Mathematics of Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Optimistic Posterior Sampling for Reinforcement Learning: Worst-Case Regret Bounds

Shipra Agrawal, Randy Jia

To cite this article:
Shipra Agrawal, Randy Jia (2022) Optimistic Posterior Sampling for Reinforcement Learning: Worst-Case Regret Bounds. Mathematics of Operations Research

Published online in Articles in Advance 06 May 2022

. https://doi.org/10.1287/moor.2022.1266

http://pubsonline.informs.org/journal/moor

MATHEMATICS OF OPERATIONS RESEARCH
Articles in Advance, pp. 1–30
ISSN 0364-765X (print), ISSN 1526-5471 (online)

Optimistic Posterior Sampling for Reinforcement Learning: Worst-Case Regret Bounds

Shipra Agrawal,[a] **Randy Jia**[a,*]

[a] Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027
*Corresponding author
Contact: sa3305@columbia.edu, https://orcid.org/0000-0003-4486-3871 (SA); rqj2000@columbia.edu, https://orcid.org/0000-0002-7101-9572 (RJ)

Received: September 26, 2020
Revised: October 4, 2021
Accepted: February 12, 2022
Published Online in Articles in Advance: May 6, 2022

MSC2020 Subject Classification: Primary: 68T05, 90-10

https://doi.org/10.1287/moor.2022.1266

Abstract. We present an algorithm based on posterior sampling (aka Thompson sampling) that achieves near-optimal worst-case regret bounds when the underlying Markov decision process (MDP) is communicating with a finite, although unknown, diameter. Our main result is a high probability regret upper bound of $\tilde{O}(DS\sqrt{AT})$ for any communicating MDP with S states, A actions, and diameter D. Here, regret compares the total reward achieved by the algorithm to the total expected reward of an optimal infinite-horizon undiscounted average reward policy in time horizon T. This result closely matches the known lower bound of $\Omega(\sqrt{DSAT})$. Our techniques involve proving some novel results about the anti-concentration of Dirichlet distribution, which may be of independent interest.

Funding: This work was supported in part by an NSF CAREER award [CMMI 1846792] awarded to author S. Agrawal.

Keywords: Thompson sampling • reinforcement learning • Markov decision process • regret bounds

1. Introduction

Reinforcement learning (RL) refers to the problem of learning and planning in sequential decision-making systems when the underlying system dynamics are unknown and may need to be learned by trying out different options and observing their outcomes. A typical model for the sequential decision-making problem is a Markov decision process (MDP), which proceeds in discrete time steps. At each time step, the system is in some state s, and the decision maker may take any available action a to obtain a (possibly stochastic) reward. The system then transitions to the next state according to a fixed state transition distribution. The reward and the next state depend on the current state s and the action a but are independent of all the previous states and actions. In the reinforcement learning problem, the underlying state transition distributions and/or reward distributions are unknown and need to be *learned* using the observed rewards and state transitions while aiming to *maximize* the cumulative reward. This requires the algorithm to manage the tradeoff between exploration versus exploitation, that is, exploring different actions in different states to learn the model more accurately versus taking actions that currently seem to be reward maximizing.

Exploration-exploitation tradeoff has been studied extensively in the context of stochastic multiarmed bandit (MAB) problems, which are essentially MDPs with a single state. The performance of MAB algorithms is typically measured through *regret*, which compares the total reward obtained by the algorithm to the total expected reward of an optimal action. Optimal regret bounds have been established for many variations of MAB (see Bubeck and Cesa-Bianchi [11] for a survey), with a large majority of results obtained using the upper confidence bound (UCB) algorithm, or more generally, the *optimism in the face of uncertainty* principle. Under this principle, the learning algorithm maintains tight overestimates (or optimistic estimates) the expected rewards for individual actions, and at any given step, picks the action with the highest optimistic estimate. More recently, posterior sampling, aka Thompson sampling (Thompson [40]), has emerged as another popular algorithm design principle in MAB, owing its popularity to a simple and extendable algorithmic structure, an attractive empirical performance (Chapelle and Li [14], Kauffman et al. [23]), and provably optimal performance bounds that have been recently obtained for many variations of MAB (Agrawal and Goyal [3], Agrawal and Goyal [4], Agrawal and Goyal [5], Bubeck and Liu [12], Russo and Van Roy [32], Russo and Van Roy [33]). In this approach, the algorithm maintains a Bayesian posterior distribution for the expected reward of every action; then at any given step, it generates an independent sample from each of these posteriors and takes the action with the highest sample value.

In this paper, we consider the RL problem in a similar regret-based framework, where the total reward of the reinforcement learning algorithm is compared with the total expected reward achieved by a single benchmark policy over a time horizon T. In our setting, the benchmark policy is the *infinite-horizon undiscounted average*

reward optimal policy for the underlying MDP. Here, the underlying MDP is assumed to have finite states S and finite actions A and is assumed be communicating with (unknown) finite diameter D. The diameter D is an upper bound on the expected time it takes to go from any state s to any other state s' using an appropriate policy, for any pair s, s'. Most algorithms studied in the literature that achieve interesting regret bounds assume the finite diameter setting. We note, however, an exception to this is the work of Fruit et al. [17], who consider potentially infinite diameter in the case of weakly-communicating or multichain MDPs. The UCRL2 algorithm of Jaksch et al. [20], which is based on the optimism principle, achieved the first finite time regret upper bound of $\tilde{O}(DS\sqrt{AT})$ for this problem. A similar bound was achieved by Bartlett and Tewari [9], although under an assumption of known diameter. Jaksch et al. [20] also established a worst-case lower bound of $\Omega(\sqrt{DSAT})$ on the regret of any algorithm for this problem. More recent works have achieved better regret bounds than the original UCRL2 work. Fruit et al. [18] demonstrated a regret bound of $\tilde{O}(S\sqrt{DAT})$ with a more careful analysis of the UCRL2 algorithm. Using slightly different problem parameters, Ortner [26] proves a regret bound of $\tilde{O}(\sqrt{t_{mix}SAT})$, where t_{mix} is a mixing time parameter of the MDP, and Talebi and Maillard [38] prove a $\tilde{O}(\sqrt{\sum_{s,a} V^*_{s,a} ST})$ regret bound, where $V^*_{s,a}$ is the variance of the bias function. The best result under our setting has been achieved by Zhang and Ji [41], where they design an algorithm that achieves near-optimal regret of $\tilde{O}(\sqrt{DSAT})$. The drawback, however, is that their algorithm is inefficient to implement, as it requires an optimization over a nonconvex constraint set.

Our main contribution is a posterior sampling–based algorithm with a high probability worst-case regret upper bound of $\tilde{O}(DS\sqrt{AT})$. Although the theoretical bound is of the same order as that of UCRL2, we believe posterior sampling–based algorithms have advantages over confidence interval–based techniques (e.g., UCRL2), in that we are no longer required to search over high-dimensional continuous intervals in solving the MDP; instead, we can simply apply standard MDP techniques under posterior sampling. In particular, this means that we are not restricted to extended value iteration (as in Jaksch et al. [20]) as the technique for solving MDPs but rather can use any black-box method. Thus, although UCRL2 with extended value iteration to solve the MDP is computationally efficient (in S and A), depending on the domain and the problem structure, other techniques may be more efficient for the given MDP and can be used in our posterior sampling–based algorithm. Our algorithm uses an "optimistic" version of the posterior sampling heuristic while using several ideas from the algorithm design structure in Jaksch et al. [20], such as an epoch-based execution and the extended MDP construction. Our algorithm proceeds in epochs, where in the beginning of every epoch, for every state and action, it generates $\psi = \tilde{O}(S)$ sample transition probability vectors from a posterior distribution and then solves an extended MDP with ψA actions and S states formed using these samples. The optimal policy computed for this extended MDP is used throughout the epoch.

The posterior sampling for RL (PSRL) approach has been studied previously in Abbasi-Yadkori and Szepesvari [1], Osband and Van Roy [27], Osband et al. [28], and Ouyang et al. [30], but in a *Bayesian regret* framework. Bayesian regret is defined as the expected regret over a known prior on the transition probability matrix. Osband and Van Roy [27] demonstrate an $\tilde{O}(H\sqrt{SAT})$ bound on the Bayesian regret for PSRL in finite-horizon *episodic* MDPs, when the episode length is H. For the nonepisodic case, Ouyang et al. [30] propose an algorithm that achieves a regret bound of $\tilde{O}(HS\sqrt{AT})$, where H here is the bound on the span of the MDP. In this paper, we consider the stronger notion of *worst-case* regret, aka minimax regret, which requires bounding the maximum regret for any instance of the problem.[1] We consider a *nonepisodic communicating MDP* setting and prove a worst-case regret bound of $\tilde{O}(DS\sqrt{AT})$, where D is the unknown diameter of the communicating MDP. In comparison with PSRL that generates a single sample from the posterior, our *optimistic* PSRL algorithm is slightly inefficient as it generates multiple ($\tilde{O}(S)$) samples (although only once in every epoch). It is not entirely clear if the extra samples are only an artifact of the analysis. In an empirical study of a multiple sample version of posterior sampling for RL, Fonteneau et al. [16] show that multiple samples can potentially improve the performance of posterior sampling in terms of probability of taking the optimal decision. Our analysis uses some ideas from the Bayesian regret analysis. However, bounding the worst-case regret requires several new technical ideas, in particular, for proving optimism of the gain of the sampled MDP. Further discussion is provided in Section 4.

PSRL (and our optimistic PSRL) approaches are referred to as "model-based" approaches, because they explicitly estimate the transition probability matrix underlying the MDP model. Another closely related line of work investigates optimistic versions of "model-free algorithms" like value-iteration (Azar et al. [8]) and Q-learning (Kakade et al. [21]). However, the setting considered in both these works is that of an *episodic MDP*, where the learning agent interacts with the system in episodes of fixed and known length H. Under this setting, both the

previously mentioned works achieve minimax (i.e., worst-case) regret bound of $\tilde{O}(\sqrt{HSAT})$ when T is large enough compared with the episode length H. To understand the challenges in our setting compared with the episodic setting, although the initial state of each episode can be arbitrary in the episodic setting, importantly, the sequence of these initial states is shared by the algorithm and any benchmark policy. In contrast, in the nonepisodic setting considered in this paper, the state trajectory of the benchmark policy over T time steps can be completely different from the algorithm's trajectory. To the best of our understanding, the shared sequence of initial states of every episode, and the fixed known length H of episodes seems to form crucial components of the analysis in the episodic settings of Azar et al. [8] and Kakade et al. [21]. Thus, it would be difficult to extend such an analysis to the nonepisodic communicating MDP setting considered in this paper.

Among other related work, Burnetas and Katehakis [13] and Tewari and Bartlett [39] present optimistic linear programming approaches that achieve logarithmic regret bounds with problem dependent constants. Strong Probably Approximately Correct (PAC) bounds have been provided in Kearns and Singh [24], Brafman and Tennenholtz [10], Kakade [22], Asmuth et al. [7], and Dann and Brunskill [15]. There, the aim is to bound the performance of the policy learned at the end of the learning horizon and not the performance during learning as quantified here by regret. Notably, the BOSS algorithm proposed in Asmuth et al. [7] is similar to the algorithm proposed here in the sense that the former also takes multiple samples from the posterior to form an extended (referred to as *merged*) MDP. Strehl and Littman [36, 37] provide an optimistic algorithm for bounding regret in a discounted reward setting, but the definition of regret is different in that it measures the difference between the rewards of an optimal policy and the rewards of the learning algorithm *on the state trajectory taken by the learning algorithm*.

2. Preliminaries and Problem Definition

2.1. Communicating MDP

We consider an MDP $\mathcal{M}$ defined by tuple $\{\mathcal{S}, \mathcal{A}, P, r, s_1\}$, where $\mathcal{S}$ is a finite state-space of size S, $\mathcal{A}$ is a finite action-space of size A, $P : \mathcal{S} \times \mathcal{A} \to \Delta^{\mathcal{S}}$ is the transition model, $r : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the reward function, and s_1 is the starting state. When an action $a \in \mathcal{A}$ is taken in a state $s \in \mathcal{S}$, a reward $r_{s,a}$ is generated, and the system transitions to the next state $s' \in \mathcal{S}$ with probability $P_{s,a}(s')$, where $\sum_{s' \in \mathcal{S}} P_{s,a}(s') = 1$.

We consider communicating MDPs with finite diameter. Here we define communicating MDPs and recall some useful known results for such MDPs.

Definition 1 (Policy). A deterministic policy $\pi : \mathcal{S} \to \mathcal{A}$ is a mapping from state space to action space.

Definition 2 (Gain of a Policy). The gain $\lambda^{\pi}(s)$ of a policy π, from starting state $s_1 = s$, is defined as the infinite horizon undiscounted average reward, given by

$$\lambda^{\pi}(s) = \mathbb{E}\left[\lim_{T \to \infty} \frac{1}{T} \sum_{i=1}^{T} r_{s_t, \pi(s_t)} \,\middle|\, s_1 = s\right].$$

The variable s_t is the state reached at time t, on executing policy π.

Definition 3 (Diameter $D(\mathcal{M})$). Diameter $D(\mathcal{M})$ of an MDP $\mathcal{M}$ is defined as the minimum time required to go from one state to another in the MDP using some deterministic policy:

$$D(\mathcal{M}) = \max_{s \neq s', s, s' \in \mathcal{S}} \min_{\pi : \mathcal{S} \to \mathcal{A}} T^{\pi}_{s \to s'},$$

where $T^{\pi}_{s \to s'}$ is the expected number of steps it takes to reach state s' when starting from state s and using policy π.

Definition 4 (Communicating MDP). An MDP $\mathcal{M}$ is communicating if and only if it has a finite diameter. That is, for every pair of states $s, s'.s \neq s'$, there exists a policy π such that the expected number of steps to reach s' from s, $T^{\pi}_{s \to s'}$, is at most D, for some constant $D \geq 0$.

Lemma 1 (Optimal Gain and Bias for a Communicating MDP). *For a communicating MDP $\mathcal{M}$ with diameter D:*

(a) *Puterman* [31] *(theorems* 8.1.2 *and* 8.3.2*): The optimal (maximum) gain λ^* is state independent and is achieved by a deterministic stationary policy π^*, that is, there exists a deterministic policy π^* such that*

$$\lambda^* := \max_{s' \in \mathcal{S}} \max_{\pi} \lambda^{\pi}(s') = \lambda^{\pi^*}(s), \ \forall s \in \mathcal{S}.$$

Here, π^* is referred to as an optimal policy for MDP $\mathcal{M}$.

(b) *Bartlett and Tewari* [9] *(theorem 4): The optimal gain> λ^* satisfies the following equations:*

$$\lambda^* = \min_{h\in\mathbb{R}^S} \max_{s,a} r_{s,a} + P_{s,a}^T h - h_s = \max_a r_{s,a} + P_{s,a}^T h^* - h_s^*, \ \forall s, \tag{1}$$

where h^*, referred to as the bias vector of MDP $\mathcal{M}$, satisfies

$$\max_s h_s^* - \min_s h_s^* \le D.$$

Given these definitions and results, we can now define the RL problem studied in this paper.

2.2. RL Problem

The RL problem proceeds in rounds $t = 1, \ldots, T$. The learning agent starts from a state s_1 at round $t = 1$. In the beginning of every round t, the agent takes an action $a_t \in \mathcal{A}$ and observes the reward r_{s_t,a_t} and the next state $s_{t+1} \sim P_{s_t,a_t}$, where r and P are the reward function and the transition model, respectively, for a communicating MDP $\mathcal{M}$ with diameter D.

The learning agent knows the state-space $\mathcal{S}$, the action space $\mathcal{A}$, and the rewards $r_{s,a}$, $\forall s \in \mathcal{S}, a \in \mathcal{A}$, for the underlying MDP, but not the transition model P or the diameter D. (The assumption of known and deterministic rewards has been made here only for simplicity of exposition because the unknown transition model is the main source of difficulty in this problem. Our algorithm and results can be extended to bounded stochastic rewards with unknown distributions using standard Thompson sampling for MAB, e.g., using the techniques in Agrawal and Goyal [4].)

The agent can use the past observations to learn the underlying MDP model and decide future actions. The goal is to maximize the total reward $\sum_{t=1}^T r_{s_t,a_t}$, or equivalently, minimize the total regret over a time horizon T, defined as

$$\mathcal{R}(T, \mathcal{M}) := T\lambda^* - \sum_{t=1}^T r_{s_t,a_t}, \tag{2}$$

where λ^* is the optimal gain of MDP $\mathcal{M}$.

We present an algorithm for the learning agent with a near-optimal upper bound on the regret $\mathcal{R}(T, \mathcal{M})$ for any communicating MDP $\mathcal{M}$ with diameter D, thus bounding the worst-case regret over this class of MDPs.

3. Algorithm Description

Our algorithm combines the ideas of posterior sampling (aka Thompson sampling) with the extended MDP construction used in Jaksch et al. [20]. Here we first describe the main components of our algorithm. Our algorithm is then summarized as Algorithm 1.

Some notation: The term $N_{s,a}^t$ denotes the total number of times the algorithm visited state s and played action a until before time t, and $N_{s,a}^t(i)$ denotes the number of time steps among these $N_{s,a}^t$ steps where the next state was i, that is, the steps where a transition from state s to i was observed. We index the states from 1 to S, so that $\sum_{i=1}^S N_{s,a}^t(i) = N_{s,a}^t$ for any t. We use the symbol $\mathbf{1}$ to denote the vector of all ones and $\mathbf{1}_i$ to denote the vector with one at the ith coordinate and zero elsewhere.

3.1. Doubling Epochs

Our algorithm uses the epoch-based execution framework of Jaksch et al. [20]. An epoch is a group of consecutive rounds. The rounds $t = 1, \ldots, T$ are broken into consecutive epochs as follows: the kth epoch begins at the round τ_k immediately after the end of the $(k-1)$th epoch and ends at the first round τ such that for some state-action pair s, a, $N_{s,a}^\tau \ge 2N_{s,a}^{\tau_k}$. The algorithm computes a new policy $\tilde{\pi}_k$ at the beginning of every epoch k and uses that policy through all the rounds in that epoch. Because the total number of visits to any state action-pair is bounded by T, it is easy to observe that, irrespective of how the policies $\{\tilde{\pi}_k\}$ are computed, the number of epochs is bounded by $SA\log_2(T)$.

3.2. Posterior Sampling

We use posterior sampling to compute the policy $\tilde{\pi}_k$ in the beginning of every epoch k. The algorithm maintains a posterior distribution over the transition probability vector $P_{s,a}$, for every $s \in \mathcal{S}, a \in \mathcal{A}$. Observe that $P_{s,a}$ specifies a categorical distribution over states $1, \ldots, S$, with parameters $P_{s,a}(i), i = 1, \ldots, S$. Dirichlet distribution is a convenient choice for maintaining a posterior over parameters $P_{s,a}$, as Dirichlet distribution is a conjugate prior for the categorical distribution. In particular, it satisfies the following useful property: given a prior Dirichlet$(\alpha_1, \ldots, \alpha_S)$ on $P_{s,a}$,

after observing a transition from state s to i (with underlying probability $P_{s,a}(i)$), the posterior distribution is given by Dirichlet$(\alpha_1,\ldots,\alpha_i+1,\ldots,\alpha_S)$. By this property, for any $s\in\mathcal{S}, a\in\mathcal{A}$, on starting from the prior Dirichlet$(\mathbf{1})$ for $P_{s,a}$, the posterior at time t is Dirichlet$(\{N^t_{s,a}(i)+1\}_{i=1,\ldots,S})$.

A direct application of the PSRL approach introduced in Osband and Van Roy [27] would involve sampling a transition probability vector from the Dirichlet posterior for each state-action pair to form a sample MDP. A sample policy $\tilde{\pi}_k$ would then be computed as an optimal policy for the sampled MDP. Our algorithm uses a modified optimistic version of this approach. At the beginning of every epoch k, for every $s\in\mathcal{S}, a\in\mathcal{A}$ such that $N^{\tau_k}_{s,a}\geq\eta$, it generates *multiple* samples for $P_{s,a}$ from a *boosted* variance posterior. Specifically, for each s, a, it generates ψ independent sample probability vectors $Q^{1,k}_{s,a},\ldots,Q^{\psi,k}_{s,a}$ as

$$Q^{j,k}_{s,a}\sim \text{Dirichlet}(\mathbf{M}^{\tau_k}_{s,a}),$$

where $\mathbf{M}^t_{s,a}$ denotes the vector $[M^t_{s,a}(i)]_{i=1,\ldots,S}$, with

$$M^t_{s,a}(i):=\frac{1}{\kappa}(N^t_{s,a}(i)+\omega),\ \text{for } i=1,\ldots,S. \tag{3}$$

Here, ψ,κ,ω,η are parameters of the algorithm. The values of these parameters are initialized as $\eta=\sqrt{TS/A}+12\omega S^4$, $\psi=\Theta(S\log(SA/\rho))$, $\kappa=\Theta(\log(T/\rho))$, $\omega=\Theta(\log(T/\rho))$, given any $\rho\in(0,1]$. In the regret analysis, we derive sufficiently large constants to be used in the definition of ψ,κ,ω to guarantee the regret bounds. In particular, for our proofs we use $\psi=(2S/C)\log(SA/\rho)$ with the constant C defined as $C=\eta^{k(\delta)}/2$, where $\eta=0.15, \delta=((1-\Phi)(1/2))/2$, $k(\delta)=2.8^2/\delta^4$, with Φ being the cumulative distribution function of the standard normal distribution.

We remark that no attempt has been made to optimize this constant, and it is likely that a much smaller constant suffices.

3.3. Simple Optimistic Sampling

Posterior sampling is used for every $s\in\mathcal{S}, a\in\mathcal{A}$ with large enough previous visits, specifically those with $N^{\tau_k}_{s,a}\geq\eta$. For every remaining s, a, that is, those with $N^{\tau_k}_{s,a}<\eta$, the algorithm uses a simple optimistic sampling, as described in Algorithm 1. Intuitively, this process overestimates one randomly selected component of the vector $P_{s,a}$ while underestimating the remaining components. This special sampling has been introduced in the algorithm to handle a technical difficulty in analyzing the anti-concentration of Dirichlet posteriors when the parameters are very small. It is possible that with a different analysis technique, this is not required to achieve the regret bounds.

Algorithm 1 (Posterior Sampling–Based Algorithm for the RL Problem)

Inputs: State space $\mathcal{S}$, Action space $\mathcal{A}$, starting state s_1, reward function r, time horizon T, parameters $\rho\in(0,1]$.

Initialize: Set $\psi:=\frac{2S}{C}\log\left(\frac{SA}{\rho}\right), \eta:=\sqrt{\frac{TS}{A}}+12\omega S^4, \omega:=720\log(T/\rho), \kappa:=120\log(T/\rho), \tau^1:=1$.

for all epochs $k=1,2,\ldots,$ **do**

Sample transition probability vectors: For each s, a, generate ψ independent sample probability vectors $Q^{j,k}_{s,a}, j=1,\ldots,\psi$, as follows:

- **(Posterior sampling)**: For s, a such that $N^{\tau_k}_{s,a}\geq\eta$, sample from the Dirichlet distribution:

$$Q^{j,k}_{s,a}\sim\text{Dirichlet}(\mathbf{M}^{\tau_k}_{s,a}),$$

  with $M^{\tau_k}_{s,a(i)}, i\in\mathcal{S}$ as defined in (3).
- **(Simple optimistic sampling)**: For s, a such that $N^{\tau_k}_{s,a}<\eta$, use the following simple optimistic sampling: let

$$P^-_{s,a}:=\hat{P}_{s,a}-\boldsymbol{\Delta},$$

  where $\hat{P}_{s,a}(i):=\frac{N^{\tau_k}_{s,a}(i)}{N^{\tau_k}_{s,a}}$, and $\Delta_i:=\min\left\{\sqrt{\frac{4\hat{P}_{s,a}(i)\log(2ST)}{N^{\tau_k}_{s,a}}}+\frac{3\log(2ST)}{N^{\tau_k}_{s,a}}, \hat{P}_{s,a}(i)\right\}$, and let $\mathbf{z}$ be a random vector picked uniformly at random from $\{\mathbf{1}_1,\ldots,\mathbf{1}_S\}$; set

$$Q^{j,k}_{s,a}=P^-_{s,a}+\left(1-\sum_{i=1}^{S}P^-_{s,a}(i)\right)\mathbf{z}.$$

Compute policy $\tilde{\pi}^k$: as the optimal gain policy for extended MDP $\tilde{\mathcal{M}}^k$ constructed using sample set $\{Q^{j,k}_{s,a}, j=1,\ldots,\psi, s\in\mathcal{S}, a\in\mathcal{A}\}$.

Execute policy $\tilde{\pi}^k$:

for all time steps $t = \tau_k, \tau_k + 1, \ldots,$ until `break epoch` **do**
Play action $a_t = \tilde{\pi}_k(s_t)$.
Observe the transition to the next state s_{t+1}.
Set $N_{s,a}^{t+1}(i), M_{s,a}^{t+1}(i)$ for all $a \in \mathcal{A}, s, i \in \mathcal{S}$ as defined (refer to Equation (3)).
If $N_{s_t,a_t}^{t+1} \geq 2N_{s_t,a_t}^{\tau_k}$, then set $\tau_{k+1} = t + 1$ and `break epoch`.
end for
end for

3.4. Extended MDP

The policy $\tilde{\pi}_k$ used in epoch k is computed as the optimal policy of an *extended MDP* $\tilde{\mathcal{M}}^k$ defined by the sampled transition probability vectors. The construction of this extended MDP is derived from a similar construction in Jaksch et al. [20]. Given sampled vectors $\{Q_{s,a}^{j,k}, j = 1, \ldots, \psi, s \in \mathcal{S}, a \in \mathcal{A}\}$, we define an extended MDP $\tilde{\mathcal{M}}^k$ by extending the original action space as follows: for every s, a, create ψ actions for every action $a \in A$, denote by a^j the action corresponding to action a and sample j; then, in MDP $\tilde{\mathcal{M}}^k$, on taking action a^j in state s, reward is $r_{s,a}$ but the state transition follows the transition probability vector $Q_{s,a}^{j,k}$.

The algorithm uses the optimal policy $\tilde{\pi}_k$ of the extended MDP $\tilde{\mathcal{M}}^k$ whose action space is technically different from the action space $\mathcal{A}$ of MDP $\tilde{\mathcal{M}}^k$. We slightly abuse the notation to say that the algorithm takes action $a_t = \tilde{\pi}(s_t) \in \mathcal{A}$ to mean that the algorithm takes action $a_t = a \in \mathcal{A}$ when $\tilde{\pi}_k(s_t) = a^j$ for some j.

Our algorithm is summarized as Algorithm 1.

4. Regret Bounds

We prove that with high probability, the regret of Algorithm 1 is bounded by $\tilde{O}(DS\sqrt{AT})$.

Theorem 1. *For any communicating MDP $\mathcal{M}$ with S states, A actions, and diameter D, for $T \geq \Omega\left(SA\log^4(SAT/\rho)\right)$, the regret of Algorithm 1 is bounded as*

$$\mathcal{R}(T, \mathcal{M}) \leq O\left(DS\sqrt{AT}\log^2(SAT/\rho) + DS^3A^2\log^2(SAT/\rho)\right),$$

with probability $1 - \rho$, for any $0 < \rho \leq 1/16S^2$. For $T \geq \Omega(S^4A^3)$, this gives a regret bound of

$$\mathcal{R}(T, \mathcal{M}) \leq O\left(DS\sqrt{AT}\log^2(SAT/\rho)\right).$$

Here $O(\cdot)$ notation hides only the absolute constants.

Proof. Here we provide a proof of the previous theorem. The proofs of the lemmas used this proof are provided in the subsequent sections.

As defined in Section 2,

$$\mathcal{R}(T, \mathcal{M}) = T\lambda^* - \sum_{t=1}^{T} r_{s_t,a_t},$$

where λ^* is the optimal gain of MDP $\mathcal{M}$, a_t is the action taken, and s_t is the state reached by the algorithm at time t. Algorithm 1 proceeds in epochs $k = 1, 2, \ldots, K$, where $K \leq SA\log(T)$. To bound its regret in time T, we separately analyze the regret in each epoch k, namely,

$$\mathcal{R}_k := (\tau_{k+1} - \tau_k)\lambda^* - \sum_{t=\tau_k}^{\tau_{k+1}-1} r_{s_t,a_t}, \tag{4}$$

where τ_k was defined as the starting time step of epoch k. The proof of epoch regret bound has two main components:

(a) **Optimism:** Recall that in every epoch k, the algorithm runs an optimal gain policy for the extended MDP $\tilde{\mathcal{M}}^k$. We show that the extended MDP $\tilde{\mathcal{M}}^k$ is optimistic, that is, its optimal gain is (close to) λ^* or higher. Specifically, let $\tilde{\lambda}_k$ be the optimal gain of the extended MDP $\tilde{\mathcal{M}}^k$. In Lemma 3 (Section 5.2), which forms a main novel technical component of our proof, we show that with probability $1 - \rho$,

$$\tilde{\lambda}_k \geq \lambda^* - O\left(D\log^2(T/\rho)\sqrt{\frac{SA}{T}}\right).$$

Substituting this upper bound on λ^* in the expression for $\mathcal{R}_k$, we obtain the following bound on the epoch regret, with probability $1-\rho$:

$$\mathcal{R}_k \le \sum_{t=\tau_k}^{\tau_{k+1}-1}\left(\tilde{\lambda}_k - r_{s_t,a_t} + O\left(D\log^2(T/\rho)\sqrt{\frac{SA}{T}}\right)\right). \tag{5}$$

(b) **Deviation bounds:** Next, the first term in the previous expression is $\tilde{\lambda}_k$, which is the gain of the algorithm's chosen policy $\tilde{\pi}_k$ on MDP $\tilde{\mathcal{M}}^k$ (with transition probability vectors $\tilde{P}_{s,a} := Q^{j,k}_{s,a}$ for some j); and the second term is the reward obtained on executing the same policy $\tilde{\pi}_k$, but on the true MDP $\mathcal{M}$ (with transition probability vectors $P_{s,a}$). We bound the difference $\sum_t(\tilde{\lambda}_k - r_{s_t,a_t})$ by bounding the deviation $(\tilde{P}_{s,a} - P_{s,a})$ for every s, a.

We use the relation between the gain, the bias vector, and the reward vector of an optimal policy for a communicating MDP, as discussed in Section 2. To use this relation for MDP $\tilde{\mathcal{M}}^k$, we show that this MDP is communicating by comparing it to the true MDP $\mathcal{M}$, which was assumed to be communicating with diameter D. Specifically, in Lemma 6 (Section 5.4), we prove a bound of $2D$ on the diameter of MDP $\tilde{\mathcal{M}}^k$ for any k with probability $1-\rho$, when $T \ge \Omega(SA\log^4(SAT/\rho))$.

Therefore, we can use the relation between the gain $\tilde{\lambda}_k$, the bias vector $\tilde{h}$, and reward vector of optimal policy $\tilde{\pi}_k$ for communicating MDP $\tilde{\mathcal{M}}^k$, as given by Lemma 1(b). According to this relation, for any state s and action $a = \tilde{\pi}_k(s)$, $\tilde{\lambda}_k = r_{s,a} + \tilde{P}_{s,a}^T\tilde{h} - \tilde{h}_s$, where $\tilde{P}_{s,a} := Q^{j,k}_{s,a}$ for some j. Because $a_t = \tilde{\pi}_k(s_t)$, using this relation:

$$\begin{aligned}\sum_{t=\tau_k}^{\tau_{k+1}-1}(\tilde{\lambda}_k - r_{s_t,a_t}) &= \sum_{t=\tau_k}^{\tau_{k+1}-1}(\tilde{P}_{s_t,a_t} - \mathbf{1}_{s_t})^T\tilde{h}\\ &= \sum_{t=\tau_k}^{\tau_{k+1}-1}(\tilde{P}_{s_t,a_t} - P_{s_t,a_t} + P_{s_t,a_t} - \mathbf{1}_{s_t})^T\tilde{h}.\end{aligned} \tag{6}$$

In Lemma 4 (Section 5.3), we prove that with probability $1-\rho$, for all s, a, and all $h \in [0, 2D]^S$:

$$(\tilde{P}_{s,a} - P_{s,a})^T h \le O\left(D\left(\sqrt{\frac{S}{N^{\tau_k}_{s,a}}} + \frac{S}{N^{\tau_k}_{s,a}}\right)\log^2(SAT/\rho)\right). \tag{7}$$

We can use this result to bound first term in (6) by observing that $\tilde{h} \in \mathbb{R}^S$, the bias vector of MDP $\tilde{\mathcal{M}}^k$ satisfies (refer to Lemma 1),

$$\max_s \tilde{h}_s - \min_s \tilde{h}_s \le D(\tilde{\mathcal{M}}^k) \le 2D,$$

where the last inequality holds with probability $1-\rho$, as shown in Lemma 6 (Section 5.4).

To bound the second term of (6), we observe that $\mathbb{E}[\mathbf{1}^T_{s_{t+1}}\tilde{h} \mid \tilde{\pi}_k, \tilde{h}, s_t] = P^T_{s_t,a_t}\tilde{h}$ and use Azuma-Hoeffding inequality to obtain with probability $1-\rho$,

$$\sum_{t=\tau_k}^{\tau_{k+1}-1}(P_{s_t,a_t} - \mathbf{1}_{s_t})^T\tilde{h} \le O\left(D\sqrt{(\tau_{k+1}-\tau_k)\log(1/\rho)}\right). \tag{8}$$

Substituting the bounds from Equations (7) and (8) into (6), and combining it with (5), we obtain the following bound on $\mathcal{R}_k$ with probability $1-3\rho$:

$$\begin{aligned}\mathcal{R}_k = O\left(\left(D(\tau_{k+1}-\tau_k)\sqrt{\frac{SA}{T}} + D\sum_{s,a}(N^{\tau_{k+1}}_{s,a} - N^{\tau_k}_{s,a})\left(\frac{\sqrt{S}}{\sqrt{N^{\tau_k}_{s,a}}} + \frac{S}{N^{\tau_k}_{s,a}}\right)\right)\log^2\left(\frac{SAT}{\rho}\right)\right)\\ + O\left(D\sqrt{(\tau_{k+1}-\tau_k)\log\left(\frac{1}{\rho}\right)}\right).\end{aligned} \tag{9}$$

We observe that (by definition of an epoch) the number of visits of any state-action pair can at most double in an epoch,

$$N_{s,a}^{\tau_{k+1}} - N_{s,a}^{\tau_k} \le N_{s,a}^{\tau_k},$$

so that

$$D\sum_{s,a}(N_{s,a}^{\tau_{k+1}} - N_{s,a}^{\tau_k})\left(\frac{S}{N_{s,a}^{\tau_k}}\right) \le DS^2A.$$

Substituting this observation, we can bound the total regret $\mathcal{R}(T) = \sum_{k=1}^{K}\mathcal{R}_k$ as the following, with probability $1-3K\rho$:

$$\sum_{k=1}^{K}\mathcal{R}_k \le O\left(\sum_{k=1}^{K}\left(D(\tau_{k+1}-\tau_k)\sqrt{\frac{SA}{T}} + D\sum_{s,a}(N_{s,a}^{\tau_{k+1}} - N_{s,a}^{\tau_k})\left(\frac{\sqrt{S}}{\sqrt{N_{s,a}^{\tau_k}}}\right) + DS^2A\right)\log^2\left(\frac{SAT}{\rho}\right)\right)$$
$$+ O\left(\sum_{k=1}^{K}D\sqrt{(\tau_{k+1}-\tau_k)\log(1/\rho)}\right).$$

Applying Lemma B.5 (see Appendix) with $z_k = N_{s,a}^{\tau_{k+1}} - N_{s,a}^{\tau_k}$ and $Z_{k-1} = N_{s,a}^{\tau_k}$, it follows that

$$D\sum_{k=1}^{K}\sum_{s,a}(N_{s,a}^{\tau_{k+1}} - N_{s,a}^{\tau_k})\left(\frac{\sqrt{S}}{\sqrt{N_{s,a}^{\tau_k}}}\right) \le D\sqrt{S}\sum_{s,a}(\sqrt{2}+1)\sqrt{N_{s,a}^{\tau_K}}.$$

Substituting, we can bound the total regret as

$$\sum_{k=1}^{K}\mathcal{R}_k \le O\left(\left(D\sqrt{SAT} + D\sqrt{S}\left(\sum_{s,a}\sqrt{N_{s,a}^{\tau_K}}\right) + KDS^2A\right)\log^2\left(\frac{SAT}{\rho}\right) + D\sqrt{KT\log\left(\frac{1}{\rho}\right)}\right),$$

where we used that $\sum_k \tau_{k+1} - \tau_k = T$ and hence $\sum_{k=1}^{K}\sqrt{\tau_{k+1}-\tau_k} \le \sqrt{KT}$.

Now, because of our epoch definition, we have that $K \le SA\log(T)$, and because $\sum_{s,a}N_{s,a}^{\tau_K} \le T$, by simple worst scenario analysis, $\sum_{s,a}\sqrt{N_{s,a}^{\tau_K}} \le \sqrt{SAT}$. Thus, we obtain

$$\mathcal{R}(T,\mathcal{M}) \le O\left(DS\sqrt{AT}\log^2\left(\frac{SAT}{\rho}\right) + DS^3A^2\log^2\left(\frac{SAT}{\rho}\right)\right).$$

For $T \ge \Omega(S^4A^3)$, this gives a regret bound of

$$\mathcal{R}(T,\mathcal{M}) \le O\left(DS\sqrt{AT}\log^2(SAT/\rho)\right). \quad \square$$

5. Proofs of the Lemmas Used in Section 4

5.1. Notation

We use the following notations repeatedly in this section. Fix an epoch k, state s, action a, and sample j. The specific values of k,j,s,a will be clear from the context in a given proof. We denote $n = N_{s,a}^{\tau_k}$, $n_i = N_{s,a}^{\tau_k}(i)$ for all $i \in \mathcal{S}$, and $m = \frac{n+\omega S}{\kappa}$. Here $\omega = 720\log(T/\rho)$ and $\kappa = 120\log(T/\rho)$, as defined in the algorithm. Also, we denote $p_i = P_{s,a}(i)$, $\hat{p}_i := \frac{n_i}{n}$, $\bar{p}_i = \frac{n_i+\omega}{n+\omega S}$, and $\tilde{p}_i = Q_{s,a}^{j,k}(i)$, for $i \in \mathcal{S}$.

When $n > \eta$, the algorithm uses Dirichlet posterior sampling to generate sample vectors $Q_{s,a}^{j,k}$, so that in this case $\tilde{p}$ is a random vector distributed as Dirichlet$(m\bar{p}_1,\ldots,m\bar{p}_S)$.

When $n < \eta$, simple optimistic sampling is used, so that $\tilde{p}$ was generated as follows: denote

$$p^- = \left[\hat{p} - \left(\sqrt{\frac{4\hat{p}_i\log(2TS)}{n}} + \frac{3\log(2TS)}{n}\right)\mathbf{1}\right]^+,$$

and let $\mathbf{z}$ be a random vector picked uniformly at random from $\{\mathbf{1}_1, \ldots, \mathbf{1}_S\}$; then

$$\tilde{p} = p^- + \left(1 - \sum_j p_j^-\right)\mathbf{z}.$$

We define

$$\delta_i := \hat{p}_i - p_i,\ \Delta_i := \hat{p}_i - p_i^- = \min\left\{\sqrt{\frac{4\hat{p}_i \log(2ST)}{n}} + \frac{3\log(2ST)}{n}, \hat{p}_i\right\}.$$

Then, using Bernstein's inequality (Corollary B.1 in the Appendix, with $Z_t = \mathbb{1}(s_t = i, s_{t-1} = s, a_{t-1} = a)$, $t = 2, \ldots \tau_k$), we have that with probability $(1 - 1/2S)$, $|\delta_i| \le \sqrt{4\hat{p}_i \log(2TS)/n} + 3\log(2TS)/n$.

Therefore,

$$\sum_i \delta_i = 0, \sum_i \Delta_i = \sum_i (\hat{p}_i - p_i^-) = 1 - \sum_i p_i^-, \text{ and } \Delta_i \ge \delta_i \left(\text{with probability } 1 - \frac{1}{2S}\right).$$

The previous notations and observations will be used repeatedly in the proofs in this section.

5.2. Optimism

The goal of this section is to show optimism, that is,

$$\tilde{\lambda}_k \ge \lambda^* - \tilde{O}\left(D\sqrt{\frac{SA}{T}}\right).$$

First, in Lemma 2, we prove for any fixed vector, for every s, a, there exists a sample transition probability vector whose projection on that vector is optimistic, with high probability. To prove this, we prove the following fundamental new result on the anticoncentration of any fixed projection of a Dirichlet random vector.

Proposition 1. *Fix any vector $h \in \mathbb{R}^S$ such that $|h_i - h_{i'}| \le D$ for any i, i'. Consider a random vector $\tilde{p}$ generated from Dirichlet distribution with parameters $(m\bar{p}_1, \ldots, m\bar{p}_S)$, where $m\bar{p}_i \ge 6$. Then, for any $\rho \in (0,1)$, with probability at least $(C/S) - 2S\rho$,*

$$(\tilde{p} - \bar{p})^T h \ge \frac{1}{8}\sqrt{\sum_{i<S} \frac{\bar{\gamma}_i \bar{c}_i^2}{m}} - \frac{2SD\log(2/\rho)}{m},$$

where $\bar{\gamma}_i := \frac{\bar{p}_i(\bar{p}_{i+1} + \ldots + \bar{p}_S)}{(\bar{p}_i + \ldots + \bar{p}_S)}$, $\bar{c}_i = (h_i - \bar{H}_{i+1})$, $\bar{H}_{i+1} = \frac{1}{\sum_{j=i+1}^S \bar{p}_j} \sum_{j=i+1}^S h_j \bar{p}_j$, for any fixed ordering on states $i = 1, \ldots, S$. Also, let constant $C = \eta^{k(\delta)}/2$, where $\eta = 0.15, \delta = ((1 - \Phi)(1/2))/2$, $k(\delta) = 2.8^2/\delta^4$, with Φ being the cdf of the standard normal distribution.

The proof is provided in the Appendix. In the Appendix, we also prove the following strong concentration bound for the empirical probability vectors.

Proposition 2. *Fix any vector $h \in \mathbb{R}^S$ such that $|h_i - h_{i'}| \le D$ for any i, i'. Fix any epoch k, state s, action a. As defined in the notations section, denote $n = N_{s,a}^{\tau_k}$, $n_i = N_{s,a}^{\tau_k}(i)$, $p_i = P_{s,a}(i)$, $\hat{p}_i := \frac{n_i}{n}$, for $i = 1, \ldots, S$. Then, if $\tau_k - 1 \ge 96$, for any $\rho \in (0,1)$, with probability $1 - \rho$,*

$$|(\hat{p} - p)^T h| \le 2\sqrt{\log(T/\rho)\sum_{i<S} \frac{\gamma_i c_i^2}{n}} + 3D\frac{\log(T/\rho)}{n},$$

where $\gamma_i = \frac{p_i(p_{i+1} + \cdots + p_S)}{(p_i + \cdots + p_S)}$, $c_i = h_i - H_{i+1}$, $H_{i+1} = \frac{1}{\sum_{j=i+1}^S p_j} \sum_{j=i+1}^S h_j p_j$, for any fixed ordering on states $i = 1, \ldots, S$.

Together the previous two results allow us to prove the following lemma.

Lemma 2. *Fix any vector $h \in \mathbb{R}^S$ such that $|h_i - h_{i'}| \le D$ for any i, i', and any epoch k. Then, given $0 < \rho \le (1/16S^2)$, for every s, a, with probability at least $1 - \rho/SA$, there exists at least one j such that*

$$(Q_{s,a}^{j,k})^T h \ge P_{s,a}^T h - O\left(D \log^2(T/\rho)\sqrt{\frac{SA}{T}}\right).$$

Proof. Fix an epoch k, state and action pair s, a, and sample j. We use the notation defined in Section 5.1 so that $\tilde{p} = Q_{s,a}^{j,k}$, $p = P_{s,a}$, and so on. We show that with probability at least $C/S - 8S\rho$, $\tilde{p}^T h \ge p^T h - O\left(D \log^2(T/\rho)\sqrt{(SA/T)}\right)$. Now assuming $\rho \le (1/16S^2)$, given large enough number ψ (specifically, given $\psi \ge (2S/C)\log\left((SA/\rho)\right) = \Theta(S\log(SA/\rho))$) of independent samples for every s, a, this result will give us the lemma statement. To prove this result, we consider two cases.

Case 1: $n > \eta$. When $n > \eta$, Dirichlet posterior sampling is used so that $\tilde{p}$ is a random vector distributed as Dirichlet$(m\bar{p}_1, \ldots, m\bar{p}_S)$, where $m = \frac{n+\omega S}{\kappa}$, $\bar{p}_i = \frac{n_i+\omega}{n+\omega S}$. We show that with probability $\Omega(1/S) - 8S\rho$, the random quantity $\tilde{p}^T h$ exceeds its mean $\bar{p}^T h$ enough to overcome the possible deviation of empirical estimate $\bar{p}^T h$ from the true value $p^T h$. This involves combining the Dirichlet anticoncentration bound from Proposition 1 to lower bound $\tilde{p}^T h$ $(m\bar{p}_i \ge (\omega/\kappa) = 6,\ \forall i \in \mathcal{S})$ and the concentration bound on empirical estimates $\hat{p}$ from Proposition 2 to lower bound $\bar{p}^T h$ $(\tau_k - 1 \ge n \ge \eta \ge 96)$, which by definition is close to $\hat{p}^T h$.

In Proposition A.1 (in the Appendix), we prove a slight modification of Proposition 1 to show that with probability $C/S - 7S\rho$,

$$(\tilde{p} - \bar{p})^T h \ge 0.148\sqrt{\kappa \sum_i \frac{\gamma_i c_i^2}{n}} - O\left(\frac{DS\omega \log(T/\rho)}{n}\right). \tag{10}$$

The above bound replaces $\bar{\gamma}_i, \bar{c}_i, m$ in the lower bound provided by Proposition 1 by γ_i, c_i, n instead. With this modification, the lower bound becomes directly comparable to the bound on the deviation $|(\hat{p} - p)^T h|$ provided by Proposition 2. To combine this lower bound with the deviation bound, we calculate

$$|(\bar{p} - \hat{p})^T h| = \left|\sum_{i=1}^S h_i\left(\frac{n\hat{p}_i + \omega}{n + \omega S} - \frac{n\hat{p}_i}{n}\right)\right| = \left|\sum_i h_i\left(\frac{\omega(1 - S\hat{p}_i)}{n + \omega S}\right)\right| \le \frac{\omega DS}{n + \omega S} \le \frac{\omega DS}{n}.$$

Then, using the above bound along with (10) and the result from Proposition 2, we have that with probability $C/S - 8S\rho$,

$$\begin{aligned}
(\tilde{p} - p)^T h &= (\tilde{p} - \bar{p})^T h + (\bar{p} - \hat{p})^T h + (\hat{p} - p)^T h \\
&\ge (\tilde{p} - \bar{p})^T h - |(\bar{p} - \hat{p})^T h| - |(\hat{p} - p)^T h| \\
&\ge 0.148\sqrt{\kappa \sum_i \frac{\gamma_i c_i^2}{n}} - 2\sqrt{\log(T/\rho)\sum_{i<S} \frac{\gamma_i c_i^2}{n}} - O\left(\frac{DS\omega \log(T/\rho)}{n}\right) \\
&\ge -O\left(\omega \frac{DS \log(T/\rho)}{n}\right) \\
&\ge -O\left(D \log^2(T/\rho)\sqrt{\frac{SA}{T}}\right),
\end{aligned}$$

where the second last inequality follows from the observation that with $\kappa = 120\log(T/\rho)$, the first term is bigger than the second. Then, substituting $\omega = 720\log(T/\rho)$ and $n \ge \eta = \sqrt{TS/A} + 12\omega S^4$, we obtain the last inequality.

Case 2: $n < \eta$. When $n < \eta$, simple optimistic sampling is used. Using notation and observations made in Section 5.1, in this case $\tilde{p} = p^- + (1 - \sum_j p_j^-)\mathbf{z}$. With probability $1/S$, $z = \mathbf{1}_i$ for an i such that $h_i = \|h\|_\infty$, and (by union bound

over all i) with probability $1 - S\frac{1}{2S} = \frac{1}{2}$, $|\delta_i| \le (\sqrt{4\hat{p}_i \log(2TS)/n} + (3\log(2TS)/n))$ for every i. Therefore, with probability at least $1/2S$:

$$\begin{aligned}\sum_i \tilde{p}_i h_i &= \sum_i p_i^- h_i + \|h\|_\infty \left(1 - \sum_j p_j^-\right) = \sum_i p_i^- h_i + \|h\|_\infty \sum_j \Delta_j \\ &= \sum_i (\hat{p}_i - \Delta_i) h_i + \|h\|_\infty \Delta_i = \sum_i \hat{p}_i h_i + (\|h\|_\infty - h_i)\Delta_i \\ &\ge \sum_i \hat{p}_i h_i + (\|h\|_\infty - h_i)\delta_i = \sum_i (\hat{p}_i - \delta_i) h_i + \|h\|_\infty \delta_i \\ &= \sum_i p_i h_i + \|h\|_\infty \sum_i \delta_i = \sum_i p_i h_i. \quad \square\end{aligned}$$

Finally, we use the previous lemma to prove the main optimism lemma (Lemma 3).

Lemma 3 (Optimism). *For every epoch k, the optimal gain $\tilde{\lambda}_k$ of the extended MDP $\tilde{\mathcal{M}}^k$ satisfies*

$$\tilde{\lambda}_k \ge \lambda^* - O\left(D \log^2(T/\rho)\sqrt{\frac{SA}{T}}\right),$$

with probability $1 - \rho$, given small enough $\rho \le (1/16S^2)$, where λ^* the optimal gain of MDP $\mathcal{M}$ and D is the diameter.

Proof. Let h^* be the bias vector for an optimal policy π^* of MDP $\mathcal{M}$ (refer to Lemma 1 in the preliminaries section). Because h^* is a fixed (although unknown) vector with $|h_i - h_j| \le D$, we can apply Lemma 2 to obtain that with probability $1 - \rho$, for all s, a, there exists a sample vector $Q_{s,a}^{j,k}$ for some $j \in \{1, \ldots, \psi\}$ such that

$$(Q_{s,a}^{j,k})^T h^* \ge P_{s,a}^T h^* - \delta,$$

where $\delta = O\left(D \log^2(T/\rho)\sqrt{SA/T}\right)$. Now, consider the policy π for MDP $\tilde{\mathcal{M}}^k$ which for any s, takes action a^j, where $a = \pi^*(s)$, and j is a sample satisfying the previous inequality. Note that π is essentially π^* but defined for an MDP with a different transition probability matrix. Let Q_π be the transition matrix for policy π, with rows formed by the vectors $Q_{s,\pi^*(s)}^{j,k}$. Let P_{π^*} be the transition matrix whose rows are formed by the vectors $P_{s,\pi^*(s)}$. This implies

$$Q_\pi h^* \ge P_{\pi^*} h^* - \delta \mathbf{1}.$$

Let Q_π^* denote the limiting matrix for Markov chain with transition matrix Q_π. Observe that Q_π is aperiodic, recurrent, and irreducible: it is aperiodic and irreducible because each entry of Q_π being a sample from Dirichlet distribution is nonzero, and it is positive recurrent because in a finite irreducible Markov chain, all states are positive and recurrent. This implies that Q_π^* is of the form $\mathbf{1}\mathbf{q}^{*T}$, where $\mathbf{q}^*$ is the stationary distribution of Q_π, and $\mathbf{1}$ is the vector of all 1s (refer to (A.6) in Puterman [31]). Also, $Q_\pi^* Q_\pi = Q_\pi$, and $Q_\pi^* \mathbf{1} = \mathbf{1}$.

Therefore, the gain of policy π,

$$\tilde{\lambda}(\pi)\mathbf{1} = (r_\pi^T \mathbf{q}^*)\mathbf{1} = Q_\pi^* r_\pi,$$

where r_π is the S-dimensional vector $[r_{s,\pi(s)}]_{s=1,\ldots,S}$. Now,

$$\begin{aligned}\tilde{\lambda}(\pi)\mathbf{1} - \lambda^*\mathbf{1} &= Q_\pi^* r_\pi - \lambda^*\mathbf{1} \\ &= Q_\pi^* r_\pi - \lambda^*(Q_\pi^*\mathbf{1}) \ldots (\text{using } Q_\pi^*\mathbf{1} = \mathbf{1}) \\ &= Q_\pi^*(r_\pi - \lambda^*\mathbf{1}) \\ &= Q_\pi^*(I - P_{\pi^*})h^* \ldots (\text{using (1)}) \\ &= Q_\pi^*(Q_\pi - P_{\pi^*})h^* \ldots (\text{using } Q_\pi^* Q_\pi = Q_\pi^*) \\ &\ge -\delta\mathbf{1} \ldots (\text{using } (Q_\pi - P_{\pi^*})h^* \ge -\delta\mathbf{1}, Q_\pi^*\mathbf{1} = \mathbf{1}).\end{aligned}$$

Then, by optimality,

$$\tilde{\lambda}_k \geq \tilde{\lambda}(\pi) \geq \lambda^* - \delta. \quad \square$$

5.3. Deviation Bounds

Lemma 4. *In every epoch k, with probability $1-\rho$, for all samples j, all s, a, and all vectors $h \in [0,H]^S$,*

$$(Q_{s,a}^{j,k} - P_{s,a})^T h \leq O\left(H\left(\sqrt{\frac{S}{N_{s,a}^{\tau_k}}} + \frac{S}{N_{s,a}^{\tau_k}}\right)\log^2(SAT/\rho)\right).$$

Proof. Fix an s,a,j,k. Let $\tilde{p} = Q_{s,a}^{j,k}$. Denote $n = N_{s,a}^{\tau_k}$ and $m = n + \omega S/\kappa$, and $n_i = N_{s,a}^{\tau_k}(i)$, $\bar{p}_i := n_i + \omega/n + \omega S$ and $\hat{p}_i := \frac{n_i}{n}$ for $i = 1,\ldots,S$. Recall that $\eta = \sqrt{TS/A} + 12\omega S^4$ and $\omega = 720\log(T/\rho)$. It suffices to prove the lemma statement for $H = 1$. We consider two cases.

Case 1: When $n > \eta$, posterior sampling is used. Therefore, $\tilde{p}$ is an S-dimensional Dirichlet random vector with parameters $m\bar{p}_i, i = 1,\ldots,S$. Let X be distributed as Gaussian with mean $\mu = \bar{p}^T h$ and variance $\sigma^2 = \frac{1}{m}$. Now, for any *fixed* $h \in [0,1]^S$, by Gaussian-Dirichlet stochastic optimism (see Lemma B.1 in the Appendix):

$$X \succeq_{so} \tilde{p}^T h.$$

Then by Gaussian concentration (Corollary B.2), for any $\rho' \in (0,1)$, and fixed $h \in [0,1]^S$, with probability at least $1-\rho'$,

$$|\tilde{p}^T h - \bar{p}^T h| \leq \sqrt{\frac{2}{m}\log\left(\frac{2}{\rho'}\right)} \leq \sqrt{\frac{140}{n}\log\left(\frac{T}{\rho}\right)\log\left(\frac{2}{\rho'}\right)}. \tag{11}$$

In the last inequality, we substituted $m \geq n/\kappa$, with $\kappa = 120\log(T/\rho)$. In Proposition 2, we proved a strong upper bound on $|\hat{p}^T h - p^T h|$ for any fixed $h \in [0,1]^S$, which was used for proving optimism. A corollary of that concentration bound (by using observations that $\gamma_i = \frac{p_i(p_{i+1}+\cdots+p_S)}{(p_i+\cdots+p_S)} \leq p_i$, and $|c_i| \leq 1$ when $h \in [0,1]^S$) is that for any $\rho' \in (0,1)$, and fixed $h \in [0,1]^S$ with probability $1-\rho'$,

$$|(\hat{p}-p)^T h| \leq 2\sqrt{\frac{\log(T/\rho')}{n}} + \frac{3\log(2/\rho')}{n}. \tag{12}$$

Also, for *all* $h \in [0,1]^S$:

$$|\hat{p}^T h - \bar{p}^T h| \leq \sum_i \left|\frac{n_i+\omega}{n+\omega S} - \frac{n_i}{n}\right| \leq \sum_i \frac{\omega S n_i}{(n+\omega S)n} \leq \frac{\omega S}{n}, \tag{13}$$

where $\omega = 720\log(T/\rho)$. Combine the bounds from Equations (11), (12), and (13) and take union bound over all fixed h on an ϵ-grid over $[0,1]^S$, with $\epsilon = 1/n$. Then, substituting ρ' by ρ'/n^S, we have that with probability $1-\rho'$, for all $h \in [0,1]^S$,

$$|\tilde{p}^T h - p^T h| \leq 14\sqrt{\frac{S\log(T/\rho')\log(T/\rho)}{n}} + 5\frac{S\log(T/\rho')}{n} + 2\frac{\omega S}{n}. \tag{14}$$

Previously, we used that for all h' not on the ϵ-grid, $\|h'-h\|_\infty \leq \epsilon = (1/n)$, so that $|\tilde{p}^T h' - p^T h'| \leq |\tilde{p}^T h - p^T h| + (1/n)$ for some h on the ϵ-grid.

Case 2: When $n \leq \eta$, simple optimistic sampling is used. Using notation in Section 5.1, in this case $\tilde{p} = p^- + (1 - \sum_{i=1}^S p_i^-)\mathbf{z}$, where $\mathbf{z}$ is a random vector picked uniformly at random from $\{\mathbf{1}_1,\ldots,\mathbf{1}_S\}$. Using Bernstein's

inequality (Corollary B.1) to bound $(\hat{p}-p)$, we have for any $\rho'' \in (0,1)$, with probability $1-\rho''$, for all $h \in [0,1]^S$:

$$\begin{aligned}(\tilde{p}^T h - p^T h) &\le (\hat{p}^T h - p^T h) + \sum_i \sqrt{\frac{3\hat{p}_i \log(4S)}{n}} + \sum_i \frac{3\log(4S)}{n} \\ &\le \|\hat{p}-p\|_1 \|h\|_\infty + \sqrt{S\frac{3\log(4S)}{n}} + \frac{3S\log(4S)}{n} \\ &\le \sqrt{\frac{4S\log(T/\rho'')}{n}} + \frac{3S\log(T/\rho'')}{n} + \sqrt{S\frac{3\log(T)}{n}} + \frac{3S\log(4S)}{n} \\ &= 4\sqrt{\frac{S\log(T/\rho'')}{n}} + \frac{3S\log(4ST/\rho'')}{n}. \end{aligned} \tag{15}$$

Equations (14) and (15) provide a bound on $|(Q_{s,a}^{j,k})^T \tilde{h} - P_{s,a}^T \tilde{h}|$ for any given s,a,j,k. Substituting $\rho' = \rho'' = \rho/(SA\psi)$, and taking a union bound over all possible values of s, a, j, we get the lemma statement. (Here $\psi = \Theta(S\log(SA/\rho))$.) □

5.4. Diameter of the Extended MDP

Algorithm 1 computes policy $\tilde{\pi}_k$ in epoch k as an optimal gain policy of the extended MDP $\tilde{\mathcal{M}}^k$. Our goal in this section is to prove that the diameter of $\tilde{\mathcal{M}}^k$ is within a constant factor of the diameter $\mathcal{M}$. We begin by deriving a bound on the diameter of $\tilde{\mathcal{M}}^k$ under certain conditions and then prove that those conditions hold with high probability.

Lemma 5. *For any state $s \in \mathcal{S}$, let $E^s \in \mathbb{R}_+^S$ be the vector of the minimum expected times to reach s from $s' \in \mathcal{S}$ in true MDP $\mathcal{M}$, that is, $E^s_{s'} = \min_\pi T^\pi_{s' \to s}$. Note that $E^s_s = 0$. For any episode k, if for every s, a there exists some j such that*

$$Q_{s,a}^{j,k} \cdot E^s \le P_{s,a} \cdot E^s + \delta, \tag{16}$$

for some $\delta \in [0,1)$, then the diameter of extended MDP $\tilde{\mathcal{M}}^k$ is at most $D/(1-\delta)$, where D is the diameter of MDP $\mathcal{M}$.

Proof. Fix an epoch k. For brevity, we omit the superscript k later.

Fix any two states $s_1 \neq s_2$. We prove the lemma statement by constructing a policy $\tilde{\pi}$ for $\tilde{\mathcal{M}}$ such that the expected time to reach s_2 from s_1 is at most $D/(1-\delta)$. Let π be the policy for MDP $\mathcal{M}$ that minimizes the expected time to reach s_2 from other states. Therefore, the time to reach s_2 from s_1 using π is at most D (because $\mathcal{M}$ has diameter D). Let E be the $|S|-1$-dimensional vector of expected times to reach s_2 from every state, except s_2 itself, using π (E is the subvector formed by removing s_2^{th} coordinate of vector E^{s_2} where E^s was defined in the lemma statement. $E^{s_2}_{s_2} = 0$). By first step analysis, E is a solution of

$$E = \mathbf{1} + P_\pi^\dagger E,$$

where $P_\pi^\dagger$ is defined as the $(S-1)\times(S-1)$ transition matrix for policy π in MDP $\mathcal{M}$, with the $(s,s')^{th}$ entry being the transition probability $P_{s,\pi(s)}(s')$ for all $s,s' \neq s_2$. Also, by choice of π, E satisfies

$$E_{s_1} \le D.$$

Now, we define $\tilde{\pi}$ using π as follows: For any state $s \neq s_2$, let $a = \pi(s)$ and jth sample satisfies the property (16) for s,a,E^{s_2}, then we define $\tilde{\pi}(s) := a^j$. Let $Q_{\tilde{\pi}}$ be the transition matrix (dimension $S \times S$) for this policy.

The term $Q_{\tilde{\pi}}$ defines a Markov chain. Next, we modify this Markov chain to construct an absorbing Markov chain with a single absorbing state s_2. Let $Q_{\tilde{\pi}}^\dagger$ be the $(S-1)\times(S-1)$ submatrix of $Q_{\tilde{\pi}}$ obtained by removing the row and column corresponding to the state s_2. Then we define Q' as (an appropriate reordering of) the following matrix:

$$Q'_{\tilde{\pi}} = \begin{bmatrix} Q_{\tilde{\pi}}^\dagger & \mathbf{q} \\ \mathbf{0} & 1 \end{bmatrix},$$

where $\mathbf{q}$ is an $(S-1)$-length column vector such that the rows of $Q'_{\tilde{\pi}}$ sum to one. Because the probabilities in $Q_{\tilde{\pi}}$ were drawn from Dirichlet distribution, they are all strictly greater than zero and less than one. Therefore, each row-sum of $Q_{\tilde{\pi}}^\dagger$ is strictly less than one, so that the vector $\mathbf{q}$ has no zero entries, and the Markov chain is indeed

an absorbing chain with single absorbing state s_2. Then we notice that $(I - Q^{\dagger}_{\tilde{\pi}})^{-1}$ is precisely the fundamental matrix of this absorbing Markov chain and hence exists and is nonnegative (Grinstead and Snell [19], theorem 11.4). Let $\tilde{E}$ be defined as the $(S-1)$-dimensional vector of expected time to reach s_2 from $s' \neq s_2$ in MDP $\tilde{\mathcal{M}}^k$ using $\tilde{\pi}$. Then, it is the same as the expected time to reach the absorbing state s_2 from $s' \neq s_2$ in the Markov chain $Q'_{\tilde{\pi}}$, given by

$$\tilde{E} = (I - Q^{\dagger}_{\tilde{\pi}})^{-1}\mathbf{1}.$$

Then, using (16) (because $E^{s_2}_{s_2} = 0$, the inequality holds for $P^{\dagger}, Q^{\dagger}$),

$$E = \mathbf{1} + P^{\dagger}_{\pi}E \geq \mathbf{1} + Q^{\dagger}_{\tilde{\pi}}E - \delta\mathbf{1} \Rightarrow (I - Q^{\dagger}_{\tilde{\pi}})E \geq (1-\delta)\mathbf{1}. \tag{17}$$

Multiplying the nonnegative matrix $(I - Q^{\dagger}_{\tilde{\pi}})^{-1}$ on both sides of this inequality, it follows that

$$E \geq (1-\delta)(I - Q^{\dagger}_{\tilde{\pi}})^{-1}\mathbf{1} = (1-\delta)\tilde{E},$$

so that $\tilde{E}_{s_1} \leq (1/(1-\delta))E_{s_1} \leq D/(1-\delta)$, proving that the expected time to reach s_2 from s_1 using policy $\tilde{\pi}$ in MDP $\tilde{\mathcal{M}}^k$ is at most $D/(1-\delta)$. □

Now we can use the previous result to prove that the diameter of the extended MDP is bounded by twice the diameter of the original MDP.

Lemma 6. *Assume $T \geq 2\bar{C}SA\log^4(SAT/\rho)$ with some large enough constant $\bar{C}$, and $0 < \rho \leq (1/16S^2)$. Then, for any epoch k, the diameter of MDP $\tilde{\mathcal{M}}^k$ is bounded by $2D$, with probability $1-\rho$.*

Proof. Fix an epoch k. For any state s, let E^s be as defined in Lemma 5. We show that with probability $1-\rho$, for all s, a, there exists some j with $Q^{j,k}_{s,a} \cdot E^s \leq P_{s,a} \cdot E^s + \delta$, with $\delta \leq 1/2$. This will allow us to apply Lemma 5 to bound the diameter of $\tilde{\mathcal{M}}^k$.

Given any s, a, j, k, we use notations and observations from Section 5.1, so that $\tilde{p} = Q^{j,k}_{s,a}, p = P_{s,a}$ and so on. Also, let $h = E^s$. Then, $\min_i h_i = 0, \max_i h_i = D$.

First consider all s, a with $n > \eta$. Using (14) (in the proof of Lemma 4), we have

$$\tilde{p}^T h - p^T h \leq 14D\sqrt{\frac{S\log(T/\rho')\log(T/\rho)}{n}} + 5D\frac{S\log(T/\rho')}{n} + 2D\frac{\omega S}{n},$$

with probability $1-\rho'$ for any $\rho' \in (0,1)$. Substituting $\rho' = \rho/(2SA\psi)$, we get that with probability $1-\frac{\rho}{2}$, for all s, a, j such that $n > \eta$, $\tilde{p}^T h - p^T h \leq \delta$, where $\delta = 14D\sqrt{2\log^2(SAT/\rho)/\eta} + 5D(2S\log(SAT/\rho'))/\eta + (2\omega SD/\eta)$. Then, using $\eta = \sqrt{TS/A} + 12\omega S^4$, and $T \geq \bar{C}SA\log^4(SAT/\rho)$ (for some constant $\bar{C}$), we get $\delta \leq 1/2$. Although no attempt has been made to optimize constants, we note that $\bar{C} \geq 28^4$ is sufficient.

For s, a such that $n \leq \eta$, simple optimistic sampling is used. Using notations introduced in Section 5.1, in this case $\tilde{p} = p^- + (1 - \sum_j p^-_j)\mathbf{z}$, where $\mathbf{z}$ is a random vector picked uniformly at random from $\{\mathbf{1}_1, \ldots, \mathbf{1}_S\}$. With probability $1/S$, $z = \mathbf{1}_i$ for i such that $h_i = \min_i h_i = 0$. Therefore, with probability at least $1/2S$:

$$\tilde{p}^T h = (p^-)^T h = \sum_i (\hat{p}_i - \Delta_i)h_i \leq \sum_i (\hat{p}_i - \delta_i)h_i = p^T h.$$

Because we have $\psi \geq (2S/C)\log(SA/\rho)$ independent samples for every s, a, with probability $1-(\rho/2)$, there exists at least one sample j such that $Q^{j,k}_{s,a} \cdot h \geq P_{s,a} \cdot h$.

Therefore, we have shown that with probability $1-\rho$, for all s, a, there exists some j such that $Q^{j,k}_{s,a} \cdot E^s \leq P_{s,a} \cdot E^s + \delta$, with $\delta \leq 1/2$. By Lemma 5, we obtain that the diameter of $\tilde{\mathcal{M}}^k$ is bounded by $D/(1-\delta) \leq 2D$ with probability $1-\rho$. □

6. Conclusions

We presented an algorithm inspired by posterior sampling that achieves near-optimal worst-case regret bounds for the reinforcement learning problem with communicating MDPs in a nonepisodic, undiscounted average reward setting. Our algorithm may be viewed as a randomized version of the UCRL2 algorithm of Jaksch et al. [20], with randomization via posterior sampling. Our analysis demonstrates that posterior sampling provides adequate amount of uncertainty in the samples, so that an optimistic policy can be obtained without excess overestimation.

Although our work surmounts some important technical difficulties in obtaining worst-case regret bounds for posterior sampling–based algorithms for communicating MDPs, the provided bound does not achieve optimal regret in S and D. Obtaining a better worst-case regret bound remains an open question. In particular, we believe that studying value functions may improve the dependence on S in the regret bound, possibly for large T (Azar et al. [8] produce an $\tilde{O}(\sqrt{HSAT})$ bound when $T \geq H^3S^3A$). We also leave as future work the analysis regarding the necessity of multiple posterior samples; in particular, whether the number of samples required in every epoch can be reduced from $\tilde{O}(S)$ to $\tilde{O}(\log(S))$ or a constant.

Acknowledgments

The authors thank Tor Lattimore for pointing out a mistake in an earlier version of this work and Ian Osband for the fruitful discussions toward resolving the said mistake.

Appendix A. Missing Proofs from Section 5.2

A.1. Anti-Concentration of Dirichlet Distribution: Proof of Proposition 1

We prove the following general result on anti-concentration of Dirichlet distributions, which will be used to prove optimism.

Proposition 1 (Restated from Main Text). *Fix any vector $h \in \mathbb{R}^S$ such that $|h_i - h_{i'}| \leq D$ for any i, i'. Consider a random vector $\tilde{p}$ generated from Dirichlet distribution with parameters $(m\bar{p}_1, \ldots, m\bar{p}_S)$, where $m\bar{p}_i \geq 6$. Then, for any $\rho \in (0,1)$, with probability at least $(C/S) - 2S\rho$,*

$$(\tilde{p} - \bar{p})^T h \geq \frac{1}{8}\sqrt{\sum_{i<S} \frac{\bar{\gamma}_i \bar{c}_i^2}{m} - \frac{2SD\log(2/\rho)}{m}},$$

where $\bar{\gamma}_i := \frac{\bar{p}_i(\bar{p}_{i+1}+\ldots+\bar{p}_S)}{(\bar{p}_i+\ldots+\bar{p}_S)}$, $\bar{c}_i = (h_i - \bar{H}_{i+1})$, $\bar{H}_{i+1} = \frac{1}{\sum_{j=i+1}^S \bar{p}_j}\sum_{j=i+1}^S h_j\bar{p}_j$, for any fixed ordering on states $i = 1, \ldots, S$. Also, let constant $C = (\eta^{k(\delta)}/2)$, where $\eta = 0.15, \delta = ((1-\Phi)((1/2))/2, k(\delta) = (2.8^2/\delta^4)$, with Φ being the cdf of the standard normal distribution.

We use an equivalent representation of a Dirichlet vector in terms of independent Beta random variables.

Fact A.1. Fix an ordering of indices $1, \ldots, S$, and define $\tilde{y}_i := \frac{\tilde{p}_i}{\tilde{p}_i+\cdots+\tilde{p}_S}, \bar{y}_i := \frac{\bar{p}_i}{\bar{p}_i+\cdots+\bar{p}_S}$. Then, for any $h \in \mathbb{R}^S$,

$$(\tilde{p} - \bar{p})^T h = \sum_i (\tilde{y}_i - \bar{y}_i)(h_i - \tilde{H}_{i+1})(\bar{p}_i + \cdots + \bar{p}_S) = \sum_i (\tilde{y}_i - \bar{y}_i)(h_i - \bar{H}_{i+1})(\tilde{p}_i + \cdots + \tilde{p}_S),$$

where $\tilde{H}_{i+1} = \frac{1}{\sum_{j=i+1}^S \tilde{p}_j}\sum_{j=i+1}^S h_j\tilde{p}_j$, $\bar{H}_{i+1} = \frac{1}{\sum_{j=i+1}^S \bar{p}_j}\sum_{j=i+1}^S h_j\bar{p}_j$.

Proof. We will prove the first equality, that is,

$$(\tilde{p} - \bar{p})^T h = \sum_i (\tilde{y}_i - \bar{y}_i)(h_i - \tilde{H}_{i+1})(\bar{p}_i + \cdots + \bar{p}_S). \tag{A.1}$$

The second equality follows analogously from the same proof steps. After substituting $\tilde{H}_{i+1}$ in the right-hand side, the coefficient of h_k for any k is given by

$$\tilde{p}_k\left(\frac{\bar{p}_k + \cdots + \bar{p}_S}{\tilde{p}_k + \cdots + \tilde{p}_S}\right) - \bar{p}_k - \sum_{i=1}^{k-1}(\tilde{y}_i - \bar{y}_i)\left(\frac{\tilde{p}_k}{\tilde{p}_{i+1} + \cdots + \tilde{p}_S}\right)(\bar{p}_i + \cdots + \bar{p}_S).$$

Therefore, to prove (A.1) holds, it suffices to show that for $k = 1, 2, \ldots, S$,

$$\sum_{i=1}^{k-1} \frac{(\bar{y}_i - \tilde{y}_i)(\bar{p}_i + \ldots + \bar{p}_S)}{\tilde{p}_{i+1} + \ldots + \tilde{p}_S} = 1 - \frac{\bar{p}_k + \ldots + \bar{p}_S}{\tilde{p}_k + \cdots + \tilde{p}_S}. \tag{A.2}$$

We prove (A.2) by induction on k, because $\tilde{p}, \bar{p}$ are probability vectors and hence sum to one. The case $k = 1$ clearly holds. For $k = 2$, we have that the left-hand side of (A.2) becomes

$$\frac{\bar{p}_1 - \tilde{p}_1}{\tilde{p}_2 + \ldots + \tilde{p}_S} = \frac{\bar{p}_1 - \tilde{p}_1}{1 - \tilde{p}_1} = \frac{1 - \tilde{p}_1 - (1 - \bar{p}_1)}{1 - \tilde{p}_1} = 1 - \frac{\bar{p}_2 + \ldots + \bar{p}_S}{\tilde{p}_2 + \ldots \tilde{p}_S}.$$

Now, assume (A.2) holds up to $k - 1$, and we will prove it holds for k. By the induction hypothesis and some algebra,

$$\begin{aligned}
\sum_{i=1}^{k} \frac{(\bar{y}_i - \tilde{y}_i)(\bar{p}_i + \ldots + \bar{p}_S)}{\tilde{p}_{i+1} + \ldots + \tilde{p}_S} &= \sum_{i=1}^{k-1} \frac{(\bar{y}_i - \tilde{y}_i)(\bar{p}_i + \ldots + \bar{p}_S)}{\tilde{p}_{i+1} + \ldots + \tilde{p}_S} + \frac{(\bar{y}_k - \tilde{y}_k)(\bar{p}_k + \ldots + \bar{p}_S)}{\tilde{p}_{k+1} + \ldots + \tilde{p}_S} \\
&= 1 - \frac{\bar{p}_k + \ldots + \bar{p}_S}{\tilde{p}_k + \ldots + \tilde{p}_S} + \frac{(\bar{y}_k - \tilde{y}_k)(\bar{p}_k + \ldots + \bar{p}_S)}{\tilde{p}_{k+1} + \ldots + \tilde{p}_S} \\
&= 1 - \frac{\bar{p}_k + \ldots + \bar{p}_S}{\tilde{p}_k + \ldots + \tilde{p}_S} + \left(\frac{\bar{p}_k}{\bar{p}_k + \ldots + \bar{p}_S} - \frac{\tilde{p}_k}{\tilde{p}_k + \ldots + \tilde{p}_S} \right) \frac{\bar{p}_k + \ldots + \bar{p}_S}{\tilde{p}_{k+1} + \ldots + \tilde{p}_S} \\
&= 1 - \frac{\bar{p}_k + \ldots + \bar{p}_S}{\tilde{p}_k + \ldots \tilde{p}_S} + \frac{\bar{p}_k}{\tilde{p}_{k+1} + \ldots + \tilde{p}_S} - \frac{\tilde{p}_k(\bar{p}_k + \ldots + \bar{p}_S)}{(\tilde{p}_k + \ldots + \tilde{p}_S)(\tilde{p}_{k+1} + \ldots + \tilde{p}_S)} \\
&= 1 + \left(\frac{\bar{p}_k(\tilde{p}_k + \ldots + \tilde{p}_S) - (\bar{p}_k + \ldots \bar{p}_S)(\tilde{p}_{k+1} + \ldots + \tilde{p}_S) - \tilde{p}_k(\bar{p}_k + \ldots + \bar{p}_S)}{(\tilde{p}_k + \ldots + \tilde{p}_S)(\tilde{p}_{k+1} + \ldots + \tilde{p}_S)} \right) \\
&= 1 + \left(\frac{\bar{p}_k(\tilde{p}_k + \ldots + \tilde{p}_S) - (\bar{p}_k + \ldots \bar{p}_S)(\tilde{p}_k + \ldots + \tilde{p}_S)}{(\tilde{p}_k + \ldots + \tilde{p}_S)(\tilde{p}_{k+1} + \ldots + \tilde{p}_S)} \right) \\
&= 1 - \frac{\bar{p}_{k+1} + \ldots + \bar{p}_S}{\tilde{p}_{k+1} + \ldots + \tilde{p}_S}
\end{aligned}$$

as desired. □

The subsequent fact follows from a known property of Dirichlet distributions (see section 1.4 of Albert and Denis [6]).

Fact A.2. For $i = 1, \ldots, S$, $\tilde{y}_i := \tilde{p}_i / \tilde{p}_i + \cdots + \tilde{p}_S$ are independent Beta random variables distributed as $\text{Beta}(m\bar{p}_i, m(\bar{p}_{i+1} + \cdots + \bar{p}_S))$, with mean

$$\mathbb{E}[\tilde{y}_i] = \frac{m\bar{p}_i}{m(\bar{p}_i + \cdots + \bar{p}_S)} = \bar{y}_i,$$

and variance

$$\bar{\sigma}_i^2 := \mathbb{E}[(\tilde{y}_i - \bar{y}_i)^2] = \frac{\bar{p}_i(\bar{p}_{i+1} + \cdots + \bar{p}_S)}{(\bar{p}_i + \cdots + \bar{p}_S)^2 (m(\bar{p}_i + \cdots + \bar{p}_S) + 1)}.$$

We derive the following basic anti-concentration inequality for Beta random variables.

Lemma A.1 (Anti-Concentration for Beta Random Variables). *Let $F_{a,b}$ denote the cdf of a Beta random variable with parameter (a, b), with $a \geq 6, b \geq 6$. Let*

$$z = \frac{a}{a+b} + C\sqrt{\frac{ab}{(a+b)^2(a+b+1)}} + \frac{C}{a+b},$$

with $C \leq 0.5$. Then,

$$1 - F_{(a,b)}(z) \geq 1 - \Phi(1) - 0.005 \geq 0.15.$$

Proof. Let $x = C\sqrt{\frac{ab}{(a+b+1)}} + C$. Then, $z = \frac{a+x}{a+b}$, $w_1 = (b(a+x)/(a+b))^{1/3}$ and $w_2 = [a(b-x)/(a+b))]^{1/3}$. Also, $z \le 2C\sqrt{\frac{ab}{a+b}}$. Also,

$$(a+b-1)(1-z) \ge (a+b-1)\left(1-\frac{a}{a+b}-C\sqrt{\frac{ab}{(a+b)^2(a+b+1)}}-\frac{C}{a+b}\right) = (a+b-1)\left(\frac{b}{a+b}-\frac{C}{a+b}\sqrt{\frac{ab}{a+b+1}}-\frac{C}{a+b}\right)$$
$$\ge \frac{a+b-1}{a+b}\left(b-C\sqrt{\frac{ab}{a+b+1}}-\frac{C}{a+b}\right) \ge \frac{11}{12}\left(b-C\sqrt{b}-\frac{C}{12}\right) \ge 0.8.$$

Hence, we can apply Fact B.5 relating Beta with Normal. We bound the numerator and denominator in the expression of y to show that the relation $I_z(a,b) \le \Phi(y) + \epsilon$ holds for some $y \le 1$:

$$\begin{aligned}
numerator(y) &= 3\left[w_1\left(1-\frac{1}{9b}\right)-w_2\left(1-\frac{1}{9a}\right)\right]\\
&= 3\left(\frac{ab}{a+b}\right)^{\frac{1}{3}}\left[\left(1+\frac{x}{a}\right)^{\frac{1}{3}}\left(1-\frac{1}{9b}\right)-\left(1-\frac{x}{b}\right)^{\frac{1}{3}}\left(1-\frac{1}{9a}\right)\right]\\
&\le 3\left(\frac{ab}{a+b}\right)^{\frac{1}{3}}\left[\left(1+\frac{x}{3a}\right)\left(1-\frac{1}{9b}\right)-\left(1-\frac{x}{3b}-\frac{2x^2}{9b^2}\right)\left(1-\frac{1}{9a}\right)\right]\\
&= 3\left(\frac{ab}{a+b}\right)^{\frac{1}{3}}\left[\left(\frac{b-a}{9ab}\right)+\left(\frac{x(a+b)}{3ab}\right)-\left(\frac{2x}{27ab}\right)\right]+3\left(\frac{ab}{a+b}\right)^{\frac{1}{3}}\left[\frac{2x^2}{9b^2}\left(1-\frac{1}{9a}\right)\right]\\
&\le 3\left(\frac{ab}{a+b}\right)^{\frac{1}{3}}\left[\left(\frac{b-a}{9ab}\right)+\left(\frac{x(a+b)}{3ab}\right)\right]+3\left(\frac{ab}{a+b}\right)^{\frac{1}{3}}\left[\frac{2x^2}{9b^2}\left(1-\frac{1}{9a}\right)\right]\\
&= \left(\frac{ab}{a+b}\right)^{\frac{1}{3}}\left(\frac{a+b}{ab}\right)\left[\left(\frac{b-a}{3(a+b)}\right)+x+\frac{2x^2}{3b^2}\left(1-\frac{1}{9a}\right)\right]\\
&\le \left(\frac{ab}{a+b}\right)^{\frac{1}{3}}\left(\frac{a+b}{ab}\right)\left[\left(\frac{b-a}{3(a+b)}\right)+\frac{2x^2}{3b^2}\left(1-\frac{1}{9a}\right)+C+C\left(\frac{ab}{a+b}\right)^{\frac{1}{2}}\right]\\
&\le \left(\frac{b-a}{3\sqrt{ab(a+b)}}+\frac{4C^2\sqrt{ab}}{b^2\sqrt{a+b}}+\frac{C\sqrt{a+b}}{\sqrt{ab}}+C\right)\left(\frac{ab}{a+b}\right)^{\frac{5}{6}}\left(\frac{a+b}{ab}\right)\\
&\le \left(\frac{1}{3\sqrt{6}}+\frac{1}{6\sqrt{6}}+\frac{1}{2\sqrt{3}}+\frac{1}{2}\right)\left(\frac{ab}{a+b}\right)^{\frac{5}{6}}\left(\frac{a+b}{ab}\right).
\end{aligned}$$

Previously, we used that $C \le (1/2)$ and $a, b \ge 6$. Similarly,

$$\begin{aligned}
denominator(y) &= \left[\frac{w_1^2}{b}+\frac{w_2^2}{a}\right]^{1/2}\\
&= \left(\frac{ab}{a+b}\right)\left[\frac{\left(1+\frac{x}{a}\right)^{\frac{2}{3}}}{b}+\frac{\left(1-\frac{x}{b}\right)^{\frac{2}{3}}}{a}\right]^{\frac{1}{2}}\\
&\ge \left(\frac{ab}{a+b}\right)^{\frac{1}{3}}\left[\frac{\left(1+\frac{2x}{3a}-\frac{x^2}{9a^2}\right)}{b}+\frac{\left(1-\frac{2x}{3b}\right)}{a}-\frac{x^2}{9a^2}\right]^{\frac{1}{2}}\\
&= \left(\frac{ab}{a+b}\right)^{\frac{1}{3}}\left[\frac{a\left(1+\frac{2x}{3a}-\frac{x^2}{9a^2}\right)+b\left(1-\frac{2x}{3b}-\frac{x^2}{9b^2}\right)}{ab}\right]^{\frac{1}{2}}\\
&= \left(\frac{ab}{a+b}\right)^{\frac{1}{3}}\left(\frac{a+b}{ab}\left(1-\frac{x^2}{9ab}\right)\right)^{\frac{1}{2}}\\
&\ge \left(\frac{ab}{a+b}\right)^{\frac{1}{3}}\left(\frac{a+b}{ab}\left(1-\frac{4C^2}{9(a+b)}\right)\right)^{\frac{1}{2}}\\
&\ge \left(\frac{ab}{a+b}\right)^{\frac{1}{3}}\left(\frac{a+b}{ab}\left(\frac{107}{108}\right)\right)^{\frac{1}{2}}.
\end{aligned}$$

Hence, we have that

$$y \le \frac{\frac{1}{3\sqrt{6}} + \frac{1}{6\sqrt{6}} + \frac{1}{2\sqrt{3}} + \frac{1}{2}}{\sqrt{\frac{107}{108}}} \le 1,$$

so that $I_z(a,b) \le \phi(1) + \epsilon$ for $\epsilon \le 0.005$. The lemma statement follows by observing that $1 - F_{(a,b)}(z) = 1 - I_z(a,b) \ge 1 - \phi(1) - \epsilon \ge 1 - 0.845 - 0.005 \ge 0.15$. □

Lemma A.2 (Corollary of Lemma A.1). *Let $\tilde{y}_i, \bar{y}_i, \bar{\sigma}_i$ be defined as in Fact A.2. If $m\bar{p}_i, m(\bar{p}_{i+1} + \cdots + \bar{p}_S) \ge 6$, then, for any positive constant $C \le \frac{1}{2}$,*

$$P\left(|\tilde{y}_i - \bar{y}_i| \ge C\bar{\sigma}_i + \frac{C}{m(\bar{p}_i + \cdots + \bar{p}_S)}\right) \ge 0.15 =: \eta.$$

Proof. By Fact A.2, $\tilde{y}_i$ is a Beta random variable with parameters $(m\bar{p}_i, m(\bar{p}_{i+1} + \cdots + \bar{p}_S))$ and mean $\bar{y}_i$. Then, by Lemma A.1 with $a = m\bar{p}_i, b = m(\bar{p}_{i+1} + \cdots + \bar{p}_S)$, we have that, for any $C \le 1/2$,

$$P\left(\tilde{y}_i \ge \bar{y}_i + C\bar{\sigma}_i + \frac{C}{m(\bar{p}_i + \ldots + \bar{p}_S)}\right) \ge 0.15. \tag{A.3}$$

Now, by symmetry of the Beta distribution, $1 - \tilde{y}_i$ is a Beta random variable with parameters $(m(\bar{p}_{i+1} + \cdots + \bar{p}_S), m\bar{p}_i)$ and mean $1 - \bar{y}_i$. Again by Lemma A.1 with $a = m(\bar{p}_{i+1} + \cdots + \bar{p}_S), b = m\bar{p}_i$, we have that, for any $C \le \frac{1}{2}$,

$$P\left((1 - \tilde{y}_i) \ge (1 - \bar{y}_i) + C\bar{\sigma}_i + \frac{C}{m(\bar{p}_i + \ldots + \bar{p}_S)}\right) \ge 0.15. \tag{A.4}$$

The result follows from (A.3) and (A.4). □

Lemma A.3 (Application of Berry-Esseen Theorem). *Let $G \subseteq \{1, \ldots, S\}$ be a set of indices, $z_i \in \mathbb{R}, i \in G$ be fixed. Let*

$$X_G := \sum_{i \in G} (\tilde{y}_i - \bar{y}_i) z_i,$$

with $\tilde{y}_i, \bar{y}_i$ defined as before, and $m\bar{p}_i \ge 6$ for all i. Let F be the cumulative distribution function of

$$\frac{X_G}{\sigma_G}, \quad \text{where,} \quad \sigma_G^2 = \sum_{i \in G} z_i^2 \bar{\sigma}_i^2,$$

with $\bar{\sigma}_i$ being the standard deviation of $\tilde{y}_i$ (refer to Fact A.2). Let Φ be the cumulative distribution function of standard normal distribution. Then, for all $\epsilon > 0$:

$$\sup_x |F(x) - \Phi(x)| \le \epsilon,$$

as long as

$$\sqrt{|G|} \ge \frac{2.8R}{\epsilon}, \quad \text{where} \quad R := \max_{i,j \in G} \frac{z_i \bar{\sigma}_i}{z_j \bar{\sigma}_j}.$$

Proof. We have that $Y_i = (\tilde{y}_i - \bar{y}_i) z_i$. Then, $Y_i, i \in G$ are independent variables, with $\mathbb{E}[Y_i] = 0$,

$$\begin{aligned}
\sigma_i^2 := \mathbb{E}[Y_i^2] &= \mathbb{E}[(\tilde{y}_i - \bar{y}_i)^2 (z_i)^2] \\
&= z_i^2 \bar{\sigma}_i^2, \\
\rho_i := \mathbb{E}[|Y_i|^3] &\le \mathbb{E}[|Y_i|^4]^{3/4} \\
&= \mathbb{E}[|\tilde{y}_i - \bar{y}_i|^4]^{3/4} z_i^3 \\
&= \kappa_i^{3/4} \mathbb{E}[|\tilde{y}_i - \bar{y}_i|^2]^{3/2} z_i^3 \\
&\le \kappa_i \bar{\sigma}_i^3 z_i^3,
\end{aligned}$$

where the first inequality is by using Jensen's inequality, and

$$\kappa_i = \frac{\mathbb{E}[(\tilde{y}_i - \bar{y}_i)^4]}{\mathbb{E}[(\tilde{y}_i - \bar{y}_i)^2]^2} \ge 1$$

is the kurtosis of $\tilde{y}_i$. It is known that the kurtosis of a $Beta(\nu\mu, \nu(1-\mu))$ random variable is

$$\kappa_i = 3 + \frac{6}{(3+\nu)}\left(\frac{(1-2\mu)^2(1+\nu)}{\mu(1-\mu)(2+\nu)} - 1\right) \le 3 + \frac{6}{\nu\mu(1-\mu)}.$$

For $\tilde{y}_i \sim Beta(m\bar{p}_i, m(\bar{p}_{i+1} + \ldots + \bar{p}_S))$, $\nu = m(\bar{p}_i + \ldots + \bar{p}_S)$ and $\mu = \bar{y}_i$. One of μ and $1-\mu$ is at least $\frac{1}{2}$, so that

$$\kappa_i \le 3 + \frac{12}{\min(m\bar{p}_i, m(\bar{p}_{i+1} + \ldots + \bar{p}_S))} \le 5,$$

because $m\bar{p}_i \ge 6$ for all i.

Now, we use Berry-Esseen theorem (Fact B.4), with

$$\psi_1 = \frac{1}{\sqrt{\sum_{i\in G}\sigma_i^2}} \max_{i\in G} \frac{\rho_i}{\sigma_i^2} \le \frac{(\max_{i\in G}\kappa_i)}{\sqrt{|G|}} \frac{\max_{i\in G} z_i\bar{\sigma}_i}{\min_{i\in G} z_i\bar{\sigma}_i}$$

to obtain

$$\sup_x |F(x) - \Phi(x)| \le 0.56\psi_1 \le \frac{2.8}{\sqrt{|G|}} \frac{\max_{i\in G} z_i\bar{\sigma}_i}{\min_{i\in G} z_i\bar{\sigma}_i}.$$

The lemma statement follows if $\sqrt{|G|} \ge 2.8R/\epsilon$, where $R := \max_{i,j\in G}(z_i\bar{\sigma}_i/z_j\bar{\sigma}_j)$. □

Lemma A.4. *Assume $m\bar{p}_i \ge 6$, $\forall i$. Then, for any fixed z_i, $i = 1, \ldots, S$,*

$$\Pr\left(\sum_{i=1}^{S}(\tilde{y}_i - \bar{y}_i)z_i \ge \frac{1}{4}\sqrt{\sum_{i=1}^{S} z_i^2\bar{\sigma}_i^2}\right) \ge \frac{\eta^{k(\delta)}}{2S},$$

where $\eta = 0.15, \delta = ((1-\Phi)(1/2))/2, k(\delta) = 2.8^2/\delta^4$, with Φ being the cdf of the standard normal distribution.

Proof. We consider two cases: if $S < k(\delta)$ and if $S \ge k(\delta)$. For the first case, when $S < k(\delta)$, we use Lemma A.2 on each index i, so that

$$\Pr\left(\sum_{i}(\tilde{y}_i - \bar{y}_i)z_i \ge 0.5\sqrt{\sum_{i} z_i^2\bar{\sigma}_i^2}\right) \ge \eta^{k(\delta)},$$

where $\eta = 0.15$.

If it is the case that $S \ge k(\delta)$, we consider the group of indices with the $k(\delta)$ largest values of $|z_i\bar{\sigma}_i|$, call it group $G(1)$, and then divide the remaining indices into groups of $\mathcal{G}$ as follows. Note that $\mathcal{G} - G(1)$ is the set of all ungrouped indices. Let index $k = \arg\max_{i\in\mathcal{G}-G(1)}|z_i\bar{\sigma}_i|$. Then the new group $G(2)$ consists of index k along with all other indices $j \in \mathcal{G} - G(1)$, where $|z_k\bar{\sigma}_k|/|z_j\bar{\sigma}_j| \le \frac{1}{\delta}$. To form successive groups, we repeat this procedure on the remaining ungrouped indices, creating new groups when necessary, until all are grouped. By construction, we have that $|z_i\bar{\sigma}_i|/|z_j\bar{\sigma}_j| \le \frac{1}{\delta}$ for all i, j in any given group G. In addition, we can define an ordering $\prec$ on groups by ordering them by maximum value of $|z_i\bar{\sigma}_i|$ in the group, that is, $G \succ G'$ if $\max_{i\in G} z_i^2\bar{\sigma}_i^2 \ge \max_{j\in G'} z_j^2\bar{\sigma}_j^2$. Then, for $G \succ G'$, we also have $\max_{i\in G} z_i^2\bar{\sigma}_i^2 \ge \frac{1}{\delta^2}\max_{j\in G'} z_j^2\bar{\sigma}_j^2$.

Recall from Lemma A.3, for every group $G \in \mathcal{G}$ of size $\sqrt{|G|} > 2.8/\delta\epsilon$, we have that the cdf of $\frac{X_G}{\sigma_G}$ is within ϵ of the normal distribution cdf. By definition of δ, it follows that $\Pr(X_G/\sigma_G \ge \frac{1}{2}) \ge 2\delta - \epsilon$. Using this result for $\epsilon = \delta$, we get that for every group of size at least $k(\delta)$, we have, for any group G where $|G| \ge (k(\delta))^2$,

$$\Pr\left(X_G \ge \frac{1}{2}\sigma_G\right) \ge \delta.$$

Now, consider the top $\log_{1/\delta}(S)$ groups (with respect to the ordering $\prec$), including $G(1)$. First consider the top groups of cardinality at least $k(\delta)$: call these "top big groups." For every top big group of (including $G(1)$), we have that

$$\sum_{i\in\text{top big groups}}(\tilde{y}_i - \bar{y}_i)z_i \ge \frac{1}{2}\sqrt{\sum_{i\in\text{top big groups}} z_i^2\bar{\sigma}_i^2}, \tag{A.5}$$

with probability at least $\delta^{\log_{1/\delta}(S)} = \frac{1}{S}$. Next, consider the remaining top groups where the cardinality is less than $k(\delta)$, we refer to these as "top small groups." For the first (i.e., lowest index) top small group, say $G(\ell)$, we use Lemma A.2 at most $k(\delta)$ times, so that

$$\Pr\left(\sum_{i\in G(\ell)} (\tilde{y}_i - \bar{y}_i) z_i \geq 0.5 \sqrt{\sum_{i\in G(\ell)} z_i^2 \bar{\sigma}_i^2}\right) \geq \eta^{k(\delta)}, \tag{A.6}$$

where $\eta \geq 0.15$. Combining (A.5) and (A.6), we have that with probability at least $\frac{1}{S}\eta^{k(\delta)}$,

$$\sum_{i\in \text{ top big groups}, G(\ell)} (\tilde{y}_i - \bar{y}_i) z_i \geq \frac{1}{2} \sqrt{\sum_{i\in \text{ top big groups}, \bar{G}} z_i^2 \bar{\sigma}_i^2}. \tag{A.7}$$

Let $G(j)$ denote the jth group according to ordering $\succ$. Also, let $|z_{(1)}\bar{\sigma}_{(1)}| = \min_{i\in G(1)} |z_i \bar{\sigma}_i|$. Then, because for any G, G', $G \succ G'$, we have that $\max_{i\in G} z_i^2 \bar{\sigma}_i^2 \geq \frac{1}{\delta^2} \max_{j\in G'} z_j^2 \bar{\sigma}_j^2$, for every remaining top small group $G(j), j > \ell$, we can bound the group's total variance by

$$k(\delta) \max_{i\in G(j)} z_i^2 \bar{\sigma}_i^2 \leq k(\delta)\delta^{2(j-\ell)} \max_{i\in G(\ell)} z_i^2 \bar{\sigma}_i^2 \leq k(\delta)\delta^{2(j-2)} z_{(1)}^2 \bar{\sigma}_{(1)}^2.$$

Therefore, the sum of the standard deviation for top small groups, excluding group $G(\ell)$, is at most

$$k(\delta) \sum_{G:\text{top small groups}-G(\ell)} \max_{i\in G} z_i^2 \bar{\sigma}_i^2 \leq k(\delta) \sum_{j=3}^{\log_{1/\delta}(S)} \delta^{2(j-2)} z_{(1)}^2 \bar{\sigma}_{(1)}^2 \leq \frac{k(\delta)\delta^2}{1-\delta^2} z_{(1)}^2 \bar{\sigma}_{(1)}^2$$

as it is a geometric series with multiplier δ^2.

For the remaining "bottom groups" (i.e., those not among the top $\log_{1/\delta}(S)$ groups), each element's variance is at most $\frac{1}{S^2} z_{(1)}^2 \bar{\sigma}_{(1)}^2$; therefore,

$$\sum_{i\in \text{top small groups}-G(\ell), \text{ bottom groups}} z_i^2 \bar{\sigma}_i^2 \leq \left(\frac{k(\delta)\delta^2}{1-\delta^2} + \frac{S}{S^2}\right) z_{(1)}^2 \bar{\sigma}_{(1)}^2 \leq \frac{k(\delta)}{25} z_{(1)}^2 \bar{\sigma}_{(1)}^2 \leq \frac{1}{25} \sum_{i\in G(1)} z_i^2 \bar{\sigma}_i^2. \tag{A.8}$$

By Cantelli's inequality (Fact B.3), with probability at least 1/2,

$$\sum_{i\in \text{top small groups}-G(\ell), \text{ bottom groups}} (\tilde{y}_i - \bar{y}_i) z_i \geq -\sqrt{\sum_{i\in \text{top small groups}-G(\ell), \text{ bottom groups}} z_i^2 \bar{\sigma}_i^2} \geq -\frac{1}{5} \sqrt{\sum_{i\in G(1)} z_i^2 \bar{\sigma}_i^2}. \tag{A.9}$$

Hence, combining (A.7) and (A.9), with probability at least $\eta^{k(\delta)}/2S = \Omega(1/S)$,

$$\begin{aligned}
\sum_i (\tilde{y}_i - \bar{y}_i) z_i &\geq \frac{1}{2} \sqrt{\sum_{i\in \text{top big groups}, G(\ell)} z_i^2 \bar{\sigma}_i^2} - \frac{1}{5} \sqrt{\sum_{i\in G(1)} z_i^2 \bar{\sigma}_i^2} \\
&\geq \frac{1}{4} \sqrt{\sum_{i\in \text{top big groups}, G(\ell)} z_i^2 \bar{\sigma}_i^2} + \frac{1}{20} \sqrt{\sum_{i\in G(1)} z_i^2 \bar{\sigma}_i^2} \\
&\geq \frac{1}{4} \sqrt{\sum_{i\in \text{top big groups}, G(\ell)} z_i^2 \bar{\sigma}_i^2} + \frac{1}{4} \sqrt{\sum_{i:\text{top small groups}-G(\ell), \text{ bottom groups}} z_i^2 \bar{\sigma}_i^2} \\
&\geq \frac{1}{4} \sqrt{\sum_i z_i^2 \bar{\sigma}_i^2},
\end{aligned}$$

where we used (A.8) in the previous third inequality. □

We are now ready to prove Proposition 1.

Proof of Proposition 1. Because $\tilde{p}$ and $\bar{p}$ are probability vectors (sum to one), it is sufficient to consider $h \in [0, D]^S$. Now, use Fact A.1 to express $(\tilde{p} - \bar{p})^T h$ as

$$(\tilde{p} - \bar{p})^T h = \sum_i (\tilde{y}_i - \bar{y}_i)(h_i - \tilde{H}_{i+1})(\tilde{p}_i + \cdots + \tilde{p}_S).$$

We note that $\tilde{H}_i$ is the scalar product of $(S-i+1)$-dimensional Dirichlet random vector $(\tilde{y}_i,\ldots,\tilde{y}_S)$ with the fixed vector $(h_i,\ldots,h_S)$, and $\bar{H}_i$ is the expected value of that product. Therefore, we can derive deviation bounds for this product using a similar argument as used in the proof of Case 1 of Lemma 4 in Section 5.3.

For any i, Let X be distributed as Gaussian with mean $\mu=\bar{H}_i/D$ and variance $\sigma^2=\frac{1}{m(\bar{p}_i+\ldots+\bar{p}_S)}$. Now, by Gaussian-Dirichlet stochastic optimism (Lemma B.1), $X\succeq_{so}\frac{1}{D}\tilde{H}_i$. Then by Gaussian concentration and Corollary B.2, for any $\rho\in(0,1)$,

$$|\tilde{H}_i-\bar{H}_i|\le D\sqrt{\frac{2\log(2/\rho)}{m(\bar{p}_i+\ldots+\bar{p}_S)}}, \tag{A.10}$$

with probability $1-\rho$.

Similarly, noting that $\tilde{y}_i$ is a Beta random variable, using Gaussian-Beta stochastic optimism (Lemma B.2), if X is distributed as Gaussian with mean $\mu=\bar{y}_i$ and variance $\sigma^2=\frac{1}{m(\bar{p}_i+\ldots+\bar{p}_S)}$, then $X\succeq_{so}\tilde{y}_i$. Then by Corollary B.2, with probability $1-\rho$,

$$|\tilde{y}_i-\bar{y}_i|\le\sqrt{\frac{2\log(2/\rho)}{m(\bar{p}_i+\ldots+\bar{p}_S)}}. \tag{A.11}$$

Therefore, with probability $1-2S\rho$,

$$\begin{aligned}
&(\tilde{p}-\bar{p})^T h-\sum_i(\tilde{y}_i-\bar{y}_i)(h_i-\bar{H}_{i+1})(\bar{p}_i+\cdots+\bar{p}_S)\\
&=\sum_i(\tilde{y}_i-\bar{y}_i)(\bar{H}_{i+1}-\tilde{H}_{i+1})(\bar{p}_i+\cdots+\bar{p}_S)\\
&\ge-\sum_i\sqrt{\frac{2\log(2/\rho)}{m(\bar{p}_i+\ldots+\bar{p}_S)}}D\sqrt{\frac{2\log(2/\rho)}{m(\bar{p}_i+\ldots+\bar{p}_S)}}(\bar{p}_i+\cdots+\bar{p}_S)\\
&=-\frac{2SD\log(2/\rho)}{m}.
\end{aligned} \tag{A.12}$$

Then, applying Lemma A.4 (given $m\bar{p}_i\ge 6$) for $z_i=(h_i-\bar{H}_{i+1})(\bar{p}_i+\cdots+\bar{p}_S), i=1,\ldots,S$, with probability at least $\eta^{k(\delta)}/2S$,

$$(\tilde{p}-\bar{p})^T h\ge\frac{1}{4}\sqrt{\sum_i z_i^2\bar{\sigma}_i^2}-\frac{2SD\log(2/\rho)}{m}.$$

Now, we observe

$$\sum_i z_i^2\bar{\sigma}_i^2=(h_i-\bar{H}_{i+1})^2(\bar{p}_i+\cdots+\bar{p}_S)^2\bar{\sigma}_i^2=\frac{\bar{c}_i^2\bar{p}_i(\bar{p}_i+\ldots,\bar{p}_S)}{m(\bar{p}_i+\ldots+\bar{p}_S)+1},$$

to obtain that with probability at least $(\eta^{k(\delta)}/2S)-2S\rho$,

$$(\tilde{p}-\bar{p})^T h\ge\frac{1}{8}\sqrt{\sum_i\frac{\bar{\gamma}_i\bar{c}_i^2}{m}}-\frac{2SD\log(2/\rho)}{m},$$

where

$$\bar{\gamma}_i=\frac{\bar{p}_i(\bar{p}_{i+1}+\ldots+\bar{p}_S)}{(\bar{p}_i+\ldots+\bar{p}_S)}.\quad\square$$

A.2. Concentration of Empirical Probability Vectors: Proof of Proposition 2

Proposition 2 (Restated from Main Text). *Fix any vector $h\in\mathbb{R}^S$ such that $|h_i-h_{i'}|\le D$ for any i,i'. Fix any epoch k, state s, action a. As defined in the notations section, denote $n=N_{s,a}^{\tau_k}$, $n_i=N_{s,a}^{\tau_k}(i)$, $p_i=P_{s,a}(i)$, $\hat{p}_i:=\frac{n_i}{n}$, for $i=1,\ldots,S$. Then, if $\tau_k-1\ge 96$, for any $\rho\in(0,1)$, with probability $1-\rho$,*

$$|(\hat{p}-p)^T h|\le 2\sqrt{\log(T/\rho)\sum_{i<S}\frac{\gamma_i c_i^2}{n}}+3D\frac{\log(T/\rho)}{n},$$

where $\gamma_i=\frac{p_i(p_{i+1}+\cdots+p_S)}{(p_i+\cdots+p_S)}$, $c_i=h_i-H_{i+1}$, $H_{i+1}=\frac{1}{\sum_{j=i+1}^S p_j}\sum_{j=i+1}^S h_j p_j$, for any fixed ordering on states $i=1,\ldots,S$.

Proof. For every $t = 2, \ldots, T, i = 1, \ldots, S$, define

$$Z_{t,i} = \left(c_i \, \mathbb{1}(s_t = i) - c_i \frac{p_i}{p_i + \cdots + p_S} \cdot \mathbb{1}(s_t \in \{i, \ldots, S\}) \right) \mathbb{1}(s_{t-1} = s, a_{t-1} = a),$$

$$Z_t = \sum_{i=1}^{S} Z_{t,i}.$$

For $i = 1, \ldots, S$, define $y_i = \frac{p_i}{(p_i + \cdots + p_S)}, \hat{y}_i = \frac{\hat{p}_i}{(\hat{p}_i + \cdots + \hat{p}_S)}$. Then,

$$\frac{\sum_{t=2}^{\tau_k} Z_t}{n} = \sum_i c_i \hat{p}_i - \sum_i \frac{c_i p_i}{p_i + \cdots + p_S} \cdot (\hat{p}_i + \ldots + \hat{p}_S) = \sum_{i=1}^{S-1} (\hat{y}_i - y_i)(\hat{p}_i + \ldots + \hat{p}_S) c_i = (\hat{p} - p)^T h,$$

where we used Fact A.1 for the last equality. Now, $E[Z_t | s_{t-1}, a_{t-1}] = \sum_i E[Z_{t,i} | s_{t-1}, a_{t-1}] = 0$. Also, we observe that for any t, $Z_{t,i}$ and $Z_{t,j}$ for any $i \neq j$ are independent given the state s_{t-1} and action a_{t-1}: (assume $j > i$ without loss of generality)

$$\begin{aligned}
\mathbb{E}[Z_{t,i} Z_{t,j} | s_{t-1} = s, a_{t-1} = a] &= c_i c_j \mathbb{E}\Bigg[\mathbb{1}(s_t = i)\, \mathbb{1}(s_t = j) - \mathbb{1}(s_t = j) \frac{p_i}{p_i + \cdots + p_S} \cdot \mathbb{1}(s_t \in \{i, \ldots, S\}) \\
&\quad - \mathbb{1}(s_t = i) \frac{p_j}{p_j + \cdots + p_S} \cdot \mathbb{1}(s_t \in \{j, \ldots, S\}) \\
&\quad + \frac{p_j p_i}{(p_j + \cdots + p_S)(p_i + \cdots + p_S)} \cdot \mathbb{1}(s_t \in \{j, \ldots, S\}) | s_{t-1} = s, a_{t-1} = a \Bigg] \\
&= c_i c_j \mathbb{E}\left[-\mathbb{1}(s_t = j) \frac{p_i}{p_i + \cdots + p_S} + \frac{p_j p_i}{(p_j + \cdots + p_S)(p_i + \cdots + p_S)} \cdot \mathbb{1}(s_t \in \{j, \ldots, S\}) | s_{t-1} = s, a_{t-1} = a \right] \\
&= c_i c_j \left(-\frac{p_j p_i}{p_i + \cdots + p_S} + \frac{p_j p_i}{(p_i + \cdots + p_S)} \right) \\
&= 0.
\end{aligned}$$

Therefore,

$$\sum_{t=2}^{\tau_k} E[Z_t^2 | s_{t-1}, a_{t-1}] = \sum_{t=2}^{\tau_k} \sum_i \mathbb{E}[Z_{t,i}^2 | s_{t-1}, a_{t-1}] = \sum_{i=1}^{S-1} c_i^2 n \gamma_i,$$

where the last equality is obtained using the following derivation:

$$\mathbb{E}\left[\sum_{t=2}^{\tau_k} Z_{t,i}^2 \,\middle|\, s_{t-1} = s, a_{t-1} = a \right] = c_i^2 \sum_{t=2}^{\tau_k} \mathbb{1}(s_{t-1} = s, a_{t-1} = a) \left(p_i - \frac{p_i^2}{(p_i + \cdots + p_S)^2} (p_i + \cdots + p_S) \right).$$

The previous expression is zero for $i = S$. For $i < S$, we get

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=2}^{\tau_k} Z_{t,i}^2 \,\middle|\, s_{t-1} = s, a_{t-1} = a \right] &= c_i^2 \sum_{t=2}^{\tau_k} \mathbb{1}(s_{t-1} = s, a_{t-1} = a) \frac{p_i (p_{i+1} + \cdots + p_S)}{p_i + \cdots + p_S} \\
&= c_i^2 n \frac{p_i (p_{i+1} + \cdots + p_S)}{p_i + \cdots + p_S} = n \gamma_i c_i^2.
\end{aligned}$$

Then, using Bernstein's inequality (refer to Corollary B.1) with $M_{\tau_k} = |\sum_{t=2}^{\tau_k} Z_t|$ and $V_{\tau_k} := \sum_{t=2}^{\tau_k} E[Z_t^2 | s_{t-1}, a_{t-1}] \leq \sum_{i<S} n \gamma_i c_i^2$, we get the desired bound on $(p - \hat{p})^T h = \frac{1}{n} \sum_{t=2}^{\tau_k} Z_t$. □

A.3. Modified Anti-Concentration Bound: Proof of Proposition 3

We use the notation described in Section 5.1. Given an epoch k, state s, action a, and sample j, we denote $n = N_{s,a}^{\tau_k}, n_i = N_{s,a}^{\tau_k}(i)$, $m = \frac{n + \omega S}{\kappa}$, where $\omega = 720 \log(T/\rho)$ and $\kappa = 120 \log(T/\rho)$, as defined in the algorithm. Then, we denote $p_i = P_{s,a}(i)$, $\hat{p}_i := \frac{n_i}{n}$, $\bar{p}_i = \frac{n_i + \omega}{n + \omega S}$, and $\tilde{p}_i = Q_{s,a}^{j,k}(i)$, for $i \in \mathcal{S}$.

Also, as defined earlier in Proposition 1 and Proposition 2, we denote

$$\bar{\gamma}_i := \frac{\bar{p}_i (\bar{p}_{i+1} + \ldots + \bar{p}_S)}{(\bar{p}_i + \ldots + \bar{p}_S)}, \bar{c}_i = (h_i - \bar{H}_{i+1}), \bar{H}_{i+1} = \frac{1}{\sum_{j=i+1}^{S} \bar{p}_j} \sum_{j=i+1}^{S} h_j \bar{p}_j,$$

and

$$\gamma_i = \frac{p_i(p_{i+1}+\cdots+p_S)}{(p_i+\cdots+p_S)}, c_i = h_i - H_{i+1}, H_{i+1} = \frac{1}{\sum_{j=i+1}^{S} p_j}\sum_{j=i+1}^{S} h_j p_j.$$

We prove the following result for s, a such that $n > \eta$. Recall that for such s, a, the algorithm uses Dirichlet posterior sampling to generate sample vectors $Q_{s,a}^{j,k}$, so that in this case $\tilde{p}$ is a random vector distributed as Dirichlet$(m\bar{p}_1, \ldots, m\bar{p}_S)$.

Proposition A.1. *Assume that $h \in [0,D]^S$, and $n > 12\omega S^2$, and states $i = 1,\ldots,S$ are ordered such that $\bar{p}_1 \le \cdots \le \bar{p}_S$. Also, $\omega = 720\log(T/\rho), \kappa = \frac{\omega}{6}$, as defined in the algorithm. Then, with probability $\frac{C}{S} - 7S\rho$,*

$$(\tilde{p}-\bar{p})^T h \ge 0.148\sqrt{\kappa\sum_i \frac{\gamma_i c_i^2}{n}} - O\left(\frac{DS\omega\log(T/\rho)}{n}\right).$$

The constant C is defined as in Proposition 1.

Proof. The proof is obtained by a modification to the proof of Proposition 1, which proves a similar bound but in terms of $\bar{\gamma}_i$ s and $\bar{c}_i$ s and m. Because $\kappa = \omega/6$, $m\bar{p}_i = (n_i+\omega)/\kappa \ge 6$.

In the proof of that proposition, we obtained (refer to Equation (A.12)) that with probability $1-2S\rho$ (given $m\bar{p}_i \ge 6$),

$$\begin{aligned}(\tilde{p}-\bar{p})^T h &\ge \sum_i (\tilde{y}_i - \bar{y}_i)(h_i - \bar{H}_{i+1})(\bar{p}_i+\cdots+\bar{p}_S) - \frac{2DS\log(2/\rho)}{m}\\ &\ge \sum_i (\tilde{y}_i - \bar{y}_i)(h_i - \bar{H}_{i+1})(\bar{p}_i+\cdots+\bar{p}_S) - O\left(\frac{DS\omega\log(T/\rho)}{n}\right),\end{aligned}$$

where $\tilde{y}_i := \frac{\tilde{p}_i}{\tilde{p}_i+\cdots+\tilde{p}_S}, \bar{y}_i := \frac{\bar{p}_i}{\bar{p}_i+\cdots+\bar{p}_S}$, $\tilde{H}_{i+1} = \frac{1}{\sum_{j=i+1}^{S}\tilde{p}_j}\sum_{j=i+1}^{S} h_j\tilde{p}_j$, $\bar{H}_{i+1} = \frac{1}{\sum_{j=i+1}^{S}\bar{p}_j}\sum_{j=i+1}^{S} h_j\bar{p}_j$. Now, breaking up the term in the summation and using Lemma A.7 to bound $|H_{i+1} - \bar{H}_{i+1}|(\bar{p}_i+\cdots+\bar{p}_S)$ (because we have that $\omega = 720\log(T/\rho)$ and $n > 12\omega S^2$ by assumption) and Lemma B.2 and Corollary B.2 to bound $|\tilde{y}_i - \bar{y}_i|$ (see (A.11) in the proof of Proposition 1), we get that for every i, with probability $1-4S\rho$,

$$\begin{aligned}&(\tilde{p}-\bar{p})^T h - \sum_i (\tilde{y}_i - \bar{y}_i)(h_i - H_{i+1})(\bar{p}_i+\cdots+\bar{p}_S) + O\left(\frac{DS\omega\log(T/\rho)}{m}\right)\\ &\ge \sum_i (\tilde{y}_i - \bar{y}_i)(\bar{H}_{i+1} - H_{i+1})(\bar{p}_i+\cdots+\bar{p}_S)\\ &\ge -\sum_i \sqrt{\frac{2\log(2/\rho)}{m(\bar{p}_i+\cdots+\bar{p}_S)}}\left(3D\sqrt{\log(T/\rho)\frac{(\bar{p}_i+\cdots+\bar{p}_S)}{n}} + 4\frac{(\omega S+\log(T/\rho))D}{n}\right)\\ (*) &\ge -\frac{6DS\sqrt{\log(2/\rho)\log(T/\rho)}}{\sqrt{mn}} - \frac{4(\omega S+\log(T/\rho))D\sqrt{2\log(2/\rho)}}{n\sqrt{m}}\sum_i \frac{1}{\sqrt{(\bar{p}_i+\cdots+\bar{p}_S)}}.\end{aligned}$$

Recall that $m = (n+\omega S)/\kappa$, so that for $n > S\omega$, $n \ge m\kappa/2 = m\omega/12 \ge m\log(2/\rho)$. Therefore, the first term of $(*)$ is at least

$$-\frac{6DS\sqrt{\log(2/\rho)\log(T/\rho)}}{\sqrt{m^2\log(2/\rho)}} = -\frac{6DS\sqrt{\log(T/\rho)}}{m} = -O\left(\frac{DS\omega\log(T/\rho)}{n}\right).$$

Then using Lemma A.5 and $m = (n+S\omega)/\kappa > 6n/\omega > 72S^2$, the second term in $(*)$ is at least

$$-\frac{8S(\omega S+\log(T/\rho))D\sqrt{2\log(2/\rho)}}{n\sqrt{72S^2}} = -O\left(\frac{DS\omega\log(T/\rho)}{n}\right).$$

Now, applying Lemma A.4 (given $m\bar{p}_i \ge 6$) for $z_i = (h_i - H_{i+1})(\bar{p}_i+\cdots+\bar{p}_S), i = 1,\ldots,S$, with probability $\eta^{k(\delta)}/2S$,

$$\sum_i (\tilde{y}_i - \bar{y}_i)z_i \ge \frac{1}{4}\sqrt{\sum_i \bar{\sigma}_i^2 z_i^2}.$$

We substitute this in the left-hand side along with the observation

$$\sum_i z_i^2\bar{\sigma}_i^2 = \sum_i (h_i - H_{i+1})^2(\bar{p}_i+\cdots+\bar{p}_S)^2\bar{\sigma}_i^2 = \sum_i \frac{c_i^2\bar{p}_i(\bar{p}_i+\ldots,\bar{p}_S)}{m(\bar{p}_i+\ldots+\bar{p}_S)+1} \ge \sum_i \frac{6}{7}\frac{\bar{\gamma}_i c_i^2}{m}.$$

Thus far, we have that with probability $(\eta^{k(\delta)}/2S) - 4S\rho$,

$$(\tilde{p} - \bar{p})^T h \geq \frac{\sqrt{6}}{4\sqrt{7}} \sqrt{\sum_i \frac{\bar{\gamma}_i c_i^2}{m}} - O\left(\frac{DS\omega \log(T/\rho)}{n}\right). \tag{A.13}$$

Now, because $n > 12\omega S^2$ and $S \geq 2$, we have that $1/n + \omega S \geq 24/25n$ and hence $1/m = \kappa/n + \omega S \geq 24\kappa/25n$. Finally, we use Lemma A.6 with $c = (12\sqrt{30})/5$ to lower bound $\bar{\gamma}_i$ by $(1/1.53)\gamma_i - O(\omega S/n)$ to get with probability $(\eta^{k(\delta)}/2S) - 7S\rho$,

$$\begin{aligned}(\tilde{p} - \bar{p})^T h &\geq \frac{1}{1.53} \frac{\sqrt{6}}{4\sqrt{7}} \frac{\sqrt{24}}{\sqrt{25}} \sqrt{\kappa \sum_i \frac{\gamma_i c_i^2}{n}} - O\left(\frac{DS\omega \log(T/\rho)}{n}\right) \\ &\geq 0.148 \sqrt{\kappa \sum_i \frac{\gamma_i c_i^2}{n}} - O\left(\frac{DS\omega \log(T/\rho)}{n}\right). \quad \square\end{aligned}$$

Lemma A.5. *Let $x \in \mathbb{R}^n$ such that $0 \leq x_1 \leq \cdots \leq x_n \leq 1$ and $\sum_i x_i = 1$. Then,*

$$\sum_{i=1}^{n} \frac{1}{\sqrt{x_i + \cdots x_n}} \leq 2n.$$

Proof. Define $f(x,j) := \frac{1}{\sqrt{x_j + \cdots + x_n}}$ for all $j = 1, \cdots, n$, and $f(x) = \sum_{j=1}^n f(x,j)$. We prove that $x^* := \left(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}\right)$ achieves the maximum value of $f(x)$. Consider any solution x'. Suppose that there exists some index pair i, j with $i < j$ and some $\epsilon > 0$ such that $x'_i \neq x'_j$ and increasing x'_i by ϵ and decreasing x'_j by ϵ preserves the ordering of the indices. This would strictly increase the objective $f(x')$, because $f(x', k)$ strictly increases for all $i < k \leq j$ and remains unchanged otherwise. Hence, x' is not an optimal solution. The only case where no such index pair (i, j) exists is when all x'_i are equal, that is, when $x' = x^*$. Because $f(x) = \sum_i f(x,i)$ is a continuous function over a compact set, it has a maximum, which therefore must be attained at x^*.

This means

$$f(x) = \sum_{i=1}^{n} \frac{1}{\sqrt{x_i + \cdots x_n}} \leq \sum_{i=1}^{n} \frac{1}{\sqrt{x_i^* + \cdots + x_n^*}} = \sum_{i=1}^{n} \sqrt{\frac{n}{i}} \leq \sqrt{n} \int_{i=0}^{n} \frac{1}{\sqrt{i}} di = 2n. \quad \square$$

Lemma A.6. *Let $6/5 < c \leq (12\sqrt{30})/5$ and $n > 12\omega S^2$, where $\omega = 720 \log(T/\rho)$. Then for any group $G \subseteq \mathcal{S}$ of indices, with probability $1 - \rho$,*

$$\left(1 - \frac{1}{c}\right) \sum_{i \in G} \bar{p}_i - \frac{2\omega S}{n} \leq \sum_{i \in G} p_i \leq \left(1 + \frac{1}{c}\right) \sum_{i \in G} \bar{p}_i + \frac{2\omega S}{n}.$$

If in the definition of $\bar{\gamma}_i$, we use an ordering of i such that $\bar{p}_S \geq \frac{1}{S}$ (e.g., if $\max \bar{p}_i$ is the last in the ordering), then for all i, with probability $1 - 3\rho$,

$$\gamma_i \leq \frac{\left(1 + \frac{1}{c}\right)^2}{1 - \frac{1}{c} - \frac{1}{6}} \bar{\gamma}_i + \frac{2\left(1 + \frac{1}{c} + \frac{1}{6}\right)\omega S}{1 - \frac{1}{c} - \frac{1}{6}} \frac{\omega S}{n}.$$

Proof. Given a group $G \subseteq \mathcal{S}$, define $p_G = \sum_{i \in G} p_i$, $\hat{p}_G = \sum_{i \in G} \hat{p}_i$, and $\bar{p}_G = \sum_{i \in G} \bar{p}_i$. Also define $Z_t = (\mathbb{1}(s_t \in G) - p_G)\, \mathbb{1}(s_{t-1} = s, a_{t-1} = a)$. Then, let $M_\tau = \sum_{t=2}^{\tau} Z_t$ and $V_\tau = \sum_{t=1}^{\tau} \mathbb{E}[(Z_t)^2 | \mathcal{F}_{t-1}]$. Note that $\mathbb{E}[Z_t | \mathcal{F}_{t-1}] = 0$ and

$$V_\tau = n(p_G(1 - p_G)^2 + (1 - p_G)(-p_G)^2) = np_G(1 - p_G),$$

so that by Bernstein's inequality (Corollary 1),

$$|M_\tau| = \left|\sum_{t=2}^{\tau} Z_t\right| \leq 2\sqrt{np_G(1 - p_G)\log(\tau/\rho)} + 3\log(\tau/\rho) \leq 2\sqrt{np_G \log(\tau/\rho)} + 3\log(\tau/\rho).$$

Noting that

$$\frac{\sum_{t=2}^{\tau} \mathbb{1}(s_t \in G)\mathbb{1}(s_{t-1} = s, a_{t-1} = a)}{n} = \sum_{i \in G} \hat{p}_i,$$

we have that

$$|\hat{p}_G - p_G| = \frac{|M_\tau|}{n} \le 2\sqrt{\frac{p_G \log(\tau/\rho)}{n}} + \frac{3\log(\tau/\rho)}{n}.$$

Substituting $p_G \le \hat{p}_G + |\hat{p}_G - p_G|$ in the right-hand side, rearranging terms, and squaring both sides, we have

$$|\hat{p}_G - p_G|^2 - 2|\hat{p}_G - p_G|\frac{3\log(\tau/\rho)}{n} + \left(\frac{3\log(\tau/\rho)}{n}\right)^2 \le \frac{4\hat{p}_G \log(\tau/\rho)}{n} + \frac{4|\hat{p}_G - p_G|\log(\tau/\rho)}{n}.$$

Then, simplifying by completing the square, the previous expression is equivalent to

$$|\hat{p}_G - p_G| \le 2\sqrt{\frac{\hat{p}_G \log(\tau/\rho)}{n}} + \frac{9\log(\tau/\rho)}{n}.$$

Because $|\bar{p}_G - \hat{p}_G| \le \omega S/n$,

$$\begin{aligned}
|p_G - \bar{p}_G| &= |p_G - \hat{p}_G + \hat{p}_G - \bar{p}_G| \\
&\le 2\sqrt{\frac{\left(\bar{p}_G + \frac{\omega S}{n}\right)\log(\tau/\rho)}{n}} + \frac{9\log(\tau/\rho)}{n} + \frac{\omega S}{n} \\
&\le 2\sqrt{\frac{\bar{p}_G \log(\tau/\rho)}{n}} + 2\sqrt{\frac{\frac{\omega S}{n}\log(\tau/\rho)}{n}} + \frac{9\log(\tau/\rho)}{n} + \frac{\omega S}{n} \\
&\le 2\sqrt{\frac{\bar{p}_G \log(\tau/\rho)}{n}} + \frac{2\omega S}{n},
\end{aligned}$$

because $\omega = 720\log(T/\rho)$. Now, for $n > 12\omega S^2$ and $S \ge 2$,

$$n\bar{p}_i = n\frac{n\hat{p}_i + \omega}{n + \omega S} \ge \frac{n\omega}{n + \omega S} \ge \frac{24\omega}{25} \ge 4c^2\log(\tau/\rho),$$

for $c \le (12\sqrt{30})/5$. Hence,

$$|p_G - \bar{p}_G| \le \bar{p}_G\sqrt{\frac{4\log(\tau/\rho)}{n\bar{p}_G}} + \frac{2\omega S}{n} \le \bar{p}_G\sqrt{\frac{4\log(\tau/\rho)}{4c^2\log(\tau/\rho)}} + \frac{2\omega S}{n} \le \frac{1}{c}\bar{p}_G + \frac{2\omega S}{n},$$

so that

$$\sum_i p_i \le \left(1 + \frac{1}{c}\right)\sum_i \bar{p}_i + \frac{2\omega S}{n},\ \sum_i p_i \ge \left(1 - \frac{1}{c}\right)\sum_i \bar{p}_i - \frac{2\omega S}{n}$$

when $k > 1$. For the second statement of the lemma, using what we just proved, we have that with probability $1 - 3\rho$,

$$\gamma_i = \frac{p_i(p_{i+1} + \cdots + p_S)}{p_i + \cdots + p_S} \le \frac{\left(1 + \frac{1}{c}\right)^2 \bar{p}_i(\bar{p}_{i+1} + \cdots + \bar{p}_S) + \frac{2\left(1+\frac{1}{c}\right)\omega S(\bar{p}_i + \cdots + \bar{p}_S)}{n} + \frac{4\omega^2 S^2}{n^2}}{\left(1 - \frac{1}{c}\right)(\bar{p}_i + \cdots + \bar{p}_S) - \frac{2\omega S}{n}}.$$

Now, if $c > \frac{6}{5}$ and indices i are ordered such that $\bar{p}_S \ge 1/S$, then $\bar{p}_i + \cdots + \bar{p}_S \ge 1/S$ for all i. Also, if $n > 12\omega S^2$, we have the following bound on the denominator from the previous expression: $(1 - 1/c)(\bar{p}_i + \cdots + \bar{p}_S) - (2\omega S/n) \ge (1 - (1/c) - (1/6))(\bar{p}_i + \cdots + \bar{p}_S)$, so that

$$\gamma_i \le \frac{\left(1 + \frac{1}{c}\right)^2}{1 - \frac{1}{c} - \frac{1}{6}}\bar{\gamma}_i + \frac{2\left(1 + \frac{1}{c} + \frac{1}{6}\right)\omega S}{1 - \frac{1}{c} - \frac{1}{6}}\frac{}{n}. \quad \square$$

Lemma A.7. *For any fixed $h \in [0, D]^S$, and i, let $\hat{H}_i = \frac{1}{\sum_{j=i}^S \hat{p}_j}\sum_{j=i}^S h_j\hat{p}_j$, $H_i = \frac{1}{\sum_{j=i}^S p_j}\sum_{j=i}^S h_j p_j$, $\bar{H}_i = \frac{1}{\sum_{j=i}^S \bar{p}_j}\sum_{j=i}^S h_j\bar{p}_j$. Then if $\tau_k - 1 \ge 96$, with probability $1 - \rho$,*

$$|(\bar{H}_i - H_i)(\bar{p}_i + \ldots + \bar{p}_S)| \le 2D\sqrt{\log(T/\rho)\frac{(p_i + \cdots + p_S)}{n}} + 3\frac{(\omega S + \log(T/\rho))D}{n}.$$

Moreover, if we also assume that $n > 12\omega S^2$, *where* $\omega = 720\log(T/\rho)$), *then with probability* $1-2\rho$,

$$|(\hat{H}_i - H_i)(\bar{p}_i + \ldots + \bar{p}_S)| \le 3D\sqrt{\log(T/\rho)\frac{(\bar{p}_i + \cdots + \bar{p}_S)}{n}} + 4\frac{(\omega S + \log(T/\rho))D}{n}.$$

Proof. For every $t = 2, \ldots, T, \ell = i, i+1, \ldots, S$, define

$$Z_{t,\ell} = \left(h_\ell \mathbb{1}(s_t = \ell) - h_\ell \frac{p_\ell}{p_i + \cdots + p_S} \cdot \mathbb{1}(s_t \in \{i, \ldots, S\})\right)\mathbb{1}(s_{t-1} = s, a_{t-1} = a),$$
$$Z_t = \sum_{\ell \ge i} Z_{t,\ell}.$$

Then,

$$\frac{\sum_{t=2}^{\tau_k} Z_t}{n} = \sum_{\ell \ge i} h_\ell \hat{p}_\ell - \sum_{\ell \ge i} h_\ell \frac{p_\ell}{p_i + \cdots + p_S} \cdot (\hat{p}_i + \ldots + \hat{p}_S) = (\hat{H}_i - H_i)(\hat{p}_i + \ldots + \hat{p}_S).$$

where we used Fact A.1 for the last equality. Now, $E[Z_t | s_{t-1}, a_{t-1}] = \sum_{\ell \ge i} E[Z_{t,\ell} | s_{t-1}, a_{t-1}] = 0$. Also, we observe that for any t, $Z_{t,\ell}$ and $Z_{t,j}$ for any $\ell \ne j, \ell, j \ge i$ are negatively correlated given the current state and action:

$$\begin{aligned}
\mathbb{E}[Z_{t,\ell} Z_{t,j} | s_{t-1}, a_{t-1}] &= h_\ell h_j \mathbb{E}\Bigg[\mathbb{1}(s_t = \ell)\mathbb{1}(s_t = j) - \mathbb{1}(s_t = j)\frac{p_\ell}{p_i + \cdots + p_S} \cdot \mathbb{1}(s_t \in \{i, \ldots, S\}) \\
&\quad - \mathbb{1}(s_t = \ell)\frac{p_j}{p_i + \cdots + p_S} \cdot \mathbb{1}(s_t \in \{i, \ldots, S\}) \\
&\quad + \frac{p_j p_\ell}{(p_i + \cdots + p_S)^2} \cdot \mathbb{1}(s_t \in \{i, \ldots, S\}) | s_{t-1}, a_{t-1}\Bigg] \\
&= h_\ell h_j \left(-\frac{2p_j p_\ell}{p_i + \cdots + p_S} + \frac{p_\ell p_j}{(p_i + \cdots + p_S)^2} \cdot \mathbb{1}(s_t \in \{i, \ldots, S\})\right) \\
&= h_\ell h_j \left(-\frac{p_j p_i}{p_i + \cdots + p_S}\right) \\
&\le 0.
\end{aligned}$$

Also,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=2}^{\tau_k} Z_{t,\ell}^2 | s_{t-1} = s, a_{t-1} = a\right] &= h_\ell^2 \sum_{t=1}^{\tau_k} \mathbb{1}(s_{t-1} = s, a_{t-1} = a)\left(p_\ell - \frac{p_\ell^2}{(p_i + \cdots + p_S)^2}(p_i + \cdots + p_S)\right) \\
&= h_\ell^2 \sum_{t=1}^{\tau_k} \mathbb{1}(s_{t-1} = s, a_{t-1} = a)\frac{p_\ell(\sum_{j \ge i, j \ne k} p_j)}{p_i + \cdots + p_S} \\
&= n h_\ell^2 \frac{p_\ell(\sum_{j \ge i, j \ne k} p_j)}{p_i + \cdots + p_S} \\
&= n h_\ell^2 \frac{p_\ell(\sum_{j \ge i, j \ne k} p_j)}{p_i + \cdots + p_S} \\
&\le n D^2 p_\ell.
\end{aligned}$$

Therefore,

$$V_{\tau_k} := \sum_{t=2}^{\tau_k} E[Z_t^2 | s_{t-1} = s, a_{t-1} = a] \le \sum_{t=1}^{\tau_k} \sum_{\ell \ge i} \mathbb{E}[Z_{t,\ell}^2 | s_{t-1} = s, a_{t-1} = a] \le n D^2 (p_i + \cdots + p_S).$$

Then, applying Bernstein's inequality (refer to Corollary 1) to bound $|M_{\tau_k}| = |\sum_{t=2}^{\tau_k} Z_t|$, we get the following bound on $(1/n)\sum_{t=2}^{\tau_k} Z_t = (\hat{H}_i - H_i)(\hat{p}_i + \ldots + \hat{p}_S)$ with probability $1-\rho$:

$$|(\hat{H}_i - H_i)(\hat{p}_i + \ldots + \hat{p}_S)| = \left|\frac{1}{n}\sum_{t=2}^{\tau_k} Z_t\right| \le 2D\sqrt{\log(T/\rho)\frac{(p_i + \cdots + p_S)}{n}} + 3D\frac{\log(T/\rho)}{n}.$$

Also,

$$|\hat{H}_i - \bar{H}_i| = \left|\sum_{\ell=i}^{S}\left(\frac{\hat{p}_\ell}{\hat{p}_i + \cdots + \hat{p}_S} h_\ell - \frac{\bar{p}_\ell}{\bar{p}_i + \cdots + \bar{p}_S} h_\ell\right)\right| \le \frac{\omega S D}{n(\hat{p}_i + \cdots + \hat{p}_S)},$$

Combining,

$$|(\tilde{H}_i - H_i)(\hat{p}_i + \ldots + \hat{p}_S)| \le 2D\sqrt{\log(T/\rho)\frac{(p_i + \cdots + p_S)}{n}} + 3D\frac{\log(T/\rho)}{n} + \frac{\omega SD}{n}.$$

Replacing $\hat{p}_i$ by $\bar{p}_i$,

$$|(\tilde{H}_i - H_i)(\bar{p}_i + \ldots + \bar{p}_S)| \le 2D\sqrt{\log(T/\rho)\frac{(p_i + \cdots + p_S)}{n}} + 3\frac{(\omega S + \log(T/\rho))D}{n},$$

with probability $1-\rho$.

Now, if we also have that $\omega = 720\log(T/\rho)$ and $n > 12\omega S^2$, using Lemma A.6 with $c = 3$ to replace p_i by $\bar{p}_i$, with probability $1-2\rho$,

$$|(\tilde{H}_i - H_i)(\bar{p}_i + \ldots + \bar{p}_S)| \le 3D\sqrt{\log(T/\rho)\frac{(\bar{p}_i + \cdots + \bar{p}_S)}{n}} + 4\frac{(\omega S + \log(T/\rho))D}{n}. \quad \square$$

Appendix B. Some Useful Facts and Known Inequalities

Fact B.1 (Bernstein's Inequality from Seldin et al. [34] Lemma 11 and Corollary 12). Let $Z_1, Z_2, \ldots, Z_\tau$ be a bounded martingale difference sequence so that $|Z_i| \le K$ and $\mathbb{E}[Z_i|\mathcal{F}_{i-1}] = 0$. Define $M_\tau = \sum_{i=1}^{\tau} Z_i$ and $V_\tau = \sum_{i=1}^{\tau} \mathbb{E}[(Z_i)^2|\mathcal{F}_{i-1}]$. For any $c > 1$ and $\delta \in (0,1)$, with probability greater than $1-\delta$, if

$$\sqrt{\frac{\ln\frac{2\nu}{\delta}}{(e-2)V_\tau}} \le \frac{1}{K},$$

then

$$|M_\tau| \le (1+c)\sqrt{(e-2)V_\tau \ln\frac{2\nu}{\delta}},$$

otherwise,

$$|M_\tau| \le 2K\ln\frac{2\nu}{\delta},$$

where

$$\nu = \left\lceil \frac{\ln\left(\sqrt{\frac{(e-2)\tau}{\ln\frac{2}{\delta}}}\right)}{\ln c} \right\rceil + 1.$$

Corollary 1 (to Bernstein's Inequality). *Let Z_i for $i = 1, \cdots, \tau$, M_τ, and V_τ as previously. For $\tau \ge 96$ and $\delta \in (0,1)$, with probability greater than $1-\delta$,*

$$|M_\tau| \le 2\sqrt{V_\tau \ln\frac{\tau}{\delta}} + 3K\ln\frac{\tau}{\delta}.$$

Proof. Applying Bernstein's inequality with $c = 1 + \frac{4}{\tau}$, with probability greater than $1-\delta$,

$$\begin{aligned}
|M_\tau| &\le (1+c)\sqrt{(e-2)V_\tau \ln\frac{2\nu}{\delta}} + 2K\ln\frac{2\nu}{\delta} \\
&\le (1+c)\sqrt{(e-2)V_\tau \ln\frac{\tau^{\frac{4}{3}}}{\delta}} + 2K\ln\frac{\tau^{\frac{4}{3}}}{\delta} \\
&\le (1+c)\sqrt{(e-2)\frac{4}{3}V_\tau \ln\frac{\tau}{\delta}} + 3K\ln\frac{\tau}{\delta} \\
&\le 2\sqrt{V_\tau \ln\frac{\tau}{\delta}} + 3K\ln\frac{\tau}{\delta},
\end{aligned}$$

where

$$\nu = \left\lceil \frac{\ln\left(\sqrt{\frac{(e-2)\tau}{\ln^2\frac{1}{\delta}}}\right)}{\ln c} \right\rceil + 1 = \left\lceil \frac{\tau}{2}\ln\left(\sqrt{\frac{(e-2)\tau}{\ln^2\frac{1}{\delta}}}\right) \right\rceil + 1 \le \frac{\tau}{2}\ln\left(\sqrt{\frac{(e-2)\tau}{\ln 2}}\right) + 2 \le \frac{1}{2}\tau^{\frac{4}{3}}. \quad \square$$

Fact B.2 (Multiplicative Chernoff Bound;Kleinberg et al. [25], Lemma 4.9). Consider n independent and identially distributed random variables $X_1, \cdots, X_n$ on $[0,1]$. Let μ be their mean and let X be their average. Then for any $\alpha > 0$, the following holds:

$$P(|X-\mu| < r(\alpha, X) < 3r(\alpha,\mu)) > 1 - e^{\Omega(\alpha)},$$

where $r(\alpha, x) = \sqrt{\frac{\alpha x}{n}} + \frac{\alpha}{n}$.

More explicitly, we have that with probability $1-\rho$,

$$|X-\mu| < \sqrt{\frac{3\log(2/\rho)X}{n}} + \frac{3\log(2/\rho)}{n}.$$

Fact B.3 (Cantelli's Inequality). Let X be a real-valued random variable with expectation μ and variance σ^2. Then $P(X-\mu \ge \lambda) \le \frac{\sigma^2}{\sigma^2+\lambda^2}$ for $\lambda > 0$ and $P(X-\mu \ge \lambda) \ge 1 - \frac{\sigma^2}{\sigma^2+\lambda^2}$ for $\lambda < 0$.

Fact B.4 (Berry-Esseen Theorem). Let $X_1, X_2, \ldots, X_n$ be independent random variables with $\mathbb{E}[X_i] = 0$, $\mathbb{E}[X_i^2] = \sigma_i^2 > 0$, and $\mathbb{E}[|X_i|^3] = \rho_i < \infty$. Let

$$S_n = \frac{X_1 + X_2 + \ldots + X_n}{\sqrt{\sigma_1^2 + \ldots + \sigma_n^2}}$$

and denote F_n the cumulative distribution function of S_n and Φ the cumulative distribution function of the standard normal distribution. Then for all n, there exists an absolute constant C_1 such that

$$sup_{x\in R}|F_n(x) - \Phi(x)| \le C_1\psi_1,$$

where $\psi_1 = (\sum_{i=1}^n \sigma_i^2)^{-1/2}\max_{1\le i\le n}\frac{\rho_i}{\sigma_i^2}$. The best upper bound on C_1 known is $C_1 \le 0.56$ (Shevtsova [35]).

Fact B.5 (Abramowitz and Stegun [2] 26.5.21). Consider the regularized incomplete Beta function $I_z(a,b)$ (cdf) for the Beta random variable with parameters (a, b). For any z such that $(a+b-1)(1-z) \ge 0.8$, $I_z(a,b) = \Phi(y) + \epsilon$, with $|\epsilon| < 0.005$ if $a+b > 6$. Here Φ is the standard normal cdf with

$$y = \frac{3\left[w_1\left(1-\frac{1}{9b}\right) - w_2\left(1-\frac{1}{9a}\right)\right]}{\left[\frac{w_1^2}{b} + \frac{w_2^2}{a}\right]^{1/2}},$$

where $w_1 = (bz)^{1/3}$ and $w_2 = [a(1-z)]^{1/3}$.

Definition B.1. For any X and Y real-valued random variables, X is stochastically optimistic for Y if for any $u : R \to R$ convex and increasing $\mathbb{E}[u(X)] \ge \mathbb{E}[u(Y)]$.

Lemma B.1 (Gaussian vs. Dirichlet Optimism, from Osband et al. [29], Lemma 1). *Let $Y = P^T V$ for $V \in [0,1]^S$ fixed and $P \sim$ Dirichlet(α) with $\alpha \in R_+^S$ and $\sum_{i=1}^S \alpha_i \ge 2$. Let $X \sim N(\mu, \sigma^2)$ with $\mu = \frac{\sum_{i=1}^S \alpha_i V_i}{\sum_{i=1}^S \alpha_i}$, $\sigma^2 = \left(\sum_{i=1}^S \alpha_i\right)^{-1}$, then X is stochastically optimistic for Y.*

Lemma B.2 (Gaussian vs. Beta Optimism from Osband et al. [29], Lemma 6). *Let $\tilde{Y} \sim$ Beta(α,β) for any $\alpha,\beta > 0$ and $X \sim N\left(\frac{\alpha}{\alpha+\beta}, \frac{1}{\alpha+\beta}\right)$. Then X is stochastically optimistic for $\tilde{Y}$ whenever $\alpha+\beta \ge 2$.*

Lemma B.3 (Dirichlet vs. Beta Optimism from Osband et al. [29], Lemma 5). *Let $y = p^T v$ for some random variable $p \sim$ Dirichlet(α) and constants $v \in \mathcal{R}^d$ and $\alpha \in \mathcal{N}^d$. Without loss of generality, assume $v_1 \le v_2 \le \cdots \le v_d$. Let $\tilde{\alpha} = \sum_{i=1}^d \alpha_i(v_i - v_1)/(v_d - v_1)$ and $\tilde{\beta} = \sum_{i=1}^d \alpha_i(v_d - v_i)/(v_d - v_1)$. Then, there exists a random variable $\tilde{p} \sim$ Beta$(\tilde{\alpha}, \tilde{\beta})$ such that, for $\tilde{y} = \tilde{p}v_d + (1-\tilde{p})v_1$, $\mathbb{E}[\tilde{y}\,|\,y] = \mathbb{E}[y]$.*

Lemma B.4. *If $\mathbb{E}[X] = \mathbb{E}[Y]$ and X is stochastically optimistic for Y, then $-X$ is stochastically optimistic for $-Y$.*

Proof. By lemma 3.3 in Osband et al. [29], X stochastically optimistic for Y is equivalent to having $X =_D Y + A + W$ with $A \geq 0$ and $\mathbb{E}[W|Y + A] = 0$ for all values $y + a$. Taking expectation of both sides, we get that $\mathbb{E}[X] = \mathbb{E}[Y] + \mathbb{E}[A] + \mathbb{E}[W]$ and because $\mathbb{E}[X] = \mathbb{E}[Y] = 0$ and $\mathbb{E}[W] = \mathbb{E}[\mathbb{E}[W|Y + A]] = 0$ we get that $\mathbb{E}[A] = 0$. Because $A \geq 0$, $A = 0$. Also, $\mathbb{E}[W|Y = y] = 0$ for all y.

Now we can show that $-X$ is stochastically optimistic for $-Y$ as follows: From the previous statements, $-X =_D -(Y + A + W) = -Y + (-W)$. Then for all y', $\mathbb{E}[-W| - Y = y'] = -\mathbb{E}[W|Y = -y'] = 0$ by definition of W. Therefore, $-X$ is stochastically optimistic for $-Y$. □

Corollary B.2. *Let Y be any distribution with mean μ such that $X \sim N(\mu, \sigma^2)$ is stochastically optimistic for Y. Then with probability $1 - \rho$,*

$$|Y - \mu| \leq \sqrt{2\sigma^2 \log(2/\rho)}.$$

Proof. For any $s > 0$, and t, and applying Markov's inequality,

$$P(Y - \mu > t) = P(Y > \mu + t) = P(e^{sY} > e^{s(\mu+t)}) \leq \frac{\mathbb{E}[e^{sY}]}{e^{s(\mu+t)}}.$$

By Definition B.1, taking $u(a) = e^{sa}$, which is a convex and increasing function, $\mathbb{E}[e^{sY}] \leq \mathbb{E}[e^{sX}]$, and hence

$$P(Y - \mu > t) \leq \frac{\mathbb{E}[e^{sX}]}{e^{s(\mu+t)}} = \frac{e^{\mu s + \frac{1}{2}\sigma^2 s^2}}{e^{s(\mu+t)}} = e^{\frac{1}{2}\sigma^2 s^2 - st}.$$

Because the previous expression holds for all $s > 0$, using $s = \frac{t}{\sigma^2}$, $P(Y - \mu > t) \leq e^{-\frac{t^2}{2\sigma^2}}$.

Similarly, for the lower tail bound, we have for any $s > 0$,

$$P(Y - \mu < -t) = P(-Y > -\mu + t) = P(e^{s(-Y)} > e^{s(-\mu+t)}) \leq \frac{\mathbb{E}[e^{s(-Y)}]}{e^{s(-\mu+t)}}.$$

By Lemma B.4, $-X$ is stochastically optimistic for $-Y$, so $\mathbb{E}[e^{s(-Y)}] \leq \mathbb{E}[e^{s(-X)}]$, and hence

$$P(Y - \mu < -t) \leq \frac{\mathbb{E}[e^{s(-X)}]}{e^{s(-\mu+t)}} = \frac{e^{-\mu s + \frac{1}{2}\sigma^2 s^2}}{e^{s(-\mu+t)}} = e^{\frac{1}{2}\sigma^2 s^2 - st}.$$

Again, letting $s = \frac{t}{\sigma^2}$, $P(Y - \mu < -t) \leq e^{-\frac{t^2}{2\sigma^2}}$.

Then, for $t = \sqrt{2\sigma^2 \log(2/\rho)}$, we have that

$$P(|Y - \mu| \leq \sqrt{2\sigma^2 \log(2/\rho)}) \geq 1 - \rho. \quad \square$$

Lemma B.5 (Lemma 19 in Jaksch et al. [20]). *For any sequence of numbers $z_1, \ldots, z_n$ with $0 \leq z_k \leq Z_{k-1} := \max\{1, \sum_{i=1}^{k-1} z_i\}$,*

$$\sum_{k=1}^{n} \frac{z_k}{\sqrt{Z_{k-1}}} \leq (\sqrt{2} + 1)\sqrt{Z_n}.$$

Endnote

[1] Worst-case regret is a strictly stronger notion of regret than Bayesian regret. However, a caveat is that the reward distributions are assumed to be bounded or sub-Gaussian to prove worst-case regret bounds. Conversely, the Bayesian regret bounds in the previously mentioned literature allow more general (known) priors on the reward distributions with possibly unbounded support. Bayesian regret bounds under such more general reward distributions are incomparable to the worst-case regret bounds presented here.

References

[1] Abbasi-Yadkori Y, Szepesvari C (2014) Bayesian optimal control of smoothly parameterized systems: The lazy posterior sampling algorithm. Preprint, submitted June 16, https://arxiv.org/abs/1406.3926.

[2] Abramowitz M, Stegun IA (1964) *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*, vol. 55 (Courier Corporation).

[3] Agrawal S, Goyal N (2012) Analysis of Thompson sampling for the multi-armed bandit problem. Mannor S, Srebro N, Williamson RC, eds. *Proc. 25th Annual Conf. on Learn. Theory* (PMLR), 39.1–39.26.

[4] Agrawal S, Goyal N (2013a) Further optimal regret bounds for Thompson sampling. Carvalho CM, Ravikumar P, eds. *Proc. 16th Internat. Conf. Artificial Intelligence and Statistics* (PMLR), 99–107.

[5] Agrawal S, Goyal N (2013b) Thompson sampling for contextual bandits with linear payoffs. Dasgupta S, McAllester D, eds. *Proc. 30th Internat. Conf. on Machine Learn.* (PMLR).

[6] Albert I, Denis JB (2012) Dirichlet and multinomial distributions: Properties and uses in jags. *Unite Mathematiques et Informatique Appliquees* (INRA), 2012–2015.

[7] Asmuth J, Li L, Littman ML, Nouri A, Wingate D (2009) A Bayesian sampling approach to exploration in reinforcement learning. *Proc. 25th Conf. on Uncertainty in Artificial Intelligence* (AUAI Press), 19–26.

[8] Azar MG, Osband I, Munos R (2017) Minimax regret bounds for reinforcement learning. Preprint, submitted July 1, https://arxiv.org/abs/1703.05449.
[9] Bartlett PL, Tewari A (2009) REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. *Proc. 25th Conf. on Uncertainty in Artificial Intelligence* (AUAI Press), 35–42.
[10] Brafman RI, Tennenholtz M (2002) R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *J. Machine Learn. Res.* 3(Oct):213–231.
[11] Bubeck S, Cesa-Bianchi N (2012) Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations Trends Machine Learn.* 5(1):1–122.
[12] Bubeck S, Liu CY (2013) Prior-free and prior-dependent regret bounds for Thompson sampling. *Adv. Neural Inform. Processing Systems* 26: 638–646.
[13] Burnetas AN, Katehakis MN (1997) Optimal adaptive policies for Markov decision processes. *Math. Oper. Res.* 22(1):222–255.
[14] Chapelle O, Li L (2011) An empirical evaluation of Thompson sampling. *Adv. Neural Inf. Processing Systems* 24:2249–2257.
[15] Dann C, Brunskill E (2015) Sample complexity of episodic fixed-horizon reinforcement learning. *Adv. Neural Inform. Processing Systems* 28: 2818–2826.
[16] Fonteneau R, Korda N, Munos R (2013) An optimistic posterior sampling strategy for bayesian reinforcement learning. *Proc. NIPS Workshop on Bayesian Optimization*.
[17] Fruit R, Pirotta M, Lazaric A (2018) Near optimal exploration-exploitation in non-communicating markov decision processes. *Adv. Neural Inform. Processing Systems* 31:31.
[18] Fruit R, Pirotta M, Lazaric A (2020) Improved analysis of ucrl2 with empirical bernstein inequality. Preprint, submitted July 10, https://arxiv.org/abs/2007.05456.
[19] Grinstead CM, Snell JL (2012) *Introduction to Probability* (American Mathematical Society, Providence, RI).
[20] Jaksch T, Ortner R, Auer P (2010) Near-optimal regret bounds for reinforcement learning. *J. Machine Learn. Res.* 11(Apr):1563–1600.
[21] Kakade SM, Wang M, Yang LF (2018) Variance reduction methods for sublinear reinforcement learning. Preprint, submitted June 27, https://arxiv.org/abs/1802.09184.
[22] Kakade SM, et al (2003) *On the sample complexity of reinforcement learning*. PhD thesis, University of London, London.
[23] Kaufmann E, Korda N, Munos R (2012) Thompson sampling: An optimal finite time analysis. *Proc. Internat. Conf. on Algorithmic Learn. Theory* (Springer).
[24] Kearns MJ, Singh SP (1999) Finite-sample convergence rates for Q-learning and indirect algorithms. *Adv. Neural Inform. Processing Systems* 11:996–1002.
[25] Kleinberg R, Slivkins A, Upfal E (2008) Multi-armed bandits in metric spaces. *Proc. 40th Annual ACM Sympos. on Theory of Comput.* (ACM, New York), 681–690.
[26] Ortner R (2020) Regret bounds for reinforcement learning via Markov chain concentration. *J. Artificial Intelligence Res.* 67:115–128.
[27] Osband I, Van Roy B (2017) Why is posterior sampling better than optimism for reinforcement learning. Preprint, submitted June 13, https://arxiv.org/abs/1607.00215.
[28] Osband I, Russo D, Van Roy B (2013) (More) efficient reinforcement learning via posterior sampling. *Adv. Neural Inform. Processing Systems* 26:3003–3011.
[29] Osband I, Van Roy B, Wen Z (2016) Generalization and exploration via randomized value functions. Preprint, submitted February 15, https://arxiv.org/abs/1402.0635.
[30] Ouyang Y, Gagrani M, Nayyar A, Jain R (2017) Learning unknown Markov decision processes: A Thompson sampling approach. Preprint, submitted September 14, https://arxiv.org/abs/1709.04570.
[31] Puterman ML (2014) *Markov Decision Processes: Discrete Stochastic Dynamic Programming* (John Wiley & Sons, New York).
[32] Russo D, Van Roy B (2014) Learning to optimize via posterior sampling. *Math. Oper. Res.* 39(4):1221–1243.
[33] Russo D, Van Roy B (2016) An information-theoretic analysis of Thompson sampling. *J. Machine Learn. Res.* 17(4):1–30.
[34] Seldin Y, Laviolette F, Cesa-Bianchi N, Shawe-Taylor J, Auer P (2012) PAC-Bayesian inequalities for martingales. *IEEE Trans. Inform. Theory* 58(12):7086–7093.
[35] Shevtsova IG (2010) An improvement of convergence rate estimates in the Lyapunov theorem. *Doklady Math.* 82(3):862–864.
[36] Strehl AL, Littman ML (2005) A theoretical analysis of model-based interval estimation. *Proc. 22nd Internat. Conf. on Machine Learn.* (ACM, New York), 856–863.
[37] Strehl AL, Littman ML (2008) An analysis of model-based interval estimation for Markov decision processes. *J. Comput. System Sci.* 74(8): 1309–1331.
[38] Talebi MS, Maillard OA (2018) Variance-aware regret bounds for undiscounted reinforcement learning in mdps. *Algorithmic Learning Theory* (PMLR), 770–805.
[39] Tewari A, Bartlett PL (2008) Optimistic linear programming gives logarithmic regret for irreducible MDPs. *Adv. Neural Inform. Processing Systems* 20:1505–1512.
[40] Thompson WR (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3/4):285–294.
[41] Zhang Z, Ji X (2019) Regret minimization for reinforcement learning by evaluating the optimal bias function. Preprint, submitted December 28, https://arxiv.org/abs/1906.05110.