follows that with probability $1-2\delta$, **BA-MC** allocates according to the following problem

$$\max_{\xi\in[0,\varepsilon_n]^K}\max_k \frac{2\beta}{x_k+T_{k,n_0}}\Big(\sum_x G_k(x)+\xi_k\Big) \quad \text{s.t.:} \quad \sum_k x_k = n-\sqrt{n}\,,$$

whose optimal solution satisfies

$$\frac{(n-\sqrt{n})\sum_x G_k(x)}{\Lambda+K\varepsilon_n} \leq x_k \leq \frac{(n-\sqrt{n})\Big(\sum_x G_k(x)+\varepsilon_n\Big)}{\Lambda}\,.$$

Recalling $\varepsilon_n = \widetilde{\mathcal{O}}(n^{-1/8})$ and noting that $T_{k,n} \geq x_k$, we obtain $\frac{T_{k,n}}{n} \to_{n\to\infty} \frac{\sum_x G_k(x)}{\Lambda}$.

It remains to show that the loss L_n(**BA-MC**) approaches Λ/n as n tends to infinity. By Lemma 1, recall that $\lim_{T_{k,n}\to\infty} T_{k,n}L_{k,n} = \sum_x G_k(x)$. So, using $\frac{T_{k,n}}{n} \to_{n\to\infty} \frac{\sum_x G_k(x)}{\Lambda}$, we conclude $\lim_{n\to\infty} nL_{k,n} = \Lambda$, and the claim follows. □