Journal of Machine Learning Research 26 (2025) 1-71 Submitted 11/24; Published 11/25

Concentration of Cumulative Reward in Markov Decision Processes

Borna Sayedana BORNA.SAYEDANA@MAIL.MCGILL.CA
Université de Montréal,
and Mila – Quebec AI Institute
Montreal, QC, H2S 3H1, Canada.

Peter E. Caines PETERC@CIM.MCGILL.CA
Department of Electrical and Computer Engineering,
McGill University,
Montreal, QC, H3A 0E9, Canada.

Aditya Mahajan ADITYA.MAHAJAN@MCGILL.CA
Department of Electrical and Computer Engineering,
McGill University,
Montreal, QC, H3A 0E9, Canada.

Editor: Maxim Raginsky

Abstract

In this paper, we investigate the concentration properties of cumulative reward in Markov Decision Processes (MDPs), focusing on both asymptotic and non-asymptotic settings. We introduce a unified approach to characterize reward concentration in MDPs, covering both infinite-horizon settings (i.e., average and discounted reward frameworks) and finite-horizon setting. Our asymptotic results include the law of large numbers, the central limit theorem, and the law of iterated logarithms, while our non-asymptotic bounds include Azuma-Hoeffding-type inequalities and a non-asymptotic version of the law of iterated logarithms. Additionally, we explore two key implications of our results. First, we analyze the sample path behavior of the difference in rewards between any two stationary policies. Second, we show that two alternative definitions of regret for learning policies proposed in the literature are *rate-equivalent*. Our proof techniques rely on a martingale decomposition of cumulative reward, properties of the solution to the policy evaluation fixed-point equation, and both asymptotic and non-asymptotic concentration results for martingale difference sequences.

Keywords: Concentration of Rewards, Markov Decision Processes, Reinforcement Learning, Average Reward Infinite-Horizon MDPs

1. Introduction

Reinforcement learning is a machine learning framework in which an agent learns to make optimal sequential decisions by repeatedly interacting with its environment. This approach is particularly effective for addressing problems with complex dynamic environments. The standard mathematical model for reinforcement learning is Markov Decision Processes (MDPs). In an MDP, the agent takes an action at each time step, receives an instantaneous

reward, and transitions to the next state based on a Markovian dynamic that depends on the current state and action.

The existing literature on MDP theory primarily focuses on analyzing and maximizing the *expected* cumulative reward, resulting in methods that emphasize the system's *average* behavior. While this approach is useful in many domains, it may fall short in high-stakes applications, where an agents' decisions may lead to costly consequences. Such scenarios arise in applications such as safety-critical engineering systems and decision-making processes in finance and healthcare. Different approaches have emerged to understand the behavior of MDPs beyond the expected reward. Broadly, these approaches can be categorized as follows: (i) risk-sensitive control, in which the agent aims to identify policies that minimize a specific *risk measure*; (ii) distributional reinforcement learning, in which the distribution of cumulative *discounted* reward is estimated and controlled; (iii) Markov reward processes, in which the *asymptotic* distributional and sample-path properties of the cumulative reward in the infinite-horizon *average* reward framework are investigated. We elaborate on each of these approaches below.

Risk-sensitive control. In the risk-neutral framework the agent's goal is to maximize the expected cumulative reward. In the risk-sensitive framework, the agent's goal is to minimize a risk measure that captures other statistical properties (e.g., variance, tail probability, etc.) of the reward in addition to the mean. There are three primary risk functionals studied in the literature (Wang and Chapman, 2022): (i) the exponential utility functional, which is an increasing and convex mapping of the cost function. Under certain conditions, this functional can be used to model the mean-variance trade-off in decision-making problems. Risk-averse control using exponential utility functional has been studied in Howard and Matheson (1972); Jacobson (1973); Whittle (1981); Coraluppi and Marcus (1999); Borkar (2002); Bäuerle and Rieder (2014). (ii) Quantile-based risk functionals such as Value at Risk (VaR) and Conditional Value at Risk (CVaR), which characterize the probability or expectation of the cost exceeds a given threshold, capturing the tail behavior of the cost distribution. Risk-averse control using quantile-based risk functionals has been studied in Bäuerle and Ott (2011); Chow et al. (2015); Miller and Yang (2021); Bäuerle and Glauner (2021); Chapman et al. (2022). (iii) Recursive risk functionals which model the risk at every stage and result in a dynamic programming type solution. Risk-averse control using recursive risk functionals has been studied in Ruszczyński (2010); Singh et al. (2018); Bäuerle and Glauner (2022). For a comprehensive survey on risk-sensitive control and RL, please refer to Wang and Chapman (2022); Biswas and Borkar (2023).

Distributional RL. The second approach focuses on estimating various statistical properties of the discounted cumulative reward in the infinite-horizon discounted reward framework. Early works such as Sobel (1982); Chung and Sobel (1987) derive a Bellman-type equation to compute the variance of the discounted cumulative reward. More recent works treat the asymptotic discounted cumulative reward as a random variable and use various methods to approximate the distribution or compute its important statistics such as quantiles. The approximate distribution is then used in reinforcement learning algorithms (Morimura et al., 2010a,b; Bellemare et al., 2017; Rowland et al., 2018; Dabney et al., 2018a,b; Rowland et al., 2019; Bellemare et al., 2019; Lyle et al., 2019; Yang et al., 2019; Farahmand, 2019; Duan et al., 2021; Lhéritier and Bondoux, 2021; Nguyen et al., 2022;

Rowland et al., 2023, 2024). For a comprehensive review of the algorithms and theoretical results in distributional RL, please refer to Bellemare et al. (2023).

Markov Chains. For a fixed stationary Markov policy, any MDP may be reduced to a Markov reward process, i.e. a Markov chain induced on the state space and an associated reward process. As a result, there is a close connection between the MDP theory and the theory of Markov chains. The asymptotic sample path and distributional behavior of functionals of Markov chains are extensively studied in the literature. For example, the CLT results for Markov chains are established in Chung (1967); Cogburn (1972); Maigret (1978); Niemi and Nummelin (1982); Kipnis and Varadhan (1986); Maxwell and Woodroofe (2000); Landim (2003); Jones (2004); Meyn and Tweedie (2012); Duflo (2013); Srikant (2025). The rate of convergence in CLT for geometrically ergodic Markov chains is studied in Kontoyiannis and Meyn (2003, 2005). In parallel, the asymptotic behavior of martingales are studied in the martingale theory, e.g., in Neveu (1975); Hall and Heyde (1980); Billingsley (2013). As discussed in Meyn and Tweedie (2012) there is a close connection between the results in Markov chains and their counterparts in martingale theory. As an example, one way to prove the CLT for Markov chains is to use the martingale decomposition arising from the Poisson equation. This approach is used in e.g., Mandl (1971); Mandl and Lausmanova (1991); Hernández-Lerma and Lasserre (2012); Duflo (2013); Maigret (1978). Recently the martingale approach is used to derive a central limit theorem for the Linear Quadratic Regulation (LQR) problem in Sayedana et al. (2024).

Our Work. In this paper, we provide a unified approach for characterizing both asymptotic and non-asymptotic reward concentration in infinite-horizon average reward, infinite-horizon discounted reward, and finite-horizon frameworks. Our results cover asymptotic concentration like LLN, CLT, and LIL, along with non-asymptotic bounds, including Azuma-Hoeffding-type inequalities and a non-asymptotic version of the Law of Iterated Logarithms for the average reward setting. Building upon these concentration results, we explore two of their key implications: (1) the sample path difference of rewards between two policies, and (2) the impact of these findings on the regret analysis of reinforcement learning algorithms. We derive similar non-asymptotic upper-bounds for discounted reward and finite-horizon setups.

Comparison. There are two key distinctions between our work and the existing literature. (i) We establish *non-asymptotic* concentration results for the cumulative reward process. In contrast, studies in distributional RL typically analyze the *asymptotic* behavior of the discounted cumulative reward, whereas in the average-reward and Markov reward process settings, prior works focus on the *asymptotic* distributions of the cumulative reward process. To the best of our knowledge, no *non-asymptotic* concentration result have been reported in the literature for MDPs. (ii) We develop a *unified framework* that enables the derivation of concentration results across the three principle MDP frameworks. In contrast, methods developed in distributional RL that rely on contraction mapping theorems do not extend naturally to the average-reward setting, and techniques based on Markov chain analysis cannot be directly applied to finite-horizon problems.

Proof Approach Our proofs rely on a martingale decomposition similar to the one originating from the Poisson equation in Markov chains (e.g. in Meyn and Tweedie (2012)). Such decomposition enables us to interpret the cumulative reward process both as a martingale and as a functional of an underlying Markov chain. In this paper, we adopt the

martingale viewpoint rather than the Markov chain one. We make such a choice since the resulting non-asymptotic bounds solely depend on the statistical properties of the value function. As a result, these bounds can be efficiently computed using existing numerical methods for value function computation.

We also use our results to clarify a nuance in the definition of regret in average reward infinite-horizon reinforcement learning. In this setting, regret is defined as the difference between the *expected* reward obtained by the optimal policy minus the (sample-path) cumulative reward obtained by the learning algorithm as a function of time. The standard results establish that this regret is lower-bounded by $\Omega(\sqrt{D|\mathcal{S}||\mathcal{A}|T})$ and upper bounded by $\tilde{\mathcal{O}}(D|\mathcal{S}|\sqrt{|\mathcal{A}|T})$ (Jaksch et al., 2010), where T denotes the horizon, $|\mathcal{S}|$ denotes the number of states, $|\mathcal{A}|$ denotes the number of actions, and D denotes the diameter of thr MDP. Various refinements of these results have been considered in the literature (Auer and Ortner, 2006; Filippi et al., 2010; Bartlett and Tewari, 2012; Russo and Van Roy, 2014; Osband et al., 2013; Lakshmanan et al., 2015; Osband et al., 2016; Ouyang et al., 2017; Theocharous et al., 2017; Agrawal and Jia, 2017; Talebi and Maillard, 2018; Fruit et al., 2018; Zhang and Ji, 2019; Qian et al., 2019; Fruit, 2019; Zanette and Brunskill, 2019; Fruit et al., 2020; Bourel et al., 2020; Zhang and Xie, 2023; Boone and Zhang, 2024).

There is a more appropriate notion of regret in applications which are driven by an independent exogenous noise process such as inventory management problems where the dynamics are driven by an exogenous demand process and linear quadratic regulation problems where the dynamics are driven by an exogenous disturbance process. In such applications, it is more appropriate to compare the cumulative reward obtained by the optimal policy with cumulative reward obtained by the learning algorithm *under the same realization of the exogenous noise*. For example, in an inventory management problem, one may ask how worse is a learning algorithm compared to the (expected-reward) optimal policy on a specific realization of the demand process. This notion of regret has received significantly less attention in the literature (Abbasi-Yadkori et al., 2019; Talebi and Maillard, 2018). We show that a consequence of our results is that the two notions of regret are rate-equivalent. A similar result was claimed without a proof in Talebi and Maillard (2018).

1.1 Contributions

The contributions of this paper can be summarized as follows:

1. We establish the asymptotic concentration of cumulative reward in average reward MDPs, deriving the law of large numbers, the central limit theorem, and the law of iterated logarithm for a class of stationary policies. Compared to the existing asymptotic results in the literature which use Markov chain theory, we provide a simpler proof which leverages a martingale decomposition for the cumulative reward along with the asymptotic concentration of measures for martingale sequences.

2. We derive policy-dependent and policy-independent non-asymptotic concentration bounds for the cumulative reward in average reward MDPs. These bounds establish an Azuma-Hoeffding-type inequality for the rewards along with a non-asymptotic version of law of iterated logarithm. Although these results apply to a broad subset of stationary policies, we show that for communicating MDPs, these bounds extend

to any stationary deterministic policy. We use the established concentration results to characterize the sample path behavior of the performance difference of any two stationary policies. As a corollary of this result, we show that the difference between cumulative reward of any two optimal policies is upper-bounded by $\mathcal{O}(\sqrt{T})$ with high probability.

3. We investigate the difference between two notions of regret in the reinforcement learning literature, cumulative regret and interim cumulative regret. By analyzing the sample path behavior, we establish that both asymptotically and non-asymptotically, this difference is upper-bounded by $\tilde{\mathcal{O}}(\sqrt{T})$. This result implies that, if a reinforcement learning algorithm has a regret upper bound of $\tilde{\mathcal{O}}(\sqrt{T})$ under one definition, the same rate applies to the other, in both of the asymptotic and non-asymptotic frameworks. While this equivalency was claimed in the literature without a proof, our concentration results provide a formal proof for this relation.

4. Lastly, we investigate several extensions of our results to other frameworks. In particular, we derive non-asymptotic concentration bounds for the cumulative reward in the infinite-horizon discounted reward and finite-horizon MDP frameworks. These bounds include an Azuma-Hoeffding-type inequality along with a non-asymptotic version of the law of iterated logarithm. Using the vanishing discount analysis, we show that under appropriate conditions, the concentration bounds for discounted reward MDPs approach to the concentration bounds for the average reward MDPs as the discount factor approaches 1. Moreover, we establish the non-asymptotic concentration bounds for models with stochastic reward, i.e., models in which reward is a function of an exogenous process in addition to state and action.

1.2 Organization

The rest of this paper is organized as follows. The problem formulation, along with the underlying assumptions, are presented in Sec. 2. The main results for the average reward setting are presented in Sec. 3. The main results for the discounted reward setting are presented in Sec. 4. The main results for the finite-horizon setting are presented in Sec. 5. The extension of our results to the case with stochastic reward is presented in Sec. 6. Our concluding remarks are presented in Sec. 7. Moreover, App. A presents a background discussion on Markov chain theory. App. B presents a background discussion on concentration of martingale sequences. Proofs of main results are presented in the remaining appendices: App. C for the average reward MDPs, App. D for the discounted reward MDPs, and App. E for the finite-horizon MDPs.

1.3 Notation

The symbols $\mathbb{R}$ and $\mathbb{N}$ denote the sets of real and natural numbers and $\mathbb{R}_+$ denotes the set of positive real numbers. The notation $\lim_{\gamma\uparrow 1}$ means the limit as γ approaches 1 from below. Given a sequence of positive numbers $\{a_t\}_{t\geq 0}$ and a function $f\colon \mathbb{N} \to \mathbb{R}$, the notation $a_T = \mathcal{O}(f(T))$ means that $\limsup_{T\to\infty} a_T/f(T) < \infty$ and $a_T = \tilde{\mathcal{O}}(f(T))$ means there exists a finite constant α such that $a_T = \mathcal{O}(\log(T)^\alpha f(T))$.

Given a finite set $\mathcal{S}$, $|\mathcal{S}|$ denotes its cardinality and $\Delta(\mathcal{S})$ denotes the space of probability measures defined on $\mathcal{S}$. For a function $V\colon \mathcal{S} \to \mathbb{R}$, the span of the function $\mathrm{sp}(V)$ is defined as

$$\mathrm{sp}(V) \coloneqq \max_{s\in\mathcal{S}} V(s) - \min_{s\in\mathcal{S}} V(s).$$

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, the notation $\mathbb{E}$ denotes the expectation operator. Given a sequence of random variables $\{S_t\}_{t\geq 0}$, $S_{0:t}$ is a short hand for $(S_0, \dots, S_t)$ and $\sigma(S_{0:t})$ is the sigma-field generated by random variables $S_{0:t}$. The notation $S \sim \rho$ denotes that the random variable S is sampled from the distribution ρ. The standard Gaussian distribution is denoted by $\mathcal{N}(0,1)$. Convergence in distribution is denoted by $\xrightarrow{(d)}$, almost sure convergence is denoted by $\xrightarrow{(a.s.)}$, and convergence in probability is denoted by $\xrightarrow{(p)}$. The phrase almost surely is abbreviated as $a.s.$ and the phrase infinitely often is abbreviated as $i.o.$ The phrases right hand side and left hand side are abbreviated as RHS and LHS, respectively.

2. Problem Formulation

2.1 System Model

Consider a Markov Decision Process (MDP) with state space $\mathcal{S}$ and action space $\mathcal{A}$. We assume that $\mathcal{S}$ and $\mathcal{A}$ are finite sets and use $S_t \in \mathcal{S}$ and $A_t \in \mathcal{A}$ to denote the state and action at time t. At time $t = 0$, the system starts at an initial state S_0, which is a random variable with probability mass function ρ. The state evolves in a controlled Markov manner with transition matrix P, i.e., for any realizations $s_{0:t+1}$ of $S_{0:t+1}$ and $a_{0:t}$ of $A_{0:t}$, we have:

$$\mathbb{P}(S_{t+1} = s_{t+1} | S_{0:t} = s_{0:t}, A_{0:t} = a_{0:t}) = P(s_{t+1} | s_t, a_t).$$

In the sequel, we will use the notation $\mathbb{E}[f(S_+)|s,a]$ to denote the expectation with respect to P, i.e.,

$$\mathbb{E}\big[f(S_+)|s,a\big] = \sum_{s_+\in\mathcal{S}} f(s_+) P(s_+|s,a).$$

At each time t, an agent observes the state of the system S_t and chooses the control action as $A_t \sim \pi_t(S_{0:t}, A_{0:t-1})$, where $\pi_t : \mathcal{S}^t \times \mathcal{A}^{t-1} \to \Delta(\mathcal{A})$ is the *decision rule* at time t. The collection $\pi = (\pi_0, \pi_1, \dots)$ is called a *policy*. We use Π to denote the set of all (history dependent and time varying) policies.

At each time t, the system yields a per-step reward $r(S_t, A_t)$, where $r : \mathcal{S}\times\mathcal{A} \to [0, R_{\max}]$. Let R_T^π denote the total reward received by policy π until time T, i.e.

$$R_T^\pi = \sum_{t=0}^{T-1} r(S_t, A_t), \quad \text{where } A_t \sim \pi(S_{0:t}, A_{0:t-1}).$$

Note that R_T^π is a random variable and we sometimes use the notation $R_T^\pi(\omega)$, $\omega \in \Omega$, to indicate its dependence on the sample path. The long-run expected average reward of a policy $\pi \in \Pi$ starting at the state $s \in \mathcal{S}$ is defined as

$$J^\pi(s) = \liminf_{T\to\infty} \frac{1}{T} \mathbb{E}^\pi\big[R_T^\pi | S_0 = s\big], \quad \forall s \in \mathcal{S},$$

where $\mathbb{E}^\pi$ is the expectation with respect to the joint distribution of all the system variables induced by π. The optimal performance J^* starting at state $s \in \mathcal{S}$ is defined as

$$J^*(s) = \sup_{\pi \in \Pi} J^\pi(s), \quad \forall s \in \mathcal{S}.$$

A policy π^* is called *optimal* if

$$J^{\pi^*}(s) = J^*(s), \quad \forall s \in \mathcal{S}.$$

2.2 The Average Reward Planning Setup

Suppose the system model $\mathcal{M} = (P, r)$ is known.

Definition 1 *Given a model $\mathcal{M} = (P, r)$, define $\Pi_{\text{SD}} \subseteq \Pi$ to be the set of all stationary deterministic Markov policies, i.e., for any $\pi = (\pi_0, \pi_1, \ldots) \in \Pi_{\text{SD}}$, we have $\pi_t : \mathcal{S} \to \mathcal{A}$ (i.e., $A_t = \pi_t(S_t)$), and π_t is the same for all t.*

With a slight abuse of notation, given a decision rule $\pi : \mathcal{S} \to \mathcal{A}$, we will denote the stationary policy $(\pi, \pi, \pi, \ldots)$ by π and interpret R_T^π and J^π as $R_T^{(\pi,\pi,\ldots)}$ and $J^{(\pi,\pi,\ldots)}$, respectively. A stationary policy $\pi \in \Pi_{\text{SD}}$ induces a time-homogeneous Markov chain on $\mathcal{S}$ with transition probability matrix

$$P^\pi(s_{t+1}|s_t) \coloneqq P(s_{t+1}|s_t, \pi(s_t)), \quad \forall s_t, s_{t+1} \in \mathcal{S}.$$

Definition 2 (AROE Solvability) *A model $\mathcal{M} = (P, r)$ is said to be* AROE *(Average Reward Optimality Equation) solvable if there exists a unique optimal long-term average reward $\lambda^* \in \mathbb{R}$ and an optimal differential value function $V^* \colon \mathcal{S} \to \mathbb{R}$ that is unique up to an additive constant that satisfy:*

$$\lambda^* + V^*(s) = \max_{a \in \mathcal{A}} \Big[r(s,a) + \mathbb{E}\big[V^*(S_+)\big|s,a\big] \Big], \quad \forall s \in \mathcal{S}. \tag{AROE}$$

Definition 3 *Given a model $\mathcal{M} = (P, r)$, a policy $\pi \in \Pi_{\text{SD}}$ is said to satisfy* ARPE *(Average Reward Policy Evaluation equation) if there exists a unique long-term average reward $\lambda^\pi \in \mathbb{R}$ and a differential value function $V^\pi \colon \mathcal{S} \to \mathbb{R}$ that is unique up to an additive constant that satisfy:*

$$\lambda^\pi + V^\pi(s) = r(s, \pi(s)) + \mathbb{E}\big[V^\pi(S_+)|s, \pi(s)\big], \quad \forall s \in \mathcal{S}. \tag{ARPE}$$

Definition 4 *Given a model $\mathcal{M} = (P, r)$, define $\Pi_{\text{AR}} \subseteq \Pi_{\text{SD}}$ to be the set of all stationary deterministic policies which satisfy* (ARPE)*.*

The next two propositions follow from standard results in MDP theory.

Proposition 5 (Bertsekas (2012a, Prop. 5.2.1.)) *Suppose model $\mathcal{M} = (P, r)$ is* AROE *solvable with a solution (λ^*, V^*). Then:*

1. *For all $s \in \mathcal{S}$, $J^*(s) = \lambda^*$.*

2. *Let $\pi^* \in \Pi_{\mathrm{SD}}$ be any policy such that $\pi^*(s)$ is an argmax of the RHS of* (AROE)*. Then π^* is optimal, i.e., for all $s \in \mathcal{S}$, $J^{\pi^*}(s) = J^*(s) = \lambda^*$.*

3. *The policy π^* in item 2 belongs to Π_{AR}. In particular, it satisfies* (ARPE) *with a solution (λ^*, V^*).*

Proposition 6 (Bertsekas (2012a, Prop. 5.2.2)) *For any policy $\pi \in \Pi_{\mathrm{AR}}$, we have $J^\pi(s) = \lambda^\pi$, for all $s \in \mathcal{S}$.*

We assume that model $\mathcal{M}$ satisfies the following assumption.

Assumption 1 *The model $\mathcal{M} = (P, r)$ is AROE solvable. Hence, there exists an optimal policy $\pi^* \in \Pi_{\mathrm{AR}}$.*

Proposition 5 implies that under Assumption 1, $J^*(s)$ is constant. In the rest of this section we assume that Assumption 1 always holds and denote $J^*(s)$ by J^*.

2.3 Classification of MDPs

We present the main results of this paper for the policy class Π_{AR} under Assumption 1. However, by imposing further assumptions on $\mathcal{M}$, we can provide a finer characterization of the set Π_{AR} and provide sufficient conditions to guarantee Assumption 1. We recall definitions of different classes of MDPs. Depending on the properties of states following the policies in Π_{SD}, we can classify MDPs to various classes.

Definition 7 (Kallenberg (2002)) *We say that $\mathcal{M}$ is*

1. ***Recurrent (or ergodic)** if for* every *policy $\pi \in \Pi_{\mathrm{SD}}$, the transition matrix P^π consists of a single recurrent class.*

2. ***Unichain** if for* every *policy $\pi \in \Pi_{\mathrm{SD}}$, the transition matrix P^π is unichain, i.e., it consists of a single recurrent class plus a possibly empty set of transient states.*

3. ***Communicating** if, for every pair of states $s, s' \in \mathcal{S}$, there exists a policy $\pi \in \Pi_{\mathrm{SD}}$ under which s' is accessible from s.*

4. ***Weakly Communicating** if there exists a closed set of states $\mathcal{S}_c$ such that (i) for every two states $s, s' \in \mathcal{S}_c$, there exists a policy $\pi \in \Pi_{\mathrm{SD}}$ under which s' is accessible from s; (ii) all states in $\mathcal{S} \setminus \mathcal{S}_c$ are transient under* every *policy.*

See App. A for the details related to the definitions of Markov chains. The following proposition shows the connections between the MDP classes defined above.

Proposition 8 (Puterman (2014, Figure 8.3.1.)) *The following statements hold:*

1. *If $\mathcal{M}$ is recurrent then it is also unichain.*

2. *If $\mathcal{M}$ is unichain then it is also weakly communicating.*

3. *If $\mathcal{M}$ is communicating then it is also weakly communicating.*

By definition, we know that $\Pi_{\text{AR}} \subseteq \Pi_{\text{SD}}$. However, providing a finer characterization of the set Π_{AR} requires further assumptions on the model $\mathcal{M}$. The following proposition presents a sufficient condition for $\mathcal{M}$ under which $\Pi_{\text{AR}} = \Pi_{\text{SD}}$, as well as conditions guaranteeing that Π_{AR} is non-empty, showing the existence of an optimal policy $\pi^* \in \Pi_{\text{AR}}$.

Proposition 9 (Puterman (2014, Table 8.3.1.)) *The following properties hold:*

1. *If $\mathcal{M}$ is recurrent or unichain, then $\Pi_{\text{SD}} = \Pi_{\text{AR}}$.*
2. *If $\mathcal{M}$ is recurrent, unichain, communicating, or weakly communicating, then there exists an optimal policy $\pi^* \in \Pi_{\text{AR}}$. Hence Π_{AR} is non-empty.*

2.4 The Average Reward Learning Setup

We now consider the case where the system model $\mathcal{M} = (P, r)$ is not known. In this case, an agent must use a history dependent policy belonging to Π to *learn* how to act. To differentiate from the planning setting, we denote such a policy by μ and refer to it as a *learning policy*. The quality of a learning policy $\mu \in \Pi$ is quantified by the regret with respect to the optimal policy π^*. There are two notions of regret in the literature, which we state below.

1. **Interim cumulative regret**[1] **of policy μ at time T**, denoted by $\bar{\mathcal{R}}_T^\mu(\omega)$, is the difference between the *average* cumulative reward (i.e., TJ^*) and the cumulative reward of the learning policy, i.e.,
$$\bar{\mathcal{R}}_T^\mu(\omega) \coloneqq TJ^* - R_T^\mu(\omega). \tag{1}$$
2. **Cumulative regret of policy μ at time T**, denoted by $\mathcal{R}_T^\mu(\omega)$, is the difference between the cumulative reward of the optimal policy and the cumulative reward of the learning policy along the *same sample trajectory*, i.e.,
$$\mathcal{R}_T^\mu(\omega) \coloneqq R_T^{\pi^*}(\omega) - R_T^\mu(\omega). \tag{2}$$

Cumulative regret compares the sample path performance of the learning policy with the sample path performance of the optimal policy *on the same sample path*, while the interim cumulative regret compares the sample path performance of the learning policy with the *average* performance of the optimal policy.

In this paper, we characterize probabilistic upper-bounds on the difference between the regret and the interim regret and establish that up to $\tilde{\mathcal{O}}(\sqrt{T})$, these two definitions are rate-equivalent under suitable assumptions.

Let $\mathcal{D}_T^\mu(\omega)$ denote the difference between the cumulative regret and the interim cumulative regret, i.e., $\mathcal{D}_T^\mu(\omega) \coloneqq \mathcal{R}_T^\mu(\omega) - \bar{\mathcal{R}}_T^\mu(\omega)$. It follows from (1)–(2) that
$$\mathcal{D}_T^\mu(\omega) = \mathcal{R}_T^{\pi^*}(\omega) - TJ^*, \tag{3}$$
which implies that $\mathcal{D}_T^\mu(\omega)$ is not a function of the learning policy μ and it only depends on the cumulative reward received by the optimal policy. Therefore, we drop the dependence

1. In the stochastic bandit literature, this definition is sometimes being refereed to as the pseudo regret

on μ in our notation and denote the difference between the cumulative regret and the interim cumulative regret by $\mathcal{D}_T(\omega)$. In this paper, we characterize asymptotic and non-asymptotic guarantees for the random sequence $\{\mathcal{D}_T(\omega)\}_{T\geq 1}$.

Remark 10 *Let $\Pi^* \subset \Pi_{AR}$ denote the set of all optimal policies that satisfy AROE. Assumption 1 implies that $\Pi^* \neq \emptyset$ but in general, $|\Pi^*|$ may be greater than 1. If that is the case, our results are applicable to all optimal policies in Π^*.*

3. Main Results for the Average Reward Setup

We first define statistical properties of the differential value function which is induced by any policy $\pi \in \Pi_{\mathrm{AR}}$.

3.1 Statistical Definitions

For any policy $\pi \in \Pi_{\mathrm{AR}}$, define the following properties of the value function V^π.

1. Span H^π, which is given by
$$H^\pi \coloneqq \mathrm{sp}(V^\pi) = \max_{s\in\mathcal{S}} V^\pi(s) - \min_{s\in\mathcal{S}} V^\pi(s). \tag{4}$$
2. Conditional standard deviation $\sigma^\pi(s)$, which is given by
$$\sigma^\pi(s) \coloneqq \Big[\mathbb{E}\big[\big(V^\pi(S_+) - \mathbb{E}\big[V^\pi(S_+)|s,\pi(s)\big]\big)^2\big|s,\pi(s)\big]\Big]^{1/2}.$$
3. Maximum absolute deviation K^π, which is given by
$$K^\pi \coloneqq \max_{s,s_+\in\mathcal{S}}\Big|V^\pi(s_+) - \mathbb{E}\big[V^\pi(S_+)\big|s,\pi(s)\big]\Big|. \tag{5}$$

For any optimal policy $\pi^* \in \Pi_{\mathrm{AR}}$, we denote the corresponding quantities by H^*, $\sigma^*(s)$, and K^*.

Remark 11 *As mentioned earlier, the solution of* (ARPE) *is unique only up to an additive constant. Adding a constant to V^π does not change the values of H^π, K^π, and σ^π. Therefore it does not matter which specific solution of* (ARPE) *is used to compute H^π, K^π, and σ^π.*

Definition 12 (Bartlett and Tewari (2009)) *Let the expected number of steps to transition from state s to state s' under a policy $\pi \in \Pi_{\mathrm{SD}}$ be denoted by $T^\pi(s,s')$. For any policy $\pi \in \Pi_{\mathrm{SD}}$, the diameter of the policy D^π is defined as*
$$D^\pi \coloneqq \max_{s,s'\in\mathcal{S}} T^\pi(s,s').$$
For the model $\mathcal{M}$, the diameter D and the worst case diameter D_w are defined as
$$D \coloneqq \max_{\substack{s,s'\in\mathcal{S}\\ s\neq s'}} \min_{\pi\in\Pi_{\mathrm{SD}}} T^\pi(s,s'), \tag{6}$$
$$D_w \coloneqq \max_{\pi\in\Pi_{\mathrm{SD}}} \max_{\substack{s,s'\in\mathcal{S}\\ s\neq s'}} T^\pi(s,s'). \tag{7}$$

Lemma 13 *Following relationships hold between the quantities H^π, K^π, and σ^π:*

1. *For any policy $\pi \in \Pi_{\mathrm{AR}}$, we have*
$$\sigma^\pi(s) \le K^\pi \le H^\pi < \infty, \quad \forall s \in \mathcal{S}. \tag{8}$$
2. *If $\mathcal{M}$ is communicating, then for any policy $\pi \in \Pi_{\mathrm{AR}}$, we have $H^\pi \le D^\pi R_{\max}$. Therefore,*
$$\sigma^\pi(s) \le K^\pi \le H^\pi \le D^\pi R_{\max} \le D_w R_{\max}, \quad \forall s \in \mathcal{S}. \tag{9}$$
3. *If $\mathcal{M}$ is weakly communicating, then for any optimal policy $\pi^* \in \Pi_{\mathrm{AR}}$, we have $H^* \le DR_{\max}$. Therefore,*
$$\sigma^*(s) \le K^* \le H^* \le DR_{\max}, \quad \forall s \in \mathcal{S}. \tag{10}$$

The proof is presented in App. C.1.3.

This section presents three families of results. In Sec. 3.2, we present a set of sample path properties for $R_T^\pi(\omega)$ for any policy $\pi \in \Pi_{\mathrm{AR}}$, depicting both asymptotic and non-asymptotic concentration of $R_T^\pi(\omega)$ around its ergodic mean. In Sec. 3.3, we apply these concentration results to characterize the sample path behavior of the difference between any two policies belonging to Π_{AR}, while in Sec. 3.4, we apply these results to the optimal policy π^* to derive the properties of the difference between the cumulative regret and the interim cumulative regret $\mathcal{D}_T(\omega)$.

3.2 Sample Path Characteristics Of Any Policy

In this section, we derive asymptotic and non-asymptotic sample path properties of $R_T^\pi(\omega)$ for any policy $\pi \in \Pi_{\mathrm{AR}}$. The following theorem characterizes the asymptotic concentration rates of $R_T^\pi(\omega)$, establishing LLN, CLT and LIL.

Definition 14 *Let $\{\Sigma_t^\pi\}_{t\ge 0}$ denote the random process defined as*
$$\Sigma_0^\pi = 0, \quad \Sigma_t^\pi = \sum_{\tau=0}^{t-1} \sigma^\pi(S_\tau)^2.$$
Corresponding to this process, define the set Ω_0^π as
$$\Omega_0^\pi \coloneqq \Big\{\omega \in \Omega \,:\, \lim_{t\to\infty} \Sigma_t^\pi(\omega) = \infty\Big\}.$$

Theorem 15 *For any policy $\pi \in \Pi_{\mathrm{AR}}$ and any initial state $s_0 \in \mathcal{S}$, we have following asymptotic characteristics:*

1. *(Law of Large Numbers) The empirical average of the cumulative reward converges almost surely to J^π, i.e.,*
$$\lim_{T\to\infty} \frac{R_T^\pi(\omega)}{T} = J^\pi, \quad a.s. \tag{11}$$

2. *(Central Limit Theorem) Assume that* $\mathbb{P}(\Omega_0^\pi) = 1$. *Let the stopping time* ν_t *be defined as* $\nu_t \coloneqq \min\Big\{T \geq 1 \,:\, \Sigma_T^\pi \geq t\Big\}$. *Then*

$$\lim_{T\to\infty} \frac{R^\pi_{\nu_T}(\omega) - \nu_T J^\pi}{\sqrt{T}} \xrightarrow{(d)} \mathcal{N}(0,1). \tag{12}$$

3. *(Law of Iterated Logarithm) For almost all* $\omega \in \Omega_0^\pi$, *we have*

$$\liminf_{T\to\infty} \frac{R_T^\pi(\omega) - TJ^\pi}{\sqrt{2\Sigma_T^\pi \log\log \Sigma_T^\pi}} = -1, \quad \limsup_{T\to\infty} \frac{R_T^\pi(\omega) - TJ^\pi}{\sqrt{2\Sigma_T^\pi \log\log \Sigma_T^\pi}} = 1. \tag{13}$$

The proof is presented in App. C.2.

Corollary 16 *For any optimal policy* $\pi^* \in \Pi^*$, *the cumulative reward* $R_T^{\pi^*}(\omega)$ *satisfies the asymptotic concentration rates in* (11)–(13), *where in the LHS,* J^π *is replaced with* J^*.

Proof Since π^* is in Π_{AR}, by Theorem 15, the optimal policy should satisfy the asymptotic concentration rates in (11)–(13). ■

The proof of Theorem 15 relies on the finiteness of K^π. However, due to the asymptotic nature of this result, the exact sample complexity dependence of these bounds on properties of the differential value function V^π is not evident. The following theorem establishes the concentration of cumulative reward around the quantity $TJ^\pi - \big(V^\pi(S_T) - V^\pi(S_0)\big)$.

Theorem 17 *For any policy* $\pi \in \Pi_{\mathrm{AR}}$, *the following upper-bounds hold:*

1. *For any* $\delta \in (0,1)$, *with probability at least* $1-\delta$, *we have*

$$\big|R_T^\pi - TJ^\pi - \big(V^\pi(S_0) - V^\pi(S_T)\big)\big| \leq K^\pi \sqrt{2T \log\frac{2}{\delta}}. \tag{14}$$

2. *For any* $\delta \in (0,1)$, *for all* $T \geq T_0^\pi(\delta) \coloneqq \Big\lceil \frac{173}{K^\pi} \log\frac{4}{\delta} \Big\rceil$, *with probability at least* $1-\delta$, *we have*

$$\big|R_T^\pi - TJ^\pi - \big(V^\pi(S_0) - V^\pi(S_T)\big)\big| \leq \max\Big\{K^\pi \sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{2}{\delta}\Big)}, (K^\pi)^2\Big\}. \tag{15}$$

The proof is presented in App. C.3.

Theorem 17 establishes a sample path dependent concentration result. The following theorem establishes a sample path independent finite-time concentration of $R_T^\pi(\omega)$ as a function of the statistical properties of V^π.

Theorem 18 *For any policy* $\pi \in \Pi_{\mathrm{AR}}$, *following upper-bounds hold:*

1. *For any* $\delta \in (0,1)$, *with probability at least* $1-\delta$, *we have*

$$\big|R_T^\pi - TJ^\pi\big| \leq K^\pi \sqrt{2T \log\frac{2}{\delta}} + H^\pi. \tag{16}$$

2. *For any* $\delta \in (0,1)$, *for all* $T \geq T_0^\pi(\delta) \coloneqq \left\lceil \frac{173}{K^\pi} \log \frac{4}{\delta} \right\rceil$, *with probability at least* $1-\delta$, *we have*

$$\left| R_T^\pi - TJ^\pi \right| \leq \max \left\{ K^\pi \sqrt{3T \left(2 \log \log \frac{3T}{2} + \log \frac{2}{\delta} \right)}, (K^\pi)^2 \right\} + H^\pi. \tag{17}$$

The proof is presented in App. C.4.

Corollary 19 *For any optimal policy* $\pi^* \in \Pi^*$, *the cumulative reward* $R_T^{\pi^*}(\omega)$ *satisfies the non-asymptotic concentration rates in* (16)–(17), *where in the LHS,* J^π *is replaced with* J^* *and in the statement and RHS,* (K^π, H^π) *are replaced with* (K^*, H^*).

Proof Since π^* is in Π_{AR}, by Theorem 18, the optimal policy should satisfy the non-asymptotic concentration rates in (16)–(17). ■

Corollary 20 *If* $\mathcal{M}$ *is unichain or recurrent, then any policy* $\pi \in \Pi_{\mathrm{SD}}$ *satisfies asymptotic concentration rates in* (11)–(13) *and non-asymptotic concentration rates in* (16)–(17).

Proof By Prop. 9, for the unichain or recurrent model $\mathcal{M}$, we have $\Pi_{\mathrm{AR}} = \Pi_{\mathrm{SD}}$. As a result, any policy π which belongs to Π_{SD} also belongs to Π_{AR}. Therefore, by Theorem 15, the asymptotic concentration rates in (11)–(13) hold for the policy π and by Theorem 18, the non-asymptotic rates in (16)–(17) hold for the policy π. ■

Corollary 21 *If* $\mathcal{M}$ *is recurrent, unichain, communicating, or weakly communicating, then every optimal policy* $\pi^* \in \Pi^*$ *satisfies asymptotic concentration rates in* (11)–(13) *and non-asymptotic concentration rates in* (16)–(17). *(Prop. 9 shows that there exists at least one such policy.)*

Proof By Prop. 9, for any model $\mathcal{M}$ which is recurrent, unichain, communicating, or weakly communicating, there exists an optimal policy π^* belonging to Π_{AR}. As a result, by Corollary 16, the asymptotic concentration rates in (11)–(13) hold for every optimal policy $\pi^* \in \Pi_{\mathrm{AR}}$. Furthermore, by Corollary 19, the non-asymptotic concentration rates in (16)–(17) hold for every optimal policy $\pi^* \in \Pi_{\mathrm{AR}}$. ■

In Theorem 18, the upper-bounds are established in terms of K^π and H^π. To compute K^π and H^π, one must solve the corresponding (ARPE) equation. At the cost of loosening these bounds, we derive upper-bounds which are in terms of the diameter of the policy D^π and the maximum reward $R_{\max}$. As a result, these upper-bounds only depend on the properties of the Markov chain induced by π and $R_{\max}$.

Corollary 22 *Suppose* $\mathcal{M}$ *is communicating. For any policy* $\pi \in \Pi_{\mathrm{AR}}$, *following upper-bounds hold:*

1. *For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\left|R_T^\pi - TJ^\pi\right| \le D^\pi R_{\max}\sqrt{2T\log\frac{2}{\delta}} + D^\pi R_{\max}. \tag{18}$$

2. *For any $\delta \in (0,1)$, for all $T \ge T_0(\delta) := \left\lceil \frac{173}{D^\pi R_{\max}} \log\frac{4}{\delta} \right\rceil$, with probability at least $1-\delta$, we have*

$$\left|R_T^\pi - TJ^\pi\right| \le \max\left\{D^\pi R_{\max}\sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{2}{\delta}\Big)}, (D^\pi R_{\max})^2\right\} + D^\pi R_{\max}. \tag{19}$$

The proof is presented in App. C.5.

Corollary 23 *If $\mathcal{M}$ is communicating or weakly communicating, then for any optimal policy $\pi^* \in \Pi^*$, the cumulative reward $R_T^{\pi^*}(\omega)$ satisfies the non-asymptotic concentration rates in (18)–(19), where in the LHS, J^π is replaced with J^* and in the RHS, D^π is replaced with D.*

The proof is presented in App. C.6.

In the Corollary 22, the dependence of upper-bounds on the parameters of $\mathcal{M}$ are reflected through $D^\pi R_{\max}$. This implies that if the diameter of the policy D^π or maximum reward $R_{\max}$ increases, these upper-bounds loosen with a linear rate.

Remark 24 *The upper-bounds derived in Corollary 22 depend on the diameter of the policy D^π and are therefore policy-dependent. If $\mathcal{M}$ is communicating or weakly communicating, by Lemma 13, Part 2, we can replace the diameter of the policy D^π with the worst case diameter D_w to get policy-independent upper-bounds. For brevity, we omit this result.*

3.3 Sample Path Behavior of the Performance Difference of Two Stationary Policies

As an implication of the results presented in the Sec. 3.2, we characterize the sample path behavior of the difference in cumulative rewards between any two stationary policies. As a consequence, we derive the non-asymptotic concentration of the difference in rewards between any two optimal policies. These concentration bounds are presented in the following two corollaries.

Corollary 25 *Consider two policies $\pi_1, \pi_2 \in \Pi_{\mathrm{AR}}$. The following upper-bounds hold for the difference between the cumulative reward received by the two policies.*

1. *For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\Big|\left|R_T^{\pi_1} - R_T^{\pi_2}\right| - \left|TJ^{\pi_1} - TJ^{\pi_2}\right|\Big| \le K^{\pi_1}\sqrt{2T\log\frac{4}{\delta}} + H^{\pi_1} + K^{\pi_2}\sqrt{2T\log\frac{4}{\delta}} + H^{\pi_2}. \tag{20}$$

2. *For any* $\delta \in (0,1)$*, for all* $T \geq T_0^{\pi}(\delta) := \max\Big\{\Big\lceil \frac{173}{K^{\pi_1}} \log \frac{8}{\delta} \Big\rceil, \Big\lceil \frac{173}{K^{\pi_2}} \log \frac{8}{\delta} \Big\rceil\Big\}$*, with probability at least* $1-\delta$*, we have*

$$\Big|\big|R_T^{\pi_1} - R_T^{\pi_2}\big| - \big|TJ^{\pi_1} - TJ^{\pi_2}\big|\Big| \leq \max\Big\{K^{\pi_1}\sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{4}{\delta}\Big)}, (K^{\pi_1})^2\Big\} + H^{\pi_1} + \max\Big\{K^{\pi_2}\sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{4}{\delta}\Big)}, (K^{\pi_2})^2\Big\} + H^{\pi_2}. \tag{21}$$

The proof is presented in App. C.7.

Corollary 26 *Consider two optimal policies* $\pi_1^*, \pi_2^* \in \Pi^*$*. Then for the difference between cumulative rewards received by the two optimal policies* $\big|R_T^{\pi_1^*} - R_T^{\pi_2^*}\big|$*, we have*

1. *For any* $\delta \in (0,1)$*, with probability at least* $1-\delta$*, we have*

$$\big|R_T^{\pi_1^*} - R_T^{\pi_2^*}\big| \leq 2\Big(K^*\sqrt{2T\log\frac{4}{\delta}} + H^*\Big). \tag{22}$$

2. *For any* $\delta \in (0,1)$*, for all* $T \geq T_0^{\pi^*}(\delta) := \Big\lceil \frac{173}{K^*} \log \frac{8}{\delta} \Big\rceil$*, with probability at least* $1-\delta$*, we have*

$$\big|R_T^{\pi_1^*} - R_T^{\pi_2^*}\big| \leq 2\Big(\max\Big\{K^*\sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{4}{\delta}\Big)}, (K^*)^2\Big\} + H^*\Big). \tag{23}$$

Proof Since both policies $\pi_1^*, \pi_2^* \in \Pi_{\mathrm{AR}}$ are optimal policies, by the definition, we have $J^{\pi_1^*} = J^{\pi_2^*} = J^*$ and therefore, $T\big|J^{\pi_1^*} - J^{\pi_2^*}\big| = 0$. As a result, by Corollary 25, the difference $\big|R_T^{\pi_1^*} - R_T^{\pi_2^*}\big|$ satisfies the non-asymptotic concentration rates in Corollary 25 with the RHS of (20)–(21) being simplified to RHS of (22)–(23). ■

Remark 27 *Similar to the Corollary 22, by imposing the assumption that* $\mathcal{M}$ *is communicating or weakly communicating, we can derive the counterpart of* (20)–(21) *and* (22)–(23) *in terms of* $D^{\pi} R_{\max}$ *respectively. For brevity, we omit this result.*

3.4 Implication for Learning

In this section, we present the consequences of our results on the regret of learning algorithms. We characterize the asymptotic and non-asymptotic sample path behavior of the difference between cumulative regret and interim cumulative regret. Recall that for any learning policy μ, this difference is defined as $\mathcal{D}_T(\omega) = \bar{\mathcal{R}}_T^{\mu}(\omega) - \mathcal{R}_T^{\mu}(\omega)$. Similar to Theorem 15, we characterize the asymptotic concentration rates of $\{\mathcal{D}_T(\omega)\}_{T\geq 1}$, establishing LLN, CLT and LIL.

Definition 28 *Let $\{\Sigma_t^*\}_{t\geq 0}$ denote the random process defined as*

$$\Sigma_0^* = 0, \quad \Sigma_t^* = \sum_{\tau=0}^{t-1} \sigma^*(S_\tau)^2.$$

Corresponding to this process, we define the set Ω_0^* as

$$\Omega_0^* \coloneqq \Big\{\omega \in \Omega \,:\, \lim_{t\to\infty} \Sigma_t^*(\omega) = \infty\Big\}.$$

Theorem 29 *For any learning policy μ, the difference $\mathcal{D}_T(\omega)$ of cumulative regret and interim cumulative regret satisfies following properties.*

1. *(Law of Large Numbers) The difference almost surely grows sub-linearly, i.e.*

$$\lim_{T\to\infty} \frac{\mathcal{D}_T(\omega)}{T} = 0, \quad a.s.$$

2. *(Central Limit Theorem) Assume that $\mathbb{P}(\Omega_0^*) = 1$. Let stopping time ν_t be defined as $\nu_t \coloneqq \min\Big\{T \geq 1 \,:\, \Sigma_T^* \geq t\Big\}$. Then*

$$\lim_{T\to\infty} \frac{\mathcal{D}_{\nu_T}(\omega)}{\sqrt{T}} \xrightarrow{(d)} \mathcal{N}(0,1).$$

3. *(Law of Iterated Logarithm) For almost all $\omega \in \Omega_0^*$, we have*

$$\liminf_{T\to\infty} \frac{\mathcal{D}_T(\omega)}{\sqrt{2\Sigma_T^* \log\log \Sigma_T^*}} = -1, \quad \limsup_{T\to\infty} \frac{\mathcal{D}_T(\omega)}{\sqrt{2\Sigma_T^* \log\log \Sigma_T^*}} = 1. \tag{24}$$

Proof is presented in App. C.8.

In addition to the asymptotic results presented in Theorem 29, we present non-asymptotic guarantees for the sequence $\{\mathcal{D}_T(\omega)\}_{T\geq 1}$. Similar to Theorem 18, we characterize the non-asymptotic concentration of $\mathcal{D}_T(\omega)$ as a function of statistical properties of V^* (i.e., K^* and H^*).

Theorem 30 *The difference of cumulative regret and interim cumulative regret $\mathcal{D}_T(\omega)$ satisfies:*

1. *For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\big|\mathcal{D}_T(\omega)\big| \leq K^* \sqrt{2T \log\frac{2}{\delta}} + H^*.$$

2. *For any $\delta \in (0,1)$, for all $T \geq T_0^*(\delta) \coloneqq \Big\lceil \frac{173}{K^*} \log\frac{4}{\delta} \Big\rceil$, with probability at least $1-\delta$, we have*

$$\big|\mathcal{D}_T(\omega)\big| \leq \max\Big\{K^* \sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{2}{\delta}\Big)}, (K^*)^2\Big\} + H^*.$$

Proof is presented in App. C.9. As mentioned earlier, the difference $\mathcal{D}_T(\omega)$ does not depend on the learning policy μ. Therefore, the results of Theorem 30 do not depend on the choice of the learning policy either.

In Theorem 30, the upper-bounds are established in terms of K^* and H^*. Similar to Corollary 22, we can derive upper-bounds in terms of model parameters D and $R_{\max}$ at the cost of loosening the upper-bounds. These bounds are presented in the following Corollary.

Corollary 31 *Suppose $\mathcal{M}$ is recurrent, unichain, communicating, or weakly communicating, then $\mathcal{D}_T(\omega)$ satisfies following properties.*

1. *For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\left|\mathcal{D}_T(\omega)\right| \leq DR_{\max}\sqrt{2T\log\frac{2}{\delta}} + DR_{\max}.$$

2. *For any $\delta \in (0,1)$, for all $T \geq T_0(\delta) \coloneqq \left\lceil \frac{173}{DR_{\max}} \log\frac{4}{\delta} \right\rceil$, with probability at least $1-\delta$, we have*

$$\left|\mathcal{D}_T(\omega)\right| \leq \max\left\{DR_{\max}\sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{2}{\delta}\Big)}, (DR_{\max})^2\right\} + DR_{\max}.$$

Proof is presented in App. C.10.

Remark 32 *Notice that conditions of Corollary 31 are weaker than the conditions of Corollary 22. As a result, Corollary 31 can be applied to broader classes of $\mathcal{M}$. This difference originates from the difference between items (2) and (3) in Lemma 13.*

In this section, we established probabilistic upper-bounds for the difference between cumulative regret and interim cumulative regret. We showed, asymptotically and non-asymptotically, the growth rate of this difference is upper-bounded by $\tilde{\mathcal{O}}(\sqrt{T})$. This implies that if we establish a regret rate of $\tilde{\mathcal{O}}(\sqrt{T})$ for a learning algorithm μ using either of the definitions, similar regret rate hold for the algorithm μ using the other definition. This result is presented in the following theorem.

Theorem 33 *For any learning policy μ we have:*

1. *The following statements are equivalent.*
   - *(a) $R_T^\mu(\omega) \leq \tilde{\mathcal{O}}(\sqrt{T})$, a.s.*
   - *(b) $\bar{R}_T^\mu(\omega) \leq \tilde{\mathcal{O}}(\sqrt{T})$, a.s.*
2. *The following statements are true.*
   - *(a) Suppose for a learning algorithm μ and any $\delta \in (0,1)$, there exists a $T_0(\delta)$ such that for all $T \geq T_0(\delta)$, with probability at least $1-\delta$, we have $R_T^\mu(\omega) \leq \tilde{\mathcal{O}}(\sqrt{T})$, where $\tilde{\mathcal{O}}(\cdot)$ notation functionally depends upon constants related to $\mathcal{M}$ and δ. Then for any $\delta \in (0,1)$, there exists $T_1(\delta)$ such that for all $T \geq T_1(\delta)$, with probability at least $1-\delta$, we have $\bar{R}_T^\mu(\omega) \leq \tilde{\mathcal{O}}(\sqrt{T})$.*

(b) Suppose for a learning algorithm μ and any $\delta \in (0,1)$, there exists a $T_0(\delta)$ such that for all $T \geq T_0(\delta)$, with probability at least $1-\delta$, we have $\bar{R}^{\mu}_T(\omega) \leq \tilde{\mathcal{O}}(\sqrt{T})$, where $\tilde{\mathcal{O}}(\cdot)$ notation functionally depends upon constants related to $\mathcal{M}$ and δ. Then for any $\delta \in (0,1)$, there exists $T_1(\delta)$ such that for all $T \geq T_1(\delta)$, with probability at least $1-\delta$, we have $R^{\mu}_T(\omega) \leq \tilde{\mathcal{O}}(\sqrt{T})$.

Proof is presented in App. C.11.

4. Main Results for the Discounted Reward Setup

In this section, we extend the non-asymptotic concentration results that we established for the average reward setup to the discounted reward setup.

4.1 System Model

Consider a discounted reward MDP with state space $\mathcal{S}$ and action space $\mathcal{A}$. Similar to Sec. 2, we assume that $\mathcal{S}$ and $\mathcal{A}$ are finite sets. The state evolves in a controlled Markov manner with transition matrix P and at each time t, the system yields a per-step reward $r(S_t, A_t) \in [0, R_{\max}]$. Let $\gamma \in (0,1)$ denote the discount factor of the model. The definitions of policies and policy sets Π and Π_{SD} are similar to Sec. 2. The discounted cumulative reward received by any policy π is given by

$$R^{\pi,\gamma}_T(\omega) \coloneqq \sum_{t=0}^{T-1} \gamma^t r(S_t, A_t), \quad \text{where, } A_t = \pi(S_{0:t}, A_{0:t-1}), \quad \omega \in \Omega.$$

Note that $R^{\pi,\gamma}_T(\omega)$ is a random variable. For this model, the long-run expected discounted reward of policy $\pi \in \Pi_{\mathrm{SD}}$ starting at the state $s \in \mathcal{S}$ is defined as

$$V^{\pi}_{\gamma}(s) \coloneqq \mathbb{E}^{\pi}\Big[\lim_{T\to\infty} R^{\pi,\gamma}_T \;\Big|\; S_0 = s \Big], \quad \forall s \in \mathcal{S},$$

where $\mathbb{E}^{\pi}$ is the expectation with respect to the joint distribution of all the system variables induced by π. We refer to the function V^{π}_{γ} as the discounted value function corresponding to the policy π. The optimal performance V^*_{γ} starting at state $s \in \mathcal{S}$ is defined as

$$V^*_{\gamma}(s) = \sup_{\pi\in\Pi} V^{\pi}_{\gamma}(s), \quad \forall s \in \mathcal{S}.$$

A policy π^* is called optimal if

$$V^{\pi^*}_{\gamma}(s) = V^*_{\gamma}(s), \quad \forall s \in \mathcal{S}.$$

Definition 34 *A discounted model $\mathcal{M}$ is said to satisfy DROE (Discounted Reward Optimality Equation) if there exists an optimal discounted value function $V^*_{\gamma} : \mathcal{S} \to \mathbb{R}$ that satisfies:*

$$V^*_{\gamma}(s) = \max_{a\in\mathcal{A}} \Big[r(s,a) + \gamma \mathbb{E}\big[V^*_{\gamma}(S_+) \;\big|\; s, a \big] \Big], \quad \forall s \in \mathcal{S}. \tag{DROE}$$

Definition 35 *Given a discounted model $\mathcal{M}$, a policy $\pi \in \Pi_{\mathrm{SD}}$ is said to satisfy DRPE (Discounted Reward Policy Evaluation equation) if there exists a discounted value function $V_\gamma^\pi : \mathcal{S} \to \mathbb{R}$ that satisfies:*

$$V_\gamma^\pi(s) = r(s, \pi(s)) + \gamma \mathbb{E}\big[V_\gamma^\pi(S_+) \;\big|\; s, \pi(s)\big], \quad \forall s \in \mathcal{S}. \tag{DRPE}$$

Proposition 36 (Bertsekas (2012a, Prop. 1.2.3–1.2.5)) *For a discounted model $\mathcal{M}$, following statements hold:*

1. *Any policy $\pi \in \Pi_{\mathrm{SD}}$ satisfies* (DRPE).
2. *Let π^* be any policy such that $\pi^*(s)$ is an argmax of the RHS of* (DROE). *Then π^* is optimal, i.e., for all $s \in \mathcal{S}$, $V_\gamma^{\pi^*}(s) = V_\gamma^*(s)$.*
3. *The policy π^* in step 2 belongs to Π_{SD}. In particular, it satisfies* (DRPE) *with a solution V_γ^*.*

4.2 Sample Path Characteristics of Any Policy

For any policy $\pi \in \Pi_{\mathrm{SD}}$, we define following statistical properties of the discounted value function V_γ^π.

1. Span of the discounted value function V_γ^π given by

$$H^{\pi,\gamma} \coloneqq \mathrm{sp}(V_\gamma^\pi) = \max_{s \in \mathcal{S}} V_\gamma^\pi(s) - \min_{s \in \mathcal{S}} V_\gamma^\pi(s). \tag{25}$$

2. Maximum absolute deviation of the discounted value function V_γ^π is given by

$$K^{\pi,\gamma} \coloneqq \max_{s, s_+ \in \mathcal{S}} \Big| V_\gamma^\pi(s_+) - \mathbb{E}\big[V_\gamma^\pi(S_+) \;\big|\; s, \pi(s)\big] \Big|. \tag{26}$$

For any optimal policy $\pi^* \in \Pi_{\mathrm{SD}}$, we denote these corresponding quantities by $H^{*,\gamma}$, and $K^{*,\gamma}$. Similar to the results in Theorem 17 for the average reward setup, we can derive non-asymptotic concentration results for the discounted reward setup. These results are presented in the following theorem. To simplify the notation, let

$$f^\gamma(T) \coloneqq \sum_{t=1}^{T} \gamma^{2t} = \frac{\gamma^2 - \gamma^{2T+2}}{1 - \gamma^2}.$$

An immediate implication of the definitions of $R_T^{\pi,\gamma}$ and $V_\gamma^\pi(s)$ is that

$$\mathbb{E}\Big[R_T^{\pi,\gamma} + \gamma^T V_\gamma^\pi(S_T) - V_\gamma^\pi(S_0)\Big] = 0.$$

In this section, we show that with high-probability $R_T^{\pi,\gamma}$ concentrates around $V_\gamma^\pi(S_0) - \gamma^T V_\gamma^\pi(S_T)$ and characterize the concentration rate.

Theorem 37 *For any policy $\pi \in \Pi_{\mathrm{SD}}$ and any $s \in \mathcal{S}$, we have:*

1. *For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\left| R_T^{\pi,\gamma} - \big(V_\gamma^\pi(S_0) - \gamma^T V_\gamma^\pi(S_T)\big) \right| \le K^{\pi,\gamma} \sqrt{2 f^\gamma(T) \log \frac{2}{\delta}}. \tag{27}$$

2. *For any $\delta \in (0,1)$, if $\lim_{T'\to\infty} f^\gamma(T') > \dfrac{173}{K^{\pi,\gamma}} \log \dfrac{4}{\delta}$, then for all $T \ge T_0(\delta) := \min\Big\{T' \ge 1 : f^\gamma(T') > \dfrac{173}{K^{\pi,\gamma}} \log \dfrac{4}{\delta}\Big\}$, with probability at least $1-\delta$, we have*

$$\begin{aligned} &\left| R_T^{\pi,\gamma} - \big(V_\gamma^\pi(S_0) - \gamma^T V_\gamma^\pi(S_T)\big) \right| \\ &\qquad \le \max\left\{ K^{\pi,\gamma} \sqrt{3 f^\gamma(T) \Big(2 \log\log\big(\frac{3}{2} f^\gamma(T)\big) + \log \frac{2}{\delta}\Big)}, (K^{\pi,\gamma})^2 \right\}. \end{aligned} \tag{28}$$

The proof is presented in App. D.1.

Corollary 38 *For any policy $\pi \in \Pi_{\mathrm{SD}}$ and any $s \in \mathcal{S}$, we have:*

1. *For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\left| R_T^{\pi,\gamma} - V_\gamma^\pi(S_0) \right| \le K^{\pi,\gamma} \sqrt{2 f^\gamma(T) \log \frac{2}{\delta}} + \frac{\gamma^T}{1-\gamma} R_{\max}. \tag{29}$$

2. *For any $\delta \in (0,1)$, if $\lim_{T'\to\infty} f^\gamma(T') > \dfrac{173}{K^{\pi,\gamma}} \log \dfrac{4}{\delta}$, then for all $T \ge T_0(\delta) := \min\Big\{T' \ge 1 : f^\gamma(T') > \dfrac{173}{K^{\pi,\gamma}} \log \dfrac{4}{\delta}\Big\}$, with probability at least $1-\delta$, we have*

$$\begin{aligned} &\left| R_T^{\pi,\gamma} - V_\gamma^\pi(S_0) \right| \\ &\qquad \le \max\left\{ K^{\pi,\gamma} \sqrt{3 f^\gamma(T) \big(2 \log\log(\frac{3}{2} f^\gamma(T)) + \log \frac{2}{\delta}\big)}, (K^{\pi,\gamma})^2 \right\} + \frac{\gamma^T}{1-\gamma} R_{\max}. \end{aligned} \tag{30}$$

The proof is presented in App. D.2.

Corollary 39 *For any optimal policy $\pi^* \in \Pi_{\mathrm{SD}}$, the discounted cumulative reward $R_T^{\pi^*,\gamma}(\omega)$ satisfies the non-asymptotic concentration rates in (27)–(30), where in the LHS, $V_\gamma^\pi(s)$ is replaced with $V_\gamma^*(s)$ and in the statement and RHS, $K^{\pi,\gamma}$ is replaced with $K^{*,\gamma}$.*

Proof Since π^* is in Π_{SD}, by Theorem 37 and Corollary 38, the optimal policy satisfies the non-asymptotic concentration rates in (27)–(30). ■

4.3 Sample Path Behavior of Performance Difference of Two Stationary Policies

As an implication of the results presented in the Sec. 4.2, we characterize the sample path behavior of the difference in discounted cumulative rewards between any two stationary policies. As a consequence, we derive the non-asymptotic concentration of the difference in rewards between any two optimal policies. These concentration bounds are presented in the following two corollaries.

Corollary 40 *Consider two policies $\pi_1, \pi_2 \in \Pi_{\text{SD}}$. Let $\{S_t^{\pi_1}\}_{t\geq 0}$ and $\{S_t^{\pi_2}\}_{t\geq 0}$ denote the random sequences of the states encountered by policy π_1 and π_2 respectively. Following upper-bounds hold for the difference between the discounted cumulative reward received by the two policies.*

1. *For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\begin{aligned}&\Big|\big|R_T^{\pi_1,\gamma} - R_T^{\pi_2,\gamma}\big| - \big|\big[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\big] - \big[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\big]\big|\Big| \\ &\leq K^{\pi_1,\gamma}\sqrt{2f^\gamma(T)\log\frac{4}{\delta}} + K^{\pi_2,\gamma}\sqrt{2f^\gamma(T)\log\frac{4}{\delta}}. \end{aligned} \tag{31}$$

2. *For any $\delta \in (0,1)$, if $\lim_{T'\to\infty} f^\gamma(T') > \frac{173}{K^{\pi,\gamma}}\log\frac{4}{\delta}$, define $T_0^{\pi_i}(\frac{\delta}{2})$ as*

$$T_0^{\pi_i}(\frac{\delta}{2}) \coloneqq \min\Big\{T' \geq 1 : f^\gamma(T') > \frac{173}{K^{\pi_i,\gamma}}\log\frac{8}{\delta}\Big\}, \quad i \in \{1,2\}. \tag{32}$$

Then, for all $T \geq T_0^\pi(\delta) \coloneqq \max\Big\{T_0^{\pi_1}(\frac{\delta}{2}), T_0^{\pi_2}(\frac{\delta}{2})\Big\}$, with probability at least $1-\delta$, we have

$$\begin{aligned}&\Big|\big|R_T^{\pi_1,\gamma} - R_T^{\pi_2,\gamma}\big| - \big|\big[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\big] - \big[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\big]\big|\Big| \\ &\leq \max\Big\{K^{\pi_1,\gamma}\sqrt{3f^\gamma(T)\Big(2\log\log\frac{3}{2}f^\gamma(T) + \log\frac{4}{\delta}\Big)}, (K^{\pi_1,\gamma})^2\Big\} \\ &+ \max\Big\{K^{\pi_2,\gamma}\sqrt{3f^\gamma(T)\Big(2\log\log\frac{3}{2}f^\gamma(T) + \log\frac{4}{\delta}\Big)}, (K^{\pi_2,\gamma})^2\Big\}. \end{aligned} \tag{33}$$

The proof if presented in App. D.3.

Corollary 41 *Consider two optimal policies $\pi_1^*, \pi_2^* \in \Pi_{\text{SD}}$. Let $\{S_t^{\pi_1^*}\}_{t\geq 0}$ and $\{S_t^{\pi_2^*}\}_{t\geq 0}$ denote the random sequences of states encountered by optimal policies π_1^* and π_2^*. To simplify the expression, we assume the system starts at a fixed initial state, i.e., $S_0^{\pi_1^*} = S_0^{\pi_2^*}$. Then for the difference between discounted cumulative rewards received by the two optimal policies $\big|R_T^{\pi_1^*,\gamma} - R_T^{\pi_2^*,\gamma}\big|$, we have:*

1. *For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\Big|\big|R_T^{\pi_1^*,\gamma} - R_T^{\pi_2^*,\gamma}\big| - \gamma^T\big|V_\gamma^*(S_T^{\pi_2^*}) - V_\gamma^*(S_T^{\pi_1^*})\big|\Big| \leq 2\Big(K^{*,\gamma}\sqrt{2f^\gamma(T)\log\frac{4}{\delta}}\Big). \tag{34}$$

2. *Consider* $T_0^{\pi^*}(\frac{\delta}{2})$ *defined in* (32)*. For any* $\delta \in (0,1)$*, for all* $T \geq T_0^{\pi^*}(\frac{\delta}{2})$*, with probability at least* $1-\delta$*, we have*

$$\Big|\big|R_T^{\pi_1^*,\gamma} - R_T^{\pi_2^*,\gamma}\big| - \gamma^T\big|V_\gamma^*(S_T^{\pi_2^*}) - V_\gamma^*(S_T^{\pi_1^*})\big|\Big| \\ \leq 2\Big(\max\Big\{K^{*,\gamma}\sqrt{3f^\gamma(T)\Big(2\log\log\big(\frac{3}{2}f^\gamma(T)\big) + \log\frac{4}{\delta}\Big)}, (K^{*,\gamma})^2\Big\}\Big). \tag{35}$$

Proof Since both policies $\pi_1^*, \pi_2^* \in \Pi_{\mathrm{SD}}$ are optimal policies, by the definition, we have

$$V_\gamma^{\pi_1^*}(s) = V_\gamma^{\pi_2^*}(s) = V_\gamma^*(s), \quad \forall s \in \mathcal{S}, \ \forall \gamma \in (0,1).$$

As a result, by the assumption that $S_0^{\pi_1^*} = S_0^{\pi_2^*}$ we have

$$\Big|V_\gamma^*(S_0^{\pi_1^*}) - V_\gamma^*(S_0^{\pi_2^*})\Big| = 0.$$

In addition, we have

$$K^{\pi_1^*,\gamma} = K^{\pi_2^*,\gamma} = K^{*,\gamma}, \quad \forall \gamma \in (0,1).$$

As a result, by Corollary 40, the difference $\big|R_T^{\pi_1^*,\gamma} - R_T^{\pi_2^*,\gamma}\big|$ satisfies the non-asymptotic concentration rates in Corollary 40 with the RHS of (31) and (33) being simplified to RHS of (34)–(35). ■

4.4 Vanishing Discount Analysis

In order to observe the connection between the upper-bounds established in Theorem 17 and Theorem 37, we investigate the asymptotic behavior of these two upper-bounds as the discount factor γ goes to 1 from below (i.e., $\gamma \uparrow 1$). This characterization is stated in the following Corollary.

Corollary 42 *For any policy* $\pi \in \Pi_{\mathrm{AR}}$*, we have the following asymptotic relations between the bounds in Theorem 17 and Theorem 37.*

1. *As* γ *goes to* 1 *from below, the quantity in the LHS of* (27)–(28) *converges to the LHS of* (14)*, i.e.,*

$$\lim_{\gamma\uparrow 1}\Big|R_T^{\pi,\gamma} - \big(V_\gamma^\pi(S_0) - \gamma^T V_\gamma^\pi(S_T)\big)\Big| = \Big|R_T^\pi - TJ^\pi + \big(V^\pi(S_0) - V^\pi(S_T)\big)\Big|.$$

2. *As* γ *goes to* 1 *from below, the RHS in* (27) *converges to the RHS in* (14)*, i.e.,*

$$\lim_{\gamma\uparrow 1}\Big[K^{\pi,\gamma}\sqrt{2f^\gamma(T)\log\frac{2}{\delta}}\Big] = K^\pi\sqrt{2T\log\frac{2}{\delta}}.$$

3. *As* γ *goes to* 1 *from below, the RHS in* (28) *converges to the RHS in* (15)*, i.e.,*

$$\lim_{\gamma\uparrow 1}\Big[\max\Big\{K^{\pi,\gamma}\sqrt{3f^\gamma(T)\big(2\log\log(\frac{3}{2}f^\gamma(T)) + \log\frac{2}{\delta}\big)}, (K^{\pi,\gamma})^2\Big\}\Big] \\ = \max\Big\{K^\pi\sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{2}{\delta}\Big)}, (K^\pi)^2\Big\}.$$

Proof is presented in App. D.4.

Remark 43 *The non-asymptotic characterizations are established in Theorem 37. Since the discounted cumulative reward $R_T^{\pi,\gamma}$ is finite for $\mathcal{M}$, we cannot provide any asymptotic characterization for this quantity. However, Corollary 42 shows that as the discount factor γ goes to 1 from below, the non-asymptotic concentration behavior of $R_T^{\pi,\gamma}$ resembles the non-asymptotic concentration of R_T^{π}. This gives a complete picture of concentration rate of $R_T^{\pi,\gamma}$ and R_T^{π}.*

5. Main Results for the Finite-Horizon Setup

In this section, we extend the non-asymptotic concentration results that we established for the average reward and discounted reward setups to the case of finite-horizon setup.

5.1 System Model

Consider an MDP with state space $\mathcal{S}$ and action space $\mathcal{A}$. Similar to Sec. 2, we assume that $\mathcal{S}$ and $\mathcal{A}$ are finite sets. The state evolves in a controlled Markov manner with transition matrix P and at each time t, the system yields a per-step reward $r(S_t, A_t) \in [0, R_{\max}]$. Let $h \in \mathbb{R}$ denote the horizon of the problem. The definitions of policy and policy set Π are similar to Sec. 2.

Definition 44 *Given a model $\mathcal{M} = (P, r, h)$, define Π_{FD} to be the set of finite-horizon deterministic policies, i.e., for any $\pi = (\pi_0, \pi_1, \ldots, \pi_h) \in \Pi_{\text{FD}}$, we have $\pi_t : \mathcal{S} \to \mathcal{A}$ (i.e., $A_t = \pi_t(S_t)$), but π_t may depend upon t.*

The cumulative reward received by any policy $\pi \in \Pi$ up to time T (T is not necessarily equal to h) is given by

$$R_T^{\pi,h}(\omega) \coloneqq \sum_{t=0}^{T-1} r(S_t, A_t), \quad \text{where, } A_t = \pi(S_{0:t}, A_{0:t-1}), \quad \omega \in \Omega, \quad T \leq h+1.$$

Note that $R_T^{\pi,h}(\omega)$ is a random variable. For this model, the expected total reward of any policy $\pi \in \Pi$ starting at the state $s \in \mathcal{S}$ is defined as

$$J^{\pi,h}(s) \coloneqq \mathbb{E}^{\pi}\Big[R_{h+1}^{\pi,h} \mid S_0 = s\Big], \quad \forall s \in \mathcal{S},$$

where $\mathbb{E}^{\pi}$ is the expectation with respect to the joint distribution of all the system variables induced by π. The optimal performance $J^{*,h}(s)$ starting at state $s \in \mathcal{S}$ is defined as

$$J^{*,h}(s) = \sup_{\pi \in \Pi} J^{\pi,h}(s), \quad \forall s \in \mathcal{S}.$$

A policy π^* is called optimal if

$$J^{\pi^*,h}(s) = J^{*,h}(s), \quad \forall s \in \mathcal{S}.$$

Definition 45 *The sequence of finite-horizon optimal value functions* $\{V_t^{*,h}\}_{t=0}^{h+1} : \mathcal{S} \to \mathbb{R}$ *is defined as follows*

$$V_{h+1}^{*,h}(s) = 0, \quad \forall s \in \mathcal{S},$$

and for $t \in \{h, h-1, \dots, 0\}$*, recursively define* $V_t^{*,h}(s)$ *based on the FHDP (Finite-Horizon Dynamic Programming equation) given by*

$$V_t^{*,h}(s) = \max_{a \in \mathcal{A}} \Big[r(s,a) + \mathbb{E}\big[V_{t+1}^{*,h}(S_+) \;\big|\; s, a \big] \Big], \quad \forall s \in \mathcal{S}. \tag{FHDP}$$

Definition 46 *Given a policy* $\pi \in \Pi_{\mathrm{FD}}$*, the sequence of finite-horizon value functions* $\{V_t^{\pi,h}\}_{t=0}^{h+1} : \mathcal{S} \to \mathbb{R}$ *corresponding to the policy* π *is defined as follows*

$$V_{h+1}^{\pi,h}(s) = 0, \quad \forall s \in \mathcal{S},$$

and for $t \in \{h, h-1, \dots, 0\}$*, recursively define* $V_t^{\pi,h}(s)$ *based on the FHPE (Finite-Horizon Policy Evaluation equation) given by*

$$V_t^{\pi,h}(s) = r(s, \pi_t(s)) + \mathbb{E}\big[V_{t+1}^{\pi,h}(S_+) \;\big|\; s, \pi_t(s) \big], \quad \forall s \in \mathcal{S}. \tag{FHPE}$$

Proposition 47 (Bertsekas (2012b)) *Let* $\pi^* = (\pi_0^*, \pi_1^*, \dots, \pi_h^*) \in \Pi_{\mathrm{FD}}$ *be a policy such that* $\pi_t^*(s_t)$ *denote the argmax of* (FHDP) *at stage* t*. Then the policy* π^* *is optimal, i.e., for all* $s \in \mathcal{S}$*,* $J^{\pi^*,h}(s) = J^{*,h}(s)$*.*

5.2 Sample Path Characteristics of Any Policy

For any policy $\pi \in \Pi_{\mathrm{FD}}$, we define following statistical properties of the sequence of finite-horizon value functions $\{V_t^{\pi,h}\}_{t=0}^{h+1}$.

1. Span of the finite-horizon value function $V_t^{\pi,h}$ is given by

$$H_t^{\pi,h} \coloneqq \mathrm{sp}(V_t^{\pi,h}), \quad \forall t \in \{0, 1, \dots, h\}. \tag{36}$$

2. Maximum absolute deviation of the finite-horizon value function $V_t^{\pi,h}$ is given by

$$K_t^{\pi,h} \coloneqq \max_{s, s_+} \Big| V_t^{\pi,h}(s_+) - \mathbb{E}\big[V_t^{\pi,h}(S_+) \;\big|\; s, \pi_t(s) \big] \Big|, \quad \forall t \in \{0, 1, \dots, h\}. \tag{37}$$

Similar to the results in Theorem 18 and Theorem 37 for the average reward and discounted reward setups, we derive non-asymptotic concentration results for the finite-horizon setup. These results are presented in the following theorem. To simplify the notation, let

$$\bar{K}_T^{\pi,h} = \max_{0 \le t \le T} K_t^{\pi,h}, \quad \bar{H}_T^{\pi,h} = \max_{0 \le t \le T} H_t^{\pi,h}, \tag{38}$$

and let

$$g^{\pi,h}(T) \coloneqq \frac{\sum_{t=1}^{T} (K_t^{\pi,h})^2}{\big(\bar{K}_T^{\pi,h}\big)^2}. \tag{39}$$

For any optimal policy $\pi^* \in \Pi_{\text{FD}}$, we denote these corresponding quantities by $H_t^{*,h}$, $K_t^{*,h}$, $\bar{H}_T^{*,h}$, $\bar{K}_T^{*,h}$, and $g^{*,h}(T)$. An immediate implication of the definitions of $R_T^{\pi,h}$ and $V_T^{\pi,h}(s)$ is that

$$\mathbb{E}\Big[R_T^{\pi,h} + V_T^{\pi,h}(S_T) - V_0^{\pi,h}(S_0)\Big] = 0.$$

In this section, we show that with high-probability $R_T^{\pi,h}$ concentrates around $V_0^{\pi,h}(S_0) - V_T^{\pi,h}(S_T)$ and characterize the concentration rate. Following theorem is analogous to the concentration bounds in average reward setup given in Theorem 17 and concentration bounds in discounted reward setup given in Theorem 37.

Theorem 48 *For any policy $\pi \in \Pi_{\text{FD}}$, we have:*

1. *For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\Big|R_T^{\pi,h} - \big(V_0^{\pi,h}(S_0) - V_T^{\pi,h}(S_T)\big)\Big| \le \bar{K}_T^{\pi,h}\sqrt{2g^{\pi,h}(T)\log\frac{2}{\delta}}.$$

2. *For any $\delta \in (0,1)$, if $g^{\pi,h}(h) \ge 173\log\frac{4}{\delta}$, define $T_0^{\pi,h}(\delta)$ to be*

$$T_0^{\pi,h}(\delta) \coloneqq \min\Big\{T' \ge 1 : g^{\pi,h}(T') \ge 173\log\frac{4}{\delta}\Big\}. \tag{40}$$

Then with probability at least $1-\delta$, for all $T_0^{\pi,h}(\delta) \le T \le h+1$, we have

$$\begin{aligned}&\Big|R_T^{\pi,h} - \big(V_0^{\pi,h}(S_0) - V_T^{\pi,h}(S_T)\big)\Big| \\ &\qquad \le \max\Big\{\bar{K}_T^{\pi,h}\sqrt{3g^{\pi,h}(T)\Big(2\log\log(\frac{3}{2}g^{\pi,h}(T)) + \log\frac{2}{\delta}\Big)}, (\bar{K}_T^{\pi,h})^2\Big\}.\end{aligned} \tag{41}$$

The proof is presented in App. E.1.

Following Corollary establishes the finite-time concentration of $R_T^{\pi,h}$ around the quantity $V_0^{\pi,h}(S_0)$. This results is analogous to the concentration bounds in the average reward setup given in Theorem 18 and concentration bounds in the discounted reward setup given in Corollary 38.

Corollary 49 *For any policy $\pi \in \Pi_{\text{FD}}$, we have:*

1. *For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\Big|R_T^{\pi,h} - V_0^{\pi,h}(S_0)\Big| \le \bar{K}_T^{\pi,h}\sqrt{2T\log\frac{2}{\delta}} + \bar{H}_T^{\pi,h}.$$

2. *For any $\delta \in (0,1)$, if $g^{\pi,h}(h) \ge 173\log\frac{4}{\delta}$, define $T_0^{\pi,h}(\delta)$ as specified in (40). Then with probability at least $1-\delta$, for all $T_0^{\pi,h}(\delta) \le T \le h+1$, we have*

$$\Big|R_T^{\pi,h} - V_0^{\pi,h}(S_0)\Big| \le \max\Big\{\bar{K}_T^{\pi,h}\sqrt{3T\big(2\log\log\big(\frac{3T}{2}\big) + \log\frac{2}{\delta}\big)}, (\bar{K}_T^{\pi,h})^2\Big\} + \bar{H}_T^{\pi,h}.$$

The proof is presented in App. E.2.

5.3 Sample Path Behavior of Performance Difference of Two Policies

As an implication of the results presented in Sec. 5.2, we characterize the sample path behavior of the difference in cumulative rewards between any two policies. As a consequence, we derive the non-asymptotic concentration of the difference in rewards between any two optimal policies. These concentration bounds are presented in the following two corollaries.

Corollary 50 *Consider two policies $\pi_1, \pi_2 \in \Pi_{\mathrm{FD}}$. Let $\{S_t^{\pi_1}\}_{t=0}^h$ and $\{S_t^{\pi_2}\}_{t=0}^h$ denote the random sequences of the states encountered by policies π_1 and π_2 respectively. Following upper-bounds hold for the difference between the cumulative reward received by the two policies $\left|R_T^{\pi_1,h} - R_T^{\pi_2,h}\right|$.*

1. *For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\begin{aligned}&\Big|\big|R_T^{\pi_1,h} - R_T^{\pi_2,h}\big| - \big|\big[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\big] - \big[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\big]\big|\Big| \\ &\le \bar{K}_T^{\pi_1,h}\sqrt{2g^{\pi_1,h}(T)\log\frac{4}{\delta}} + \bar{K}_T^{\pi_2,h}\sqrt{2g^{\pi_2,h}(T)\log\frac{4}{\delta}}. \end{aligned} \tag{42}$$

2. *For any $\delta \in (0,1)$, if $\min\left\{g^{\pi_1,h}(h), g^{\pi_2,h}(h)\right\} \ge 173\log\frac{8}{\delta}$, define $T_0^{\pi,h}(\delta)$ as specified in (40) and let*

$$T_0^h(\delta) \coloneqq \max\left\{T_0^{\pi_1,h}(\frac{\delta}{2}), T_0^{\pi_2,h}(\frac{\delta}{2})\right\}.$$

Then, with probability at least $1-\delta$, for all $T_0^h(\delta) \le T \le h+1$, we have

$$\begin{aligned}&\Big|\big|R_T^{\pi_1,h} - R_T^{\pi_2,h}\big| - \big|\big[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\big] - \big[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\big]\big|\Big| \\ &\le \max\left\{\bar{K}_T^{\pi_1,h}\sqrt{3g^{\pi_1,h}(T)\Big(2\log\log(\frac{3}{2}g^{\pi_1,h}(T)) + \log\frac{4}{\delta}\Big)}, (\bar{K}_T^{\pi_1,h})^2\right\} \\ &+ \max\left\{\bar{K}_T^{\pi_2,h}\sqrt{3g^{\pi_2,h}(T)\Big(2\log\log(\frac{3}{2}g^{\pi_2,h}(T)) + \log\frac{4}{\delta}\Big)}, (\bar{K}_T^{\pi_2,h})^2\right\}. \end{aligned} \tag{43}$$

The proof is presented in App. E.3.

Corollary 51 *Consider two optimal policies $\pi_1^*, \pi_2^* \in \Pi_{\mathrm{FD}}$. Let $\{S_t^{\pi_1^*}\}_{t=0}^h$ and $\{S_t^{\pi_2^*}\}_{t=0}^h$ denote the random sequences of states encountered by optimal policies π_1^* and π_2^*. To simplify the expression, we assume the system starts at a fixed initial state, i.e., $S_0^{\pi_1^*} = S_0^{\pi_2^*}$. Then for the difference between the cumulative rewards received by the two optimal policies $\left|R_T^{\pi_1^*,h} - R_T^{\pi_2^*,h}\right|$, we have:*

1. *For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\Big|\big|R_T^{\pi_1^*,h} - R_T^{\pi_2^*,h}\big| - \big|V_T^{*,h}(S_T^{\pi_2^*}) - V_T^{*,h}(S_T^{\pi_1^*})\big|\Big| \le 2\Big(\bar{K}_T^{*,h}\sqrt{2g^{*,h}(T)\log\frac{4}{\delta}}\Big). \tag{44}$$

2. *For any $\delta \in (0,1)$, if $g^{*,h}(h) \geq 173 \log \frac{4}{\delta}$, define $T_0^{\pi^*,h}(\delta)$ as specified in (40). Then with probability at least $1-\delta$, for all $T_0^{\pi^*,h}(\delta) \leq T \leq h+1$, we have*

$$\begin{aligned}&\Big|\big|R_T^{\pi_1^*,h} - R_T^{\pi_2^*,h}\big| - \big|V_T^{*,h}(S_T^{\pi_2^*}) - V_T^{*,h}(S_T^{\pi_1^*})\big|\Big| \\ \leq& 2\Big(\max\Big\{\bar{K}_T^{*,h}\sqrt{3g^{*,h}(T)\Big(2\log\log(\frac{3}{2}g^{*,h}(T)) + \log\frac{4}{\delta}\Big)}, (\bar{K}_T^{*,h})^2\Big\}\Big). \end{aligned} \tag{45}$$

Proof Since both policies $\pi_1^*, \pi_2^* \in \Pi_{\mathrm{FD}}$ are optimal policies, by the definition, we have

$$V_t^{\pi_1^*,h}(s) = V_t^{\pi_2^*,h}(s) = V_t^{*,h}(s), \quad \forall s \in \mathcal{S}, \quad \forall t \in \{0,1,\ldots,h+1\}.$$

As a result, by the assumption that $S_0^{\pi_1^*} = S_0^{\pi_2^*}$, we have

$$\Big|V_0^{*,h}(S_0^{\pi_1^*}) - V_0^{*,h}(S_0^{\pi_2^*})\Big| = 0.$$

In addition, we have

$$\bar{K}_T^{\pi_1^*,h} = \bar{K}_T^{\pi_2^*,h} = \bar{K}_T^{*,h} \quad \text{and} \quad g^{\pi_1^*,h}(T) = g^{\pi_2^*,h}(T) = g^{*,h}(T).$$

As a result, by Corollary 50, the difference $\big|R_T^{\pi_1^*,h} - R_T^{\pi_2^*,h}\big|$ satisfies the non-asymptotic concentration rates in Corollary 50 with the RHS of (42)–(43) being simplified to RHS of (44)–(45). ■

6. Extension of the Results to Random Reward

In this section, we extend the results of Sec. 2 to settings where the reward at time t depends not only on the state-action pair (s,a) but also on an exogenous process. Consider an average reward MDP with state space $\mathcal{S}$ and action space $\mathcal{A}$. Similar to Sec. 2, we assume that $\mathcal{S}$ and $\mathcal{A}$ are finite sets. The state evolves in a controlled Markov manner with transition matrix P. Let $\{E_t\}_{t\geq 0}$ denote an exogenous process which satisfies the following property

$$\mathbb{E}\big[E_t\big|S_{0:t}, A_{0:t}\big] = \mathbb{E}\big[E_t\big|S_t, A_t\big], \quad \forall t \geq 0. \tag{46}$$

At each time t, the system yields a per-step random reward $\tilde{r}(S_t, A_t, E_t)$, where $\tilde{r} : \mathcal{S} \times \mathcal{A} \times \mathcal{E} \to [0, R_{\max}]$. We use the notation $\tilde{r}$ to denote the new definition of reward function distinguishing it from r in Sec. 2, and denote this model by $\tilde{M} = (P, \tilde{r})$. Let $\tilde{R}_T^\pi$ denote the total reward received by policy π until time T, i.e.,

$$\tilde{R}_T^\pi = \sum_{t=0}^{T-1} \tilde{r}(S_t, A_t, E_t), \quad A_t \sim \pi(S_{0:t}, A_{0:t-1}).$$

Compared to the model in Sec. 2, the only change in this model is the use of $\tilde{r}(S_t, A_t, E_t)$ as the per-step reward, where the process $\{E_t\}_{t\geq 0}$ satisfies the conditional independence property in (46). As a result, by defining reward function $r(s,a)$ as

$$r(s,a) = \mathbb{E}\big[\tilde{r}(S_t, A_t, E_t)\big|S_t = s, A_t = a\big], \tag{47}$$

model $\tilde{\mathcal{M}} = (P, \tilde{r})$ reduces to the model $\mathcal{M} = (P, r)$ in Sec. 2. Therefore by the results in Sec. 3, we can establish the concentration behavior of the process R_T^π, where

$$R_T^\pi = \sum_{t=0}^{T-1} r(S_t, A_t) = \sum_{t=0}^{T-1} \mathbb{E}\big[\tilde{r}(S_t, A_t, E_t) \big| S_t, A_t = \pi(S_t)\big].$$

However, in this setting, we are interested in the concentration of cumulative reward process $\tilde{R}_T^\pi = \sum_{t=0}^{T-1} \tilde{r}(S_t, A_t, E_t)$. To simplify the analysis, we define the quantities associated with the model $\tilde{\mathcal{M}}$ based on the reduced model $\mathcal{M}$.

Definition 52 *Let $\mathcal{M} = (P, r)$ with $r(s, a)$ defined in (47) be the reduced model of $\tilde{\mathcal{M}} = (P, \tilde{r})$. For the model $\tilde{\mathcal{M}}$, we define policies $\pi \in \tilde{\Pi}$, policy sets $\tilde{\Pi}_{\mathrm{AR}}$ and $\tilde{\Pi}_{\mathrm{SD}}$, long-run expected reward function $\tilde{J}^\pi$, differential value function $\tilde{V}^\pi$, optimal performance $\tilde{J}^*$, and optimal policy $\tilde{\pi}^*$ as the corresponding quantities of the reduced model $\mathcal{M}$ in Sec. 2.*

For any policy $\pi \in \tilde{\Pi}_{\mathrm{AR}}$, we define the maximum absolute deviation of value function $\tilde{K}^\pi$ similar to Sec. 3, i.e.,

$$\tilde{K}^\pi \coloneqq \max_{s, s_+ \in \mathcal{S}} \Big| \tilde{V}^\pi(s_+) - \mathbb{E}\big[\tilde{V}^\pi(S_+) \big| s, \pi(s)\big] \Big|. \tag{48}$$

In this section, we define a similar quantity for the reward function $\tilde{r}(S_t, A_t, E_t)$.

Definition 53 *For any policy $\pi \in \tilde{\Pi}_{\mathrm{AR}}$, we define the maximum absolute deviation of reward function $\tilde{K}_r^\pi$ as follows*

$$\tilde{K}_r^\pi \coloneqq \max_{\substack{s, s' \in \mathcal{S} \\ a' \in \mathcal{A} \\ e' \in \mathcal{E}}} \Big| \tilde{r}(s', a', e') - \mathbb{E}\big[\tilde{r}(S, A, E) \big| S = s, A = \pi(s)\big] \Big|. \tag{49}$$

The following theorem establishes the concentration of cumulative reward $\tilde{R}_T^\pi$ around the quantity $T\tilde{J}^\pi - \big(\tilde{V}^\pi(S_T) - \tilde{V}^\pi(S_0)\big)$. Compared to the results in Sec. 3, the concentration of cumulative reward is characterized by a weaker bound due to the added stochasticity in the cumulative reward process.

Theorem 54 *For any policy $\pi \in \tilde{\Pi}_{\mathrm{AR}}$, the following upper-bounds hold:*

1. *For any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have*

$$\Big| \tilde{R}_T^\pi - T\tilde{J}^\pi - \big(\tilde{V}^\pi(S_0) - \tilde{V}^\pi(S_T)\big) \Big| \le \tilde{K}^\pi \sqrt{2T \log \frac{4}{\delta}} + \tilde{K}_r^\pi \sqrt{2T \log \frac{4}{\delta}}. \tag{50}$$

2. *For any $\delta \in (0, 1)$, for all $T \ge T_0(\delta) \coloneqq \max\Big\{ \Big\lceil \frac{173}{\tilde{K}^\pi} \log \frac{8}{\delta} \Big\rceil, \Big\lceil \frac{173}{\tilde{K}_r^\pi} \log \frac{8}{\delta} \Big\rceil \Big\}$, with probability at least $1 - \delta$, we have*

$$\Big| \tilde{R}_T^\pi - T\tilde{J}^\pi - \big(\tilde{V}^\pi(S_0) - \tilde{V}^\pi(S_T)\big) \Big| \tag{51}$$

$$\le \max\Big\{ \tilde{K}^\pi \sqrt{3T\Big(2 \log\log \frac{3T}{2} + \log \frac{4}{\delta}\Big)}, (\tilde{K}^\pi)^2 \Big\} \tag{52}$$

$$+ \max\Big\{ \tilde{K}_r^\pi \sqrt{3T\Big(2 \log\log \frac{3T}{2} + \log \frac{4}{\delta}\Big)}, (\tilde{K}_r^\pi)^2 \Big\}. \tag{53}$$

The proof is presented in App. F.2.

Remark 55 *By following arguments analogous to those used in this section, similar concentration bounds can be established for the cumulative reward process under any stationary randomized policy. For brevity, we omit these results.*

7. Conclusion

In this paper, we investigated the sample path behavior of cumulative reward in Markov decision processes. In particular, we established the asymptotic concentration of rewards, including the law of large numbers, the central limit theorem, and the law of iterated logarithm. Moreover, non-asymptotic concentrations of rewards were obtained, including an Azuma-Hoeffding-type inequality and a non-asymptotic version of the law of iterated logarithm, all applicable to a general class of stationary policies. Using these results, we characterized the relationship between two notions of regret in the literature, cumulative regret and interim cumulative regret. We showed that, in both the asymptotic and non-asymptotic settings, the two definitions are *rate equivalent* as long as either of the regrets is upper-bounded by $\tilde{\mathcal{O}}(\sqrt{T})$. Moreover, we extended our results to three different frameworks: (i) the infinite-horizon discounted reward setting, where we established non-asymptotic concentration of the cumulative discounted reward; (ii) the finite-horizon setting, where we established non-asymptotic concentration of the cumulative total reward; (iii) the infinite-horizon average setting with stochastic reward, where we established the non-asymptotic concentration of the cumulative reward process. Our proof technique in the third extension may also be applied to the discounted reward and finite-horizon frameworks; however, these extensions are omitted for brevity.

The contributions of this work are twofold: (i) It unifies two sets of literature, showing that if an algorithm achieves a regret of $\tilde{\mathcal{O}}(\sqrt{T})$ under one definition, the same rate applies to the other, thereby resolving an existing gap in our theoretical understanding. (ii) The asymptotic and non-asymptotic concentration bounds found in this work can be used to evaluate the probabilistic performance of a policy, allowing for the assessment of risk and safety in the MDP setup. Such assessments may be used in decision making problems involving high-stakes costs including safety-critical engineering systems, finance and healthcare. As a result, we believe our analysis paves the way for risk-aware decision making and reinforcement learning in such applications.

8. Disclosure of Funding

This research was supported in part by Fonds de Recherche du Québec, Nature et Technologies (FRQNT) Grant 316558 (B. Sayedana), Air Force OSR Grant FA9550-23-1-0015 (P.E. Caines), NSERC Grants RGPIN-2019-0533 (P.E. Caines) and RGPIN-2021-03511 (A. Mahajan), and Alliance Grant ALLRP 592356 (A. Mahajan).

9. Acknowledgment

The authors thank Reza Alvandi for helpful feedback and comments.

References

Yasin Abbasi-Yadkori, Peter Bartlett, Kush Bhatia, Nevena Lazic, Csaba Szepesvari, and Gellért Weisz. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, pages 3692–3702. PMLR, 2019.

Shipra Agrawal and Randy Jia. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Robert B. Ash and Catherine A. Doléans-Dade. *Probability and Measure Theory*. Academic Press, San Diego, CA, 2nd edition, 2000. ISBN 978-0-12-065202-0.

Peter Auer and Ronald Ortner. Logarithmic online regret bounds for undiscounted reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

Akshay Balsubramani. Sharp finite-time iterated-logarithm martingale concentration. *arXiv preprint arXiv:1405.2639*, 2014.

Peter L. Bartlett and Ambuj Tewari. Regal: A regularization-based algorithm for reinforcement learning in weakly communicating MDPs. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 35–42. AUAI Press, 2009.

Peter L Bartlett and Ambuj Tewari. Regal: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. *arXiv preprint arXiv:1205.2661*, 2012.

Nicole Bäuerle and Alexander Glauner. Minimizing spectral risk measures applied to Markov decision processes. *Mathematical Methods of Operations Research*, 93(3):499–521, 2021.

Nicole Bäuerle and Alexander Glauner. Markov decision processes with recursive risk measures. *European Journal of Operational Research*, 298:462–480, 2022.

Nicole Bäuerle and Jonathan Ott. Markov decision processes with average-value-at-risk criteria. *Mathematical Methods of Operations Research*, 74(3):457–478, 2011.

Nicole Bäuerle and Ulrich Rieder. More risk-sensitive Markov decision processes. *Mathematics of Operations Research*, 39(1):37–55, 2014.

Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 449–458. PMLR, 06–11 Aug 2017.

Marc G. Bellemare, Nicolas Le Roux, Rémi Munos, Mark Rowland, and Will Dabney. Distributional reinforcement learning with linear function approximation. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pages 2208–2216, Naha, Okinawa, Japan, 2019. PMLR. URL `https://proceedings.mlr.press/v89/bellemare19a.html`.

Marc G. Bellemare, Will Dabney, and Mark Rowland. *Distributional Reinforcement Learning*. MIT Press, 2023. http://www.distributional-rl.org.

Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control, Volume II*. Athena Scientific, Belmont, MA, 4th edition, 2012a. ISBN 978-1-886529-44-1.

Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control, Volume I*. Athena Scientific, Belmont, MA, 4th edition, 2012b. ISBN 978-1-886529-44-1.

Patrick Billingsley. *Convergence of Probability Measures*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ, 2nd edition, 2013. ISBN 978-1-118-12237-2.

Anup Biswas and Vivek S. Borkar. Ergodic risk-sensitive control-a survey. *Annual Reviews in Control*, 55:118–141, 2023. doi: 10.1016/j.arcontrol.2023.03.001. Review article – published March 2023.

Victor Boone and Zihan Zhang. Achieving tractable minimax optimal regret in average reward MDPs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

V. S. Borkar. Q-learning for risk-sensitive control. *Mathematics of Operations Research*, 27 (2):294–311, 2002.

Hippolyte Bourel, Odalric Maillard, and Mohammad Sadegh Talebi. Tightening exploration in upper confidence reinforcement learning. In *International Conference on Machine Learning*, pages 1056–1066. PMLR, 2020.

Pierre Brémaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*, volume 31 of *Texts in Applied Mathematics*. Springer Science & Business Media, New York, 2013. ISBN 978-1-4757-4386-2.

M. P. Chapman, R. Bonalli, K. M. Smith, I. Yang, M. Pavone, and C. J. Tomlin. Risk-sensitive safety analysis using conditional value-at-risk. *IEEE Transactions on Automatic Control*, 67(12):6521–6536, Dec 2022. doi: 10.1109/TAC.2021.3131149.

Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. Risk-sensitive and robust decision-making: a CVaR optimization approach. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1522–1530. Curran Associates, Inc., 2015.

Kai Lai Chung. *Markov Chains with Stationary Transition Probabilities: 2d Ed*. Springer, 1967.

Kun-Jen Chung and Matthew J. Sobel. Discounted MDPs: Distribution functions and exponential utility maximization. *SIAM Journal on Control and Optimization*, 25(1): 49–62, 1987. doi: 10.1137/0325004. URL https://doi.org/10.1137/0325004.

Robert Cogburn. The central limit theorem for Markov processes. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, volume 6, pages 485–513. University of California Press, 1972.

Stefano P. Coraluppi and Steven I. Marcus. Risk-sensitive and minimax control of discrete-time, finite-state Markov decision processes. *Automatica*, 35(2):301–309, 1999.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 1096–1105, 2018a. URL `https://proceedings.mlr.press/v80/dabney18a.html`.

Will Dabney, Mark Rowland, Marc G. Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, 2018b.

Jingliang Duan, Yang Guan, Shengbo Eben Li, Yangang Ren, and Bo Cheng. Distributional soft actor-critic: Off-policy reinforcement learning for addressing value estimation errors. *IEEE Transactions on Neural Networks and Learning Systems*, 33(11):6584–6598, 2021. doi: 10.1109/TNNLS.2021.3056130. URL `https://arxiv.org/abs/2001.02811`.

Marie Duflo. *Random Iterative Models*, volume 34 of *Applications of Mathematics*. Springer Science & Business Media, Berlin, Heidelberg, 2013. ISBN 978-3-662-03203-0.

Amir-massoud Farahmand. Value function in frequency domain and the characteristic value iteration algorithm. In *Advances in Neural Information Processing Systems*, Vancouver, Canada, 2019.

Sarah Filippi, Olivier Cappé, and Aurélien Garivier. Optimism in reinforcement learning and kullback-leibler divergence. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 115–122. IEEE, 2010.

Ronan Fruit. *Exploration-exploitation dilemma in Reinforcement Learning under various form of prior knowledge*. Theses, Université de Lille 1, Sciences et Technologies; CRIStAL UMR 9189, November 2019. URL `https://theses.hal.science/tel-02388395`.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *International Conference on Machine Learning*, pages 1578–1586. PMLR, 2018.

Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Improved analysis of UCRL2 with empirical bernstein inequality. *arXiv preprint arXiv:2007.05456*, 2020.

Peter Hall and Christopher C. Heyde. *Martingale Limit Theory and Its Application*. Probability and Mathematical Statistics. Academic Press, New York, 1980. ISBN 9780123193501.

Onésimo Hernández-Lerma and Jean B. Lasserre. *Further Topics on Discrete-Time Markov Control Processes*, volume 42 of *Applications of Mathematics*. Springer Science & Business Media, Berlin, Heidelberg, 2012. ISBN 978-1-4612-7067-1.

Ronald A. Howard and James E. Matheson. Risk-sensitive Markov decision processes. *Management Science*, 18(7):356–369, 1972.

David H. Jacobson. Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games. *IEEE Transactions on Automatic Control*, 18(2):124–131, 1973.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(51):1563–1600, 2010.

Galin L. Jones. On the Markov chain central limit theorem. *Probability Surveys*, 1:299–320, 2004. URL `http://eudml.org/doc/223812`.

L.C.M. Kallenberg. Classification problems in MDPs. In *Markov Processes and Controlled Markov Chains*, pages 151–165. Springer, Boston, MA, 2002.

Claude Kipnis and S. R. S. Varadhan. Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Communications in Mathematical Physics*, 104(1):1–19, 1986. doi: 10.1007/BF01210778.

Ioannis Kontoyiannis and Sean Meyn. Large deviations asymptotics and the spectral theory of multiplicatively regular Markov processes. *Electronic Journal of Probability*, 10:61–123, 2005.

Ioannis Kontoyiannis and Sean P Meyn. Spectral theory and limit theorems for geometrically ergodic Markov processes. *The Annals of Applied Probability*, 13(1):304–362, 2003.

Kailasam Lakshmanan, Ronald Ortner, and Daniil Ryabko. Improved regret bounds for undiscounted continuous reinforcement learning. In *International conference on machine learning*, pages 524–532. PMLR, 2015.

Claudio Landim. Central limit theorem for Markov processes. In *From Classical to Modern Probability: CIMPA Summer School 2001*, pages 145–205. Springer, 2003.

Alix Lhéritier and Nicolas Bondoux. A Cramér distance perspective on quantile regression based distributional reinforcement learning. *arXiv preprint arXiv:2110.00535*, 2021.

Clare Lyle, Pablo Samuel Castro, and Marc G. Bellemare. A comparative analysis of expected and distributional reinforcement learning. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)*, pages 4504–4511, Honolulu, Hawaii, USA, 2019. doi: 10.1609/aaai.v33i01.33014504.

Nelly Maigret. Théorème de limite centrale fonctionnel pour une chaîne de Markov récurrente au sens de harris et positive. In *Annales de l'institut Henri Poincaré. Section B. Calcul des probabilités et statistiques*, volume 14, pages 425–440, 1978.

Petr Mandl. On the variance in controlled Markov chains. *Kybernetika*, 7(1):1–12, 1971. URL `https://eudml.org/doc/27906`.

Petr Mandl and Michaela Lausmanova. Two extensions of asymptotic methods in controlled Markov chains. *Annals of Operations Research*, 28:67–80, 1991. doi: 10.1007/BF02055575.

Michael Maxwell and Michael Woodroofe. Central limit theorems for additive functionals of Markov chains. *Annals of probability*, pages 713–724, 2000.

Sean P. Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge Mathematical Library. Springer Science & Business Media, Cambridge, 2012. ISBN 978-1-4612-4244-9.

Christopher W. Miller and Insoon Yang. Optimal control of conditional value-at-risk in continuous time. *Mathematics of Operations Research*, 46(4):1691–1711, 2021.

Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Nonparametric return distribution approximation for reinforcement learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 799–806, Haifa, Israel, 2010a.

Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Parametric return density estimation for reinforcement learning. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 368–375, Catalina Island, California, USA, 2010b.

Jacques Neveu. *Discrete-Parameter Martingales*, volume 10 of *North-Holland Mathematical Library*. North-Holland, Amsterdam, 1975. ISBN 978-0-7204-2830-5.

Tung Nguyen, Quang Ngo, Shinji Hasegawa, and Tetsuya Asai. Distributional reinforcement learning via moment matching. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162 of *Proceedings of Machine Learning Research*, pages 16684–16700. PMLR, 2022. URL `https://proceedings.mlr.press/v162/nguyen22c.html`.

Seppo Niemi and Esa Nummelin. *Central Limit Theorems for Markov Random Walks*, volume 54 of *Commentationes Physico-Mathematicae*. Societas Scientiarum Fennica, Helsinki, 1982.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (More) Efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2377–2386, New York, New York, USA, 20–22 Jun 2016. PMLR.

Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown Markov decision processes: A Thompson sampling approach. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Hoboken, NJ, 2014. ISBN 978-1-118-62013-9.

Jian Qian, Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Exploration bonus for regret minimization in discrete and continuous average reward MDPs. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Maxim Raginsky and Igal Sason. *Concentration of Measure Inequalities in Information Theory, Communications, and Coding*, volume 10 of *Foundations and Trends in Communications and Information Theory*. Now Publishers Inc., 2014. ISBN 978-1-60198-839-5.

Mark Rowland, Marc G. Bellemare, Will Dabney, Rémi Munos, and Yee Whye Teh. An analysis of categorical distributional reinforcement learning. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 29–37, 2018. URL `https://proceedings.mlr.press/v84/rowland18a.html`.

Mark Rowland, Robert Dadashi, Saurabh Kumar, Rémi Munos, Marc G. Bellemare, and Will Dabney. Statistics and samples in distributional reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 5528–5536, Long Beach, California, USA, 2019. PMLR.

Mark Rowland, Yunhao Tang, Clare Lyle, Rémi Munos, Marc G Bellemare, and Will Dabney. The statistical benefits of quantile temporal-difference learning for value estimation. In *International Conference on Machine Learning*, pages 29210–29231. PMLR, 2023.

Mark Rowland, Rémi Munos, Mohammad Gheshlaghi Azar, Yunhao Tang, Georg Ostrovski, Anna Harutyunyan, Karl Tuyls, Marc G Bellemare, and Will Dabney. An analysis of quantile temporal-difference learning. *Journal of Machine Learning Research*, 25(163): 1–47, 2024.

Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

Andrzej Ruszczyński. Risk-averse dynamic programming for Markov decision processes. *Mathematical programming*, 125:235–261, 2010.

Borna Sayedana, Peter E. Caines, and Aditya Mahajan. Asymptotic normality of cumulative cost in linear quadratic regulators. In *2024 IEEE 63rd Conference on Decision and Control (CDC)*, pages 1856–1862. IEEE, 2024.

Sumeet Singh, Yinlam Chow, Anirudha Majumdar, and Marco Pavone. A framework for time-consistent, risk-sensitive model predictive control: Theory and algorithms. *IEEE Transactions on Automatic Control*, 63(10):3328–3343, 2018.

Matthew J. Sobel. The variance of discounted Markov decision processes. *Journal of Applied Probability*, 19(4):794–802, 1982.

Rayadurgam Srikant. Rates of convergence in the central limit theorem for Markov chains, with an application to td learning. *Mathematics of Operations Research*, 2025.

William F. Stout. *Almost Sure Convergence*. Academic Press, New York, 1974. ISBN 978-0-12-672950-4.

Mohammad Sadegh Talebi and Odalric-Ambrym Maillard. Variance-aware regret bounds for undiscounted reinforcement learning in MDPs. In *Algorithmic Learning Theory*, pages 770–805. PMLR, 2018.

Georgios Theocharous, Zheng Wen, Yasin Abbasi-Yadkori, and Nikos Vlassis. Posterior sampling for large scale reinforcement learning. *arXiv preprint arXiv:1711.07979*, 2017.

Yuheng Wang and Margaret P. Chapman. Risk-averse autonomous systems: A brief history and recent developments from the perspective of optimal control. *Artificial Intelligence*, 311:103743, October 2022. ISSN 0004-3702. doi: 10.1016/j.artint.2022.103743. URL `https://www.sciencedirect.com/science/article/pii/S0004370222000832`.

Peter Whittle. Risk-sensitive linear/quadratic/Gaussian control. *Advances in Applied Probability*, 13(4):764–777, 1981.

Derek Yang, Li Zhao, Zichuan Lin, Tao Qin, Jiang Bian, and Tie-Yan Liu. Fully parameterized quantile function for distributional reinforcement learning. *Advances in neural information processing systems*, 32, 2019.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pages 7304–7312. PMLR, 2019.

Zihan Zhang and Xiangyang Ji. Regret minimization for reinforcement learning by evaluating the optimal bias function. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Zihan Zhang and Qiaomin Xie. Sharper model-free reinforcement learning for average-reward Markov decision processes. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5476–5477. PMLR, 2023.

Appendix Contents

Appendix A. Background on Markov Chain Theory

Consider a time-homogeneous Markov chain defined on a finite state space $\mathcal{S}$. Let P denote the state transition probability and P^k denote the k-step state transition probability. Then we use the following terminology.

- Given $s, s' \in \mathcal{S}$, state s' is said to be *accessible from* s, if there exists a finite time $k \geq 0$ such that $P^k(s'|s) > 0$.
- States s and s' in $\mathcal{S}$ are said to *communicate* if s is accessible from s' and s' is accessible from s.
- Communication relation is reflexive, symmetric, and transitive. Therefore, communication relation is an equivalence relation, and it generates a partition of the state space $\mathcal{S}$ into disjoint equivalence classes called *communication classes* (Brémaud, 2013).
- Let T_s denote the hitting time of state s. State s is called *recurrent* if

$$\mathbb{P}\big(T_s < \infty \;\big|\; S_0 = s\big) = 1,$$

  and otherwise it is called *transient*.
- A *recurrent class* is a communication class where every state within the class is recurrent.
- A *transient class* is a communication class where every state within the class is transient.

Appendix B. Background on Martingales

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A *filtration* $\{\mathcal{F}_t\}_{t\geq 0}$ is a non-decreasing family of sub-sigma fields of $\mathcal{F}$. A random sequence $\{X_t\}_{t\geq 0}$ is called *integrable* if $\mathbb{E}[|X_t|] < \infty$ for all $t \geq 0$. A random sequence $\{X_t\}_{t\geq 0}$ is called *adapted* to the filtration $\{\mathcal{F}_t\}_{t\geq 0}$ if X_t is $\mathcal{F}_t$-measurable for all $t \geq 0$.

Definition 56 *An integrable sequence $\{X_t\}_{t\geq 0}$ adapted to the filtration $\{\mathcal{F}_t\}_{t\geq 0}$ is called a martingale if*

$$\mathbb{E}[X_{t+1}|\mathcal{F}_t] = X_t, \quad a.s. \quad \forall t \geq 0.$$

Definition 57 *Let $\{c_t\}_{t\geq 1}$ be a sequence of real numbers and C be a positive real number. A real integrable sequence $\{Y_t\}_{t\geq 1}$ adapted to the filtration $\{\mathcal{F}_t\}_{t\geq 0}$ is called:*

1. *Martingale Difference Sequence (MDS) if*

$$\mathbb{E}[Y_t|\mathcal{F}_{t-1}] = 0, \quad a.s. \quad \forall t \geq 1.$$

2. *Sequentially bounded MDS with respect to the sequence $\{c_t\}_{t\geq 1}$ if it is an MDS and*

$$|Y_t| \leq c_t, \quad a.s. \quad \forall t \geq 1.$$

3. Uniformly bounded MDS with respect to the constant C if it is an MDS and

$$|Y_t| \leq C, \quad a.s. \quad \forall t \geq 0.$$

There is a unique MDS corresponding to a martingale and vise versa. In particular, given a martingale $\{X_t\}_{t\geq 0}$, the corresponding MDS $\{Y_t\}_{t\geq 1}$ is defined as

$$Y_t \coloneqq X_t - X_{t-1}, \quad \forall t \geq 1.$$

Moreover, given an MDS $\{Y_t\}_{t\geq 1}$, the corresponding martingale sequence $\{X_t\}_{t\geq 0}$ is defined as

$$X_0 = 0, \quad X_T = \sum_{t=1}^{T} Y_t, \quad \forall T \geq 1.$$

Consider a martingale $\{X_t\}_{t\geq 0}$ such that $\{X_t^2\}_{t\geq 0}$ is integrable. The *increasing process* $\{A_t\}_{t\geq 1}$ associated with the sequence $\{X_t^2\}_{t\geq 0}$ is defined as

$$A_1 = \mathbb{E}[X_1^2|\mathcal{F}_0] - X_1^2, \quad A_t = \mathbb{E}[X_t^2|\mathcal{F}_{t-1}] - X_{t-1}^2 + A_{t-1}, \quad \forall t \geq 2.$$

Let $\{Y_t\}_{t\geq 1}$ be the MDS corresponding to $\{X_t\}_{t\geq 0}$. Then, we can express $\{A_t\}_{t\geq 1}$ in terms of $\{Y_t^2\}_{t\geq 1}$. In particular, we have

$$\begin{aligned} A_t &= \mathbb{E}[X_t^2|\mathcal{F}_{t-1}] - X_{t-1}^2 + A_{t-1} \\ &= \mathbb{E}[X_{t-1}^2|\mathcal{F}_{t-1}] + 2\mathbb{E}[Y_t|\mathcal{F}_{t-1}]X_{t-1} + \mathbb{E}[Y_t^2|\mathcal{F}_{t-1}] - X_{t-1}^2 + A_{t-1} \\ &= \mathbb{E}[Y_t^2|\mathcal{F}_{t-1}] + A_{t-1}. \end{aligned}$$

As a result, we have

$$A_T = \sum_{t=1}^{T} \mathbb{E}[Y_t^2|\mathcal{F}_{t-1}], \quad \forall T \geq 1.$$

Therefore, we sometimes say that $\{A_t\}_{t\geq 1}$ is the increasing sequence associated with $\{Y_t^2\}_{t\geq 1}$.

Martingale sequences are an important class of stochastic processes. Both asymptotic and non-asymptotic concentration of martingale sequences have been well studied. In Sec. B.1 and B.2, we present the asymptotic and non-asymptotic concentration characteristics of martingales with bounded MDS.

B.1 Asymptotic Concentration

B.1.1 Strong Law of Large Numbers

The first asymptotic results presented in this section is a version of Strong Law of Large numbers for martingale difference sequences.

Theorem 58 (see (Stout, 1974, Theorem 3.3.1)) *Let $\{Y_t\}_{t\geq 1}$ be an MDS and $\{a_t\}_{t\geq 1}$ be a sequence of positive and $\mathcal{F}_{t-1}$-measurable real numbers such that $\lim_{t\to\infty} a_t = \infty$. If for some $0 < p \leq 2$, we have:*

$$\sum_{t=1}^{\infty} \frac{\mathbb{E}\big(|Y_t|^p|\mathcal{F}_{t-1}\big)}{a_t^p} < \infty.$$

Then:

$$\frac{\sum_{t=t}^{T} Y_t}{T} \to 0, \quad a.s.$$

B.1.2 Central Limit Theorem

Following theorem characterizes a version of Central Limit Theorem for martingale sequences with corresponding bounded MDS.

Theorem 59 (see (Billingsley, 2013, Theorem 35.11)) *Let $\{Y_t\}_{t\geq 1}$ be a sequentially bounded MDS with respect to the sequence $\{c_t\}_{t\geq 1}$. Let $\{A_t\}_{t\geq 1}$ be the increasing process associated with $\{Y_t^2\}_{t\geq 1}$, i.e.*

$$A_T = \sum_{t=1}^{T} \mathbb{E}[Y_t^2 | \mathcal{F}_{t-1}], \quad \forall T \geq 1.$$

Define the stopping time ν_t as

$$\nu_t := \min\big\{T \geq 1 : A_T \geq t\big\}.$$

Let $\Omega_0 = \{\omega \in \Omega : \lim_{T\to\infty} A_T = \infty\}$. If $\mathbb{P}(\Omega_0) = 1$, then

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{\nu_T} Y_t \xrightarrow{(d)} \mathcal{N}(0,1).$$

B.1.3 Law of Iterated Logarithm

Following theorem characterizes a version of Law of Iterated Logarithm for uniformly bounded MDS.

Theorem 60 (see (Neveu, 1975, Proposition VII-2-7)) *Let $\{Y_t\}_{t\geq 1}$ be a uniformly bounded MDS with respect to the constant C. Furthermore, let $\{A_t\}_{t\geq 1}$ and Ω_0 be as defined in Theorem 59. Then, for almost all $\omega \in \Omega_0$, we have*

$$\liminf_{T\to\infty} \frac{\sum_{t=1}^{T} Y_t}{\sqrt{2A_T \log\log A_T}} = -1, \quad \limsup_{T\to\infty} \frac{\sum_{t=1}^{T} Y_t}{\sqrt{2A_T \log\log A_T}} = 1.$$

Non-asymptotic high-probability bounds with similar functional dependence on the horizon T also exist for martingales. These bounds are presented in Sec. B.2.

B.2 Non-Asymptotic Concentration

B.2.1 Azuma-Hoeffding Inequality

A famous non-asymptotic concentration for martingale sequences is Azuma-Hoeffding inequality.

Theorem 61 (see (Raginsky and Sason, 2014, Theorem 2.2.1)) *Let $\{Y_t\}_{t\geq 1}$ be a sequentially bounded MDS with respect to the sequence $\{c_t\}_{t\geq 1}$. Then for all $T \geq 1$ and for all $\epsilon > 0$, we have*

$$\mathbb{P}\left(\left|\sum_{t=1}^{T} Y_t\right| \geq \epsilon\right) \leq 2\exp\left(\frac{-\epsilon^2}{2\sum_{t=1}^{T} c_t^2}\right).$$

By rewriting the statement of Theorem 61, we get following equivalent form of Azuma-Hoeffding inequality.

Corollary 62 *We have following statements*

1. *Let $\{Y_t\}_{t\geq 1}$ be a sequentially bounded MDS with respect to the sequence $\{c_t\}_{t\geq 1}$. For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\Big|\sum_{t=1}^{T} Y_t\Big| \leq \sqrt{2\sum_{t=1}^{T} c_t^2 \log\frac{2}{\delta}}.$$

2. *Let $\{Y_t\}_{t\geq 1}$ be a uniformly bounded MDS with respect to the constant C. For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\Big|\sum_{t=1}^{T} Y_t\Big| \leq C\sqrt{2T \log\frac{2}{\delta}}.$$

The proof of Part 1 follows by equating the RHS of Theorem 61 to δ and solving for ϵ. The proof of Part 2 follows by substituting the sequence $\{c_t\}_{t\geq 1}$ with the constant C in the RHS of Part 1.

B.2.2 Non-Asymptotic Law of Iterated Logarithm

Following result is a finite-time analogue of Law of Iterated Logarithm. This result shows that for a large enough horizon T, the growth rate of a martingale sequence is of the order $\mathcal{O}\Big(\sqrt{T\log\log(T)}\Big)$ with high probability.

Theorem 63 (see (Balsubramani, 2014, Theorem 4)) *Let $\{Y_t\}_{t\geq 1}$ be a sequentially bounded MDS with respect to the sequence $\{c_t\}_{t\geq 1}$. For any $\delta \in (0,1)$, for all $T \geq T_0(\delta) \coloneqq \min\Big\{T \,:\, \sum_{t=1}^{T} c_t^2 \geq 173\log\frac{4}{\delta}\Big\}$, with probability at least $1-\delta$, we have*

$$\Big|\sum_{t=1}^{T} Y_t\Big| \leq \sqrt{3\Big(\sum_{t=1}^{T} c_t^2\Big)\Big(2\log\log\frac{3\sum_{t=1}^{T} c_t^2}{2\big|\sum_{t=1}^{T} Y_t\big|} + \log\frac{2}{\delta}\Big)}. \tag{54}$$

For the simplicity of the analysis, we state a slightly simplified version of this theorem in the following corollary.

Corollary 64 *Let $\{Y_t\}_{t\geq 1}$ be a uniformly bounded MDS with respect to the constant C. For any $\delta \in (0,1)$, for all $T \geq T_0(\delta) \coloneqq \Big\lceil \frac{173}{C}\log\frac{4}{\delta}\Big\rceil$, with probability at least $1-\delta$, we have*

$$\Big|\sum_{t=1}^{T} Y_t\Big| \leq C\max\Bigg\{\sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{2}{\delta}\Big)}, C\Bigg\}. \tag{55}$$

Proof This corollary follows from Theorem 63, by substituting the sequence $\{c_t\}_{t\geq 1}$ with the constant C on the RHS of (54). There are two cases: either $\Big|\sum_{t=1}^{T} Y_t\Big| \leq C^2$ or $\Big|\sum_{t=1}^{T} Y_t\Big| \geq C^2$. If $\Big|\sum_{t=1}^{T} Y_t\Big| \geq C^2$, by Theorem 63, with probability at least $1-\delta$, we get:

$$\Big|\sum_{t=1}^{T} Y_t\Big| \leq C\sqrt{3T\Big(2\log\log\frac{3TC^2}{2|\sum_{t=1}^{T} Y_t|}+\log\frac{2}{\delta}\Big)} \leq C\sqrt{3T\Big(2\log\log\frac{3T}{2}+\log\frac{2}{\delta}\Big)}.$$

Otherwise, we have $\Big|\sum_{t=1}^{T} Y_t\Big| \leq C^2$. As a result, we can summarize these two cases and get that with probability at least $1-\delta$, we have

$$\Big|\sum_{t=1}^{T} Y_t\Big| \leq \max\Big\{C\sqrt{3T\Big(2\log\log\frac{3T}{2}+\log\frac{2}{\delta}\Big)}, C^2\Big\}. \tag{56}$$

■

Appendix C. Proof of Main Results for the Average Reward Setup

C.1 Preliminary Results

C.1.1 Martingale Decomposition

We first present a few preliminary lemmas. To simplify the notation, we define following martingale difference sequence.

Definition 65 *Let filtration $\mathcal{F} = \{\mathcal{F}_t\}_{t\geq 0}$ be defined as $\mathcal{F}_t \coloneqq \sigma(S_{0:t}, A_{0:t})$. For any policy $\pi \in \Pi_{\mathrm{AR}}$, let V^π denote the corresponding differential value function. We define the sequence $\{M_t^\pi\}_{t\geq 1}$ as follows*

$$M_t^\pi \coloneqq V^\pi(S_t) - \mathbb{E}\big[V^\pi(S_t) \;\big|\; S_{t-1}, \pi(S_{t-1})\big], \quad \forall t \geq 1, \tag{57}$$

where $\{S_t\}_{t\geq 0}$ denotes the random sequence of states encountered along the current sample path.

Lemma 66 *Sequence $\{M_t^\pi\}_{t\geq 1}$ is an MDS.*

Proof By the definition of $\{\mathcal{F}_t\}_{t\geq 0}$, we have that S_{t-1} is $\mathcal{F}_{t-1}$-measurable. As a result, we have

$$\begin{aligned}\mathbb{E}\Big[M_t^\pi \;\Big|\; \mathcal{F}_{t-1}\Big] &= \mathbb{E}\Big[V^\pi(S_t) - \mathbb{E}\big[V^\pi(S_t) \;\big|\; S_{t-1}, \pi(S_{t-1})\big] \;\Big|\; \mathcal{F}_{t-1}\Big] \\ &= \mathbb{E}\Big[V^\pi(S_t) \;\Big|\; \mathcal{F}_{t-1}\Big] - \mathbb{E}\Big[V^\pi(S_t) \;\Big|\; S_{t-1}, \pi(S_{t-1})\Big] = 0,\end{aligned}$$

which shows that $\{M_t^\pi\}_{t\geq 0}$ is an MDS with respect to the filtration $\{\mathcal{F}_t\}_{t\geq 0}$. ■

We now present a martingale decomposition of the cumulative reward $R_T^\pi(\omega)$.

Lemma 67 *Given any policy $\pi \in \Pi_{\mathrm{AR}}$, we can rewrite the cumulative reward R_T^π as follows*

$$R_T^\pi = TJ^\pi + \sum_{t=1}^{T} M_t^\pi + V^\pi(S_0) - V^\pi(S_T). \tag{58}$$

Proof Since $\pi \in \Pi_{\mathrm{AR}}$, (ARPE) implies that along the trajectory of states $\{S_t\}_{t=0}^T$ induced by the policy π, we have

$$r\big(S_t, \pi(S_t)\big) = J^\pi + V^\pi(S_t) - \mathbb{E}\Big[V^\pi(S_{t+1}) \;\Big|\; S_t, \pi(S_t)\Big], \quad \forall t \geq 1.$$

As a result, we have

$$\begin{aligned}
R_T^\pi =&\; TJ^\pi + \sum_{t=0}^{T-1} \Big[V^\pi(S_t) - \mathbb{E}\big[V^\pi(S_{t+1}) \;\big|\; S_t, \pi(S_t)\big]\Big] \\
\overset{(a)}{=}&\; TJ^\pi + \sum_{t=0}^{T-1} \Big[V^\pi(S_t) - \mathbb{E}\big[V^\pi(S_{t+1}) \;\big|\; S_t, \pi(S_t)\big]\Big] + V^\pi(S_T) - V^\pi(S_T) \\
\overset{(b)}{=}&\; TJ^\pi + \sum_{t=0}^{T-1} \Big[V^\pi(S_{t+1}) - \mathbb{E}\big[V^\pi(S_{t+1}) \;\big|\; S_t, \pi(S_t)\big]\Big] + V^\pi(S_0) - V^\pi(S_T) \\
\overset{(c)}{=}&\; TJ^\pi + \sum_{t=1}^{T} M_t^\pi + V^\pi(S_0) - V^\pi(S_T),
\end{aligned}$$

where (a) follows from adding and subtracting $V^\pi(S_T)$, (b) follows from re-arranging the terms in the summation, and (c) follows from the definition of $\{M_t^\pi\}_{t\geq 0}$ in (57). ■

C.1.2 A Consequence of The Union Bound

Lemma 68 *Suppose for any $\delta_1 \in (0,1)$, for all $T \geq T_1(\delta_1)$, with probability at least $1-\delta_1$, the random sequence $\{X_T\}_{T\geq 0}$ satisfies*

$$|X_T| \leq h_1(T, \delta_1).$$

Moreover, suppose for any $\delta_2 \in (0,1)$, for all $T \geq T_2(\delta_2)$, with probability at least $1-\delta_2$, the random sequence $\{Y_T\}_{T\geq 0}$ satisfies

$$|Y_T| \leq h_2(T, \delta_2).$$

Then for any $\delta \in (0,1)$, for all $T \geq T_0(\delta) \coloneqq \max\{T_1(\frac{\delta}{2}), T_2(\frac{\delta}{2})\}$, with probability at least $1-\delta$, the random sequence $\{X_T + Y_T\}_{T\geq 0}$ satisfies

$$|X_T + Y_T| \leq h_1(T, \delta/2) + h_2(T, \delta/2).$$

Proof For a given $\delta \in (0,1)$, by the lemma's assumption, for all $T \geq T_1(\delta/2)$, we have

$$\mathbb{P}\Big(|X_T| > h_1(T,\delta/2)\Big) < \frac{\delta}{2}. \tag{59}$$

Similarly, we have that for all $T \geq T_2(\delta/2)$, we have

$$\mathbb{P}\Big(|Y_T| > h_2(T,\delta/2)\Big) < \frac{\delta}{2}. \tag{60}$$

Now $|X_T+Y_T| \geq h_1(T,\delta/2)+h_2(T,\delta/2)$ implies that $|X_T| > h_1(T,\delta/2)$ or $|Y_T| > h_2(T,\delta/2)$. As a result, by applying the union bound and (59)–(60), we get

$$\mathbb{P}\Big(|X_T + Y_T| \geq h_1(T,\delta/2) + h_2(T,\delta/2)\Big) \leq \delta.$$

■

C.1.3 Proof of Lemma 13

Proof of Part 1 Recall that for any policy $\pi \in \Pi_{\mathrm{AR}}$, the claim is the following chain of inequalities

$$\sigma_\pi(s) \overset{(a)}{\leq} K^\pi \overset{(b)}{\leq} H^\pi \overset{(c)}{\leq} \infty, \quad \forall s \in \mathcal{S}. \tag{61}$$

Proof of Part 1-(a): By the definition of K^π in Eq. (5), we have

$$\Big|V^\pi(S_+) - \mathbb{E}\big[V^\pi(S_+) \;\big|\; s,\pi(s)\big]\Big| \leq K^\pi, \quad \forall s \in \mathcal{S}, \quad a.s.$$

As a result, we have

$$\begin{aligned}
&\mathbb{E}\Big[\big(V^\pi(S_+) - \mathbb{E}[V^\pi(S_+) \mid s,\pi(s)]\big)^2 \mid s,\pi(s)\Big] \\
&= \sum_{s' \in \mathcal{S}} \big(V^\pi(s') - \mathbb{E}\big[V^\pi(S_+) \mid s,\pi(s))\big]\big)^2 P(s'|s,\pi(s)) \leq (K^\pi)^2, \quad \forall s \in \mathcal{S}.
\end{aligned}$$

Proof of Part 1-(b): By the definition of expectation operator, we have

$$\min_{s\in\mathcal{S}} V^\pi(s) \leq \mathbb{E}[V^\pi(S_+)|s,\pi(s)] \leq \max_{s\in\mathcal{S}} V^\pi(s).$$

As a result, we have

$$V^\pi(s) - \mathbb{E}[V^\pi(S_+)|s,\pi(s)] \leq V^\pi(s) - \min_{s\in\mathcal{S}} V^\pi(s) \leq \max_{s\in\mathcal{S}} V^\pi(s) - \min_{s\in\mathcal{S}} V^\pi(s), \quad \forall s \in \mathcal{S}. \tag{62}$$

Similarly, we have

$$\mathbb{E}[V^\pi(S_+)|s,\pi(s)] - V^\pi(s) \leq \max_{s\in\mathcal{S}} V^\pi(s) - V^\pi(s) \leq \max_{s\in\mathcal{S}} V^\pi(s) - \min_{s\in\mathcal{S}} V^\pi(s), \quad \forall s \in \mathcal{S}. \tag{63}$$

Therefore (62)–(63) imply that

$$\big|V^\pi(S_+) - \mathbb{E}[V^\pi(S_+) \;\big|\; s,\pi(s)]\big| \leq \mathrm{sp}(V^\pi) = H^\pi.$$

Proof of Part 1-(c): Since policy $\pi \in \Pi_{\mathrm{AR}}$, by (ARPE), we know $V^\pi : \mathcal{S} \to \mathbb{R}$ is a real-valued function and therefore, $H^\pi < \infty$.

Proof of Part 2 We prove that if $\mathcal{M}$ is communicating, then for any policy $\pi \in \Pi_{\mathrm{AR}}$, we have $H^\pi \le D^\pi R_{\max}$. Consider $s, s' \in \mathcal{S}$ where $s \neq s'$. By Puterman (2014), we have:

$$V^\pi(s) = \mathbb{E}\Big[\sum_{t=0}^{\infty} [r(S_t, A_t) - J^\pi] \;\Big|\; S_0 = s \Big]. \tag{64}$$

Now consider the stopping time τ_0 where $S = s'$ for the first time. We can rewrite $V^\pi(s)$ as follows

$$\begin{aligned}
V^\pi(s) &\overset{(a)}{=} \mathbb{E}\Big[\sum_{t=0}^{\tau_0 - 1} [r(S_t, A_t) - J^\pi] + \sum_{t=\tau_0}^{\infty} [r(S_t, A_t) - J^\pi] \;\Big|\; S_0 = s \Big]. \\
&\overset{(b)}{=} \mathbb{E}\Big[\sum_{t=0}^{\tau_0 - 1} [r(S_t, A_t) - J^\pi] \;\Big|\; S_0 = s \Big] + \mathbb{E}\Big[\sum_{t=\tau_0}^{\infty} [r(S_t, A_t) - J^\pi] \;\Big|\; S_0 = s \Big] \\
&\overset{(c)}{=} \mathbb{E}\Big[\sum_{t=0}^{\tau_0 - 1} [r(S_t, A_t) - J^\pi] \;\Big|\; S_0 = s \Big] + \mathbb{E}\Big[\sum_{t=\tau_0}^{\infty} [r(S_t, A_t) - J^\pi] \;\Big|\; S_{\tau_0} = s' \Big] \\
&\overset{(d)}{=} \mathbb{E}\Big[\sum_{t=0}^{\tau_0 - 1} [r(S_t, A_t) - J^\pi] \;\Big|\; S_0 = s \Big] + V^\pi(s'),
\end{aligned}$$

where (a) follows from splitting the summation with the stopping time τ_0; (b) follows from linearity of expectation and the fact that first and second term of RHS of (b) are finite; (c) follows from the strong Markov property and (d) follows from definition of $V^\pi(s')$. Therefore, we have

$$\begin{aligned}
V^\pi(s) - V^\pi(s') = \mathbb{E}\Big[\sum_{t=0}^{\tau_0 - 1} [r(S_t, A_t) - J^\pi] \Big] &\le \mathbb{E}\Big[\sum_{t=0}^{\tau_0 - 1} [r(S_t, A_t)] \Big] \\
&\overset{(e)}{\le} T^\pi(s, s') R_{\max} \overset{(f)}{\le} D^\pi R_{\max} \overset{(g)}{\le} D_w R_{\max} \overset{(h)}{<} \infty,
\end{aligned}$$

where (e) follows from the definition of $T^\pi(s, s')$, (f) follows from the definition of D^π, (g) follows from the definition of D_w, and (h) follows by the fact that $\mathcal{M}$ is communicating. Since one can repeat the same argument with any two pairs of (s, s'), it implies that $H^\pi \le D^\pi R_{\max} \le D_w R_{\max} < \infty$.

Proof of Part 3 The result of this part follows from Bartlett and Tewari (2012, Theorem 4), where it is shown that for weakly communicating $\mathcal{M}$, we have $H^* \le D R_{\max}$.

C.2 Proof of Theorem 15

C.2.1 Proof of Part 1

By Lemma 67, for any policy $\pi \in \Pi_{\mathrm{AR}}$, we can rewrite the cumulative reward R_T^π as follows

$$R_T^\pi = T J^\pi + \sum_{t=1}^{T} M_t^\pi + V^\pi(S_0) - V^\pi(S_T).$$

By (5) and Lemma 13, we have

$$\left|M_t^\pi\right| \leq K^\pi < \infty, \quad \forall t \geq 1.$$

Therefore

$$\sum_{t=1}^{\infty} \frac{(M_t^\pi)^2}{t^2} \leq K^\pi \sum_{t=1}^{\infty} \frac{1}{t^2} < \infty.$$

As a result by choosing $p = 2$ and $a_t = t$ in Theorem 58, we have

$$\lim_{T\to\infty} \frac{\sum_{t=1}^{T} M_t^\pi}{T} = 0, \quad a.s.$$

Furthermore, Lemma 13 implies that random variable $V^\pi(S_t)$ has bounded support, therefore,

$$\lim_{T\to\infty} \frac{V^\pi(S_0) - V^\pi(S_T)}{T} = 0, \quad a.s.$$

As a result, we have

$$\lim_{T\to\infty} \frac{R_T^\pi}{T} = \lim_{T\to\infty} \frac{\sum_{t=1}^{T} M_t^\pi + V^\pi(S_0) - V^\pi(S_T) + TJ^\pi}{T} = J^\pi, \quad a.s.$$

C.2.2 Proof of Part 2

To prove this part, we verify the conditions of Theorem 59 for the MDS $\{M_t^\pi\}_{t\geq 0}$. By Lemma 13, we have

$$\left|M_t^\pi\right| \leq K^\pi < \infty, \quad \forall t \geq 1.$$

As a result, the MDS $\{M_t^\pi\}_{t\geq 0}$ is a uniformly bounded MDS with respect to the constant K^π. By the theorem's assumption we have $\mathbb{P}(\Omega_0^\pi) = 1$, as a result,

$$\sum_{t=1}^{\infty} \mathbb{E}\Big[\left(M_t^\pi\right)^2 \mid \mathcal{F}_{t-1}\Big] = \infty, \quad a.s.$$

Therefore, for the stopping time $\{\nu_t\}_{t\geq 0}$ defined in Theorem 15, we have

$$\frac{\sum_{t=1}^{\nu_T} M_t^\pi}{\sqrt{T}} \xrightarrow{(d)} \mathcal{N}(0,1). \tag{65}$$

Since by Lemma 13, $V^\pi(S_t)$ has bounded support for all $t \geq 1$, we get

$$\frac{V^\pi(S_0) - V^\pi(S_T)}{\sqrt{T}} \to 0, \quad a.s. \tag{66}$$

By combining (65) and (66) and by using Theorem 78, we get

$$\lim_{T\to\infty} \frac{R_{\nu_T}^\pi(\omega) - \nu_T J^\pi}{\sqrt{T}} \xrightarrow{(d)} \mathcal{N}(0,1).$$

C.2.3 Proof of Part 3

We verify the conditions of Theorem 60 for the MDS $\{M_t^\pi\}_{t\geq 0}$. By Lemma 13, we have

$$\big|M_t^\pi\big| \leq K^\pi < \infty, \quad \forall t \geq 1.$$

As a result, MDS $\{M_t^\pi\}_{t\geq 0}$ is a uniformly bounded MDS with respect to the constant K^π. On the set Ω_0^π, we have

$$\sum_{t=1}^{\infty} \mathbb{E}\Big[\big(M_t^\pi\big)^2 \mid \mathcal{F}_{t-1}\Big] = \infty.$$

As a result, by using the definition of increasing process $\{\Sigma_t^\pi\}_{t\geq 0}$ and Theorem 60, we get

$$\liminf_{T\to\infty} \frac{\sum_{t=1}^T M_t^\pi}{\sqrt{2\Sigma_T^\pi \log\log \Sigma_T^\pi}} = -1, \quad \limsup_{T\to\infty} \frac{\sum_{t=1}^T M_t^\pi}{\sqrt{2\Sigma_T^\pi \log\log \Sigma_T^\pi}} = 1. \tag{67}$$

Since by Lemma 13, $V^\pi(S_t)$ has bounded support for all $t \geq 1$, we get

$$\lim_{T\to\infty} \frac{V^\pi(S_0) - V^\pi(S_T)}{\sqrt{2\Sigma_T^\pi \log\log \Sigma_T^\pi}} = 0, \quad \text{a.s.} \tag{68}$$

By combining (67) and (68), we get

$$\liminf_{T\to\infty} \frac{R_T^\pi(\omega) - TJ^\pi}{\sqrt{2\Sigma_T^\pi \log\log \Sigma_T^\pi}} = -1, \quad \limsup_{T\to\infty} \frac{R_T^\pi(\omega) - TJ^\pi}{\sqrt{2\Sigma_T^\pi \log\log \Sigma_T^\pi}} = 1.$$

C.3 Proof of Theorem 17

C.3.1 Proof of Part 1

By Lemma 67, for any policy $\pi \in \Pi_{\text{AR}}$, we can rewrite the cumulative reward $R_T^\pi(\omega)$ as follows

$$R_T^\pi(\omega) = TJ^\pi + \sum_{t=1}^T M_t^\pi + V^\pi(S_0) - V^\pi(S_T).$$

As a result, we have

$$\big|R_T^\pi(\omega) - TJ^\pi - \big(V^\pi(S_0) - V^\pi(S_T)\big)\big| = \big|\sum_{t=1}^T M_t^\pi\big|. \tag{69}$$

In order to upper-bound the term $\big|\sum_{t=1}^T M_t^\pi\big|$, we verify the conditions of Corollary 62. By (5) and Lemma 13, we have

$$\big|M_t^\pi\big| \leq K^\pi < \infty, \quad \forall t \geq 1.$$

As a result, MDS $\{M_t^\pi\}_{t\geq 1}$ is a uniformly bounded MDS with respect to the constant K^π. Therefore, Corollary 62 implies that for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\Big|\sum_{t=1}^T M_t^\pi\Big| \leq \sqrt{2T(K^\pi)^2 \log(\frac{2}{\delta})}. \tag{70}$$

By combining (69) and (70), with probability at least $1-\delta$, we have

$$\left|R_T^\pi(\omega) - TJ^\pi - \left(V^\pi(S_0) - V^\pi(S_T)\right)\right| \le K^\pi \sqrt{2T \log \frac{2}{\delta}}.$$

C.3.2 Proof of Part 2

Similar to the proof of Part 1, by lemma 67, we have

$$\left|R_T^\pi(\omega) - TJ^\pi - \left(V^\pi(S_0) - V^\pi(S_T)\right)\right| = \left|\sum_{t=1}^T M_t^\pi\right| \tag{71}$$

Moreover, MDS $\{M_t^\pi\}_{t\ge 0}$ is a uniformly bounded MDS with respect to the constant K^π. Therefore, Corollary 64 implies that for any $\delta \in (0,1)$, for all $T \ge T_0(\delta) \coloneqq \left\lceil \frac{173}{K^\pi} \log \frac{4}{\delta} \right\rceil$, with probability at least $1-\delta$, we have

$$\left|\sum_{t=1}^T M_t^\pi\right| \le \max\left\{K^\pi \sqrt{3T\left(2\log\log\frac{3T}{2} + \log\frac{2}{\delta}\right)}, (K^\pi)^2\right\}. \tag{72}$$

By combining (71) and (72), with probability at least $1-\delta$, we have

$$\left|R_T^\pi(\omega) - TJ^\pi - \left(V^\pi(S_0) - V^\pi(S_T)\right)\right| \le \max\left\{K^\pi \sqrt{3T\left(2\log\log\frac{3T}{2} + \log\frac{2}{\delta}\right)}, (K^\pi)^2\right\}.$$

C.4 Proof of Theorem 18

C.4.1 Proof of Part 1

By lemma 67, for any policy $\pi \in \Pi_{\mathrm{AR}}$, we can rewrite the cumulative reward $R_T^\pi(\omega)$ as follows

$$R_T^\pi(\omega) = TJ^\pi + \sum_{t=1}^T M_t^\pi + V^\pi(S_0) - V^\pi(S_T).$$

As a result, we have

$$\begin{aligned}
\left|R_T^\pi(\omega) - TJ^\pi\right| &= \left|\sum_{t=1}^T M_t^\pi + V^\pi(S_0) - V^\pi(S_T)\right| \\
&\overset{(a)}{\le} \left|\sum_{t=1}^T M_t^\pi\right| + \left|V^\pi(S_0) - V^\pi(S_T)\right| \\
&\overset{(b)}{\le} \left|\sum_{t=1}^T M_t^\pi\right| + H^\pi,
\end{aligned} \tag{73}$$

where (a) follows from the triangle inequality and (b) follows from the definition of H^π. In order to upper-bound the term $\left|\sum_{t=1}^T M_t^\pi\right|$, we verify the conditions of Corollary 62. By (5) and Lemma 13 we have

$$\left|M_t^\pi\right| \le K^\pi < \infty, \quad \forall t \ge 1.$$

As a result, MDS $\{M_t^\pi\}_{t\geq 1}$ is a uniformly bounded MDS with respect to the constant K^π. Therefore, Corollary 62 implies that for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\Big|\sum_{t=1}^{T} M_t^\pi\Big| \leq \sqrt{2T(K^\pi)^2 \log(\frac{2}{\delta})}. \tag{74}$$

By combining (73) and (74), with probability at least $1-\delta$, we have

$$|R_T^\pi(\omega) - TJ^\pi| \leq K^\pi \sqrt{2T \log \frac{2}{\delta}} + H^\pi.$$

C.4.2 Proof of Part 2

Similar to the proof of Part 1, by lemma 67, we have

$$\big|R_T^\pi(\omega) - TJ^\pi\big| \leq \Big|\sum_{t=1}^{T} M_t^\pi\Big| + H^\pi. \tag{75}$$

Moreover, MDS $\{M_t^\pi\}_{t\geq 0}$ is a uniformly bounded MDS with respect to the constant K^π. Therefore, Corollary 64 implies that for any $\delta \in (0,1)$, for all $T \geq T_0(\delta) \coloneqq \Big\lceil \frac{173}{K^\pi} \log \frac{4}{\delta} \Big\rceil$, with probability at least $1-\delta$, we have

$$\Big|\sum_{t=1}^{T} M_t^\pi\Big| \leq \max\Big\{K^\pi \sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{2}{\delta}\Big)}, (K^\pi)^2\Big\}. \tag{76}$$

By combining (75) and (76), with probability at least $1-\delta$, we have

$$\big|R_T^\pi(\omega) - TJ^\pi\big| \leq \max\Big\{K^\pi \sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{2}{\delta}\Big)}, (K^\pi)^2\Big\} + H^\pi.$$

C.5 Proof of Corollary 22

C.5.1 Proof of Part 1

Since $\mathcal{M}$ is communicating, by Lemma 13, Part 2, for any policy $\pi \in \Pi_{\mathrm{AR}}$, we have

$$\big|M_t^\pi\big| \leq K^\pi \leq D^\pi R_{\max}, \quad \forall t \geq 1. \tag{77}$$

As a result, the MDS $\{M_t^\pi\}_{t\geq 1}$ is a uniformly bounded MDS with respect to the constant $D^\pi R_{\max}$. Therefore, by repeating the arguments of the proof of Theorem 18, Part 1, and substituting H^π with $D^\pi R_{\max}$ in the RHS of (73) and replacing K^π with $D^\pi R_{\max}$ in the RHS of (74), with probability at least $1-\delta$, we have:

$$\big|R_T^\pi(\omega) - TJ^\pi\big| \leq D^\pi R_{\max} \sqrt{2T \log \frac{2}{\delta}} + D^\pi R_{\max}.$$

C.5.2 Proof of Part 2

Since $\mathcal{M}$ is communicating, by Lemma 13, Part 2 for any policy $\pi \in \Pi_{\mathrm{AR}}$, we have

$$\left|M_t^\pi\right| \le K^\pi \le D^\pi R_{\max}, \quad \forall t \ge 1. \tag{78}$$

As a result, the MDS $\{M_t^\pi\}_{t\ge 1}$ is a uniformly bounded MDS with respect to the constant $D^\pi R_{\max}$. Therefore, by repeating the arguments of the proof of Theorem 18, Part 2, and substituting H^π with $D^\pi R_{\max}$ in the RHS of (75) and substituting K^π with $D^\pi R_{\max}$ in the RHS of (76), we prove the claim, i.e, for any $\delta \in (0,1)$, for all $T \ge T_0(\delta) \coloneqq \left\lceil \frac{173}{D^\pi R_{\max}} \log \frac{4}{\delta} \right\rceil$, with probability at least $1-\delta$, we have

$$\left|R_T^\pi(\omega) - TJ^\pi\right| \le \max\left\{ D^\pi R_{\max} \sqrt{3T\Big(2\log\log \frac{3T}{2} + \log\frac{2}{\delta}\Big)}, (D^\pi R_{\max})^2 \right\} + D^\pi R_{\max}.$$

C.6 Proof of Corollary 23

By Prop. 8, Part 3, if $\mathcal{M}$ is communicating, it is also weakly communicating. In the case of weakly communicating $\mathcal{M}$, by Lemma 13, Part 3, for any optimal policy $\pi^* \in \Pi_{\mathrm{AR}}$, we have

$$\left|M_t^{\pi^*}\right| = \left|V^*(S_t) - \mathbb{E}\big[V^*(S_t) \mid S_{t-1}, \pi(S_{t-1})\big]\right| \le K^* \le DR_{\max}, \quad \forall t \ge 1. \tag{79}$$

As a result, the MDS $\{M_t^{\pi^*}\}_{t\ge 1}$ is uniformly bounded MDS with respect to the constant $DR_{\max}$. Therefore, by repeating the arguments of the proof of Corollary 22, Part 1 and Part 2 for the optimal policy $\pi^* \in \Pi_{\mathrm{AR}}$, we prove that $\left|R_T^{\pi^*}(\omega) - TJ^*\right|$ satisfies the non-asymptotic concentration rates in (18)–(19), where in the RHS, D^π is replaced with D.

C.7 Proof of Corollary 25

C.7.1 Proof of Part 1

Consider two policies $\pi_1, \pi_2 \in \Pi_{\mathrm{AR}}$. Then we have

$$\begin{aligned}\left|R_T^{\pi_1} - R_T^{\pi_2}\right| =& \left|R_T^{\pi_1} - TJ^{\pi_1} + TJ^{\pi_1} - TJ^{\pi_2} + TJ^{\pi_2} - R_T^{\pi_2}\right| \\ \overset{(a)}{\le}& \left|R_T^{\pi_1} - TJ^{\pi_1}\right| + \left|TJ^{\pi_1} - TJ^{\pi_2}\right| + \left|TJ^{\pi_2} - R_T^{\pi_2}\right|,\end{aligned} \tag{80}$$

where (a) follows from the triangle inequality. Similarly, we have

$$\begin{aligned}\left|TJ^{\pi_1} - TJ^{\pi_2}\right| =& \left|TJ^{\pi_1} - R_T^{\pi_1} + R_T^{\pi_1} - R_T^{\pi_2} + R_T^{\pi_2} - TJ^{\pi_2}\right| \\ \overset{(b)}{\le}& \left|TJ^{\pi_1} - R_T^{\pi_1}\right| + \left|R_T^{\pi_1} - R_T^{\pi_2}\right| + \left|R_T^{\pi_2} - TJ^{\pi_2}\right|,\end{aligned} \tag{81}$$

where (b) follows from the triangle inequality. (80)–(81) imply that

$$\left|\left|R_T^{\pi_1} - R_T^{\pi_2}\right| - \left|TJ^{\pi_1} - TJ^{\pi_2}\right|\right| \le \left|R_T^{\pi_1} - TJ^{\pi_1}\right| + \left|R_T^{\pi_2} - TJ^{\pi_2}\right|. \tag{82}$$

By Theorem 18, we know that for any $\delta_1 \in (0,1)$, with probability at least $1-\delta_1$, we have

$$\left|R_T^{\pi_1} - TJ^{\pi_1}\right| \leq K^{\pi_1}\sqrt{2T\log\frac{2}{\delta_1}} + H^{\pi_1}.$$

Similarly, we have that for any $\delta_2 \in (0,1)$, with probability at least $1-\delta_2$, we have

$$\left|R_T^{\pi_2} - TJ^{\pi_2}\right| \leq K^{\pi_2}\sqrt{2T\log\frac{2}{\delta_2}} + H^{\pi_2}.$$

As a result, by applying Lemma 68 and (82), for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\begin{aligned}\left|\left|R_T^{\pi_1} - R_T^{\pi_2}\right| - \left|TJ^{\pi_1} - TJ^{\pi_2}\right|\right| &\leq \left|R_T^{\pi_1} - TJ^{\pi_1}\right| + \left|TJ^{\pi_1} - TJ^{\pi_2}\right| \\ &\leq K^{\pi_1}\sqrt{2T\log\frac{4}{\delta}} + H^{\pi_1} + K^{\pi_2}\sqrt{2T\log\frac{4}{\delta}} + H^{\pi_2}.\end{aligned}$$

C.7.2 Proof of Part 2

As we showed in the proof of part 1, for any two policies $\pi_1, \pi_2 \in \Pi_{\mathrm{AR}}$, we have

$$\left|\left|R_T^{\pi_1} - R_T^{\pi_2}\right| - \left|TJ^{\pi_1} - TJ^{\pi_2}\right|\right| \leq \left|R_T^{\pi_1} - TJ^{\pi_1}\right| + \left|R_T^{\pi_2} - TJ^{\pi_2}\right|.$$

By Theorem 18, we have that for any $\delta_1 \in (0,1)$, for all $T \geq T_0^{\pi_1}(\delta) \coloneqq \left\lceil \frac{173}{K^{\pi_1}}\log\frac{4}{\delta_1}\right\rceil$, with probability at least $1-\delta_1$, we have

$$\left|R_T^{\pi_1} - TJ^{\pi_1}\right| \leq \max\left\{K^{\pi_1}\sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{2}{\delta_1}\Big)}, (K^{\pi_1})^2\right\} + H^{\pi_1}.$$

Similarly, we have that for any $\delta_2 \in (0,1)$, for all $T \geq T_0^{\pi_2}(\delta) \coloneqq \left\lceil \frac{173}{K^{\pi_2}}\log\frac{4}{\delta_2}\right\rceil$, with probability at least $1-\delta_2$, we have

$$\left|R_T^{\pi_2} - TJ^{\pi_2}\right| \leq \max\left\{K^{\pi_2}\sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{2}{\delta_2}\Big)}, (K^{\pi_2})^2\right\} + H^{\pi_2}.$$

As a result, by applying Lemma 68 and (82), for all $T \geq T_0(\delta) \coloneqq \max\left\{\left\lceil \frac{173}{K^{\pi_1}}\log\frac{8}{\delta}\right\rceil, \left\lceil \frac{173}{K^{\pi_2}}\log\frac{8}{\delta}\right\rceil\right\}$, with probability at least $1-\delta$, we have

$$\begin{aligned}\left|\left|R_T^{\pi_1} - R_T^{\pi_2}\right| - \left|TJ^{\pi_1} - TJ^{\pi_2}\right|\right| &\leq \left|R_T^{\pi_1} - TJ^{\pi_1}\right| + \left|TJ^{\pi_2} - R_T^{\pi_2}\right| \\ &\leq \max\left\{K^{\pi_1}\sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{4}{\delta}\Big)}, (K^{\pi_1})^2\right\} + H^{\pi_1} \\ &\quad + \max\left\{K^{\pi_2}\sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{4}{\delta}\Big)}, (K^{\pi_2})^2\right\} + H^{\pi_2}.\end{aligned}$$

C.8 Proof of Theorem 29

By Corollary 16, for any optimal policy $\pi^* \in \Pi_{\mathrm{AR}}$, the quantity R^{π^*} satisfies the asymptotic concentration rates in (11)–(13). On the other hand, by (3), for any learning policy μ, we have

$$\mathcal{D}_T(\omega) = R_T^{\pi^*} - TJ^*.$$

As a result, by substituting $\mathcal{D}_T(\omega)$ in the LHS of (11)–(13), we get that for any learning policy μ, these asymptotic concentration rates also hold for the difference $\mathcal{D}_T(\omega)$ of cumulative regret and interim cumulative regret.

C.9 Proof of Theorem 30

By Corollary 19, for any optimal policy $\pi^* \in \Pi_{\mathrm{AR}}$, the quantity $|R_T^{\pi^*} - TJ^*|$ satisfies the asymptotic concentration rates in (16)–(17). On the other hand, by (3), for any learning policy μ, we have

$$\mathcal{D}_T(\omega) = R_T^{\pi^*} - TJ^*.$$

As a result, by substituting $\mathcal{D}_T(\omega)$ in the LHS of (16)–(17), we get that for any learning policy μ, these non-asymptotic concentration rates also hold for the difference $\mathcal{D}_T(\omega)$ of cumulative regret and interim cumulative regret.

C.10 Proof of Corollary 31

By Corollary 23, for the weakly communicating $\mathcal{M}$, for any optimal policy $\pi^* \in \Pi_{\mathrm{AR}}$, the quantity $|R_T^{\pi^*} - TJ^*|$ satisfies the non-asymptotic concentration rates in (18)–(19), where in the RHS, D^π is replaced with D. On the other hand, by (3), for any learning policy μ, we have

$$\mathcal{D}_T(\omega) = R_T^{\pi^*} - TJ^*.$$

As a result, by substituting $\mathcal{D}_T(\omega)$ in the LHS of (18)–(19), we get that for the weakly communicating $\mathcal{M}$, for any learning policy μ, these non-asymptotic concentration rates also hold for the difference $\mathcal{D}_T(\omega)$ of cumulative regret and interim cumulative regret. At last by Prop. 8, we have that if $\mathcal{M}$ is recurrent, unichain, or communicating it is also weakly communication. As a result, these non-asymptotic concentration bounds hold for all the cases.

C.11 Proof of Theorem 33

C.11.1 Proof of Part 1

This part of the theorem is a consequence of Theorem 29. Recall that by definition, we have

$$\mathcal{D}_T(\omega) = \mathcal{R}_T^\mu(\omega) - \bar{\mathcal{R}}_T^\mu(\omega). \tag{83}$$

On the other hand, we can rewrite the law of iterated logarithm in Theorem 29 using the $\tilde{\mathcal{O}}(\cdot)$ notation as follows

$$\mathcal{D}_T(\omega) \leq \tilde{\mathcal{O}}(\sqrt{T}), \quad a.s. \tag{84}$$

As a result, for any learning policy μ that satisfies $R_T^\mu(\omega) \leq \tilde{\mathcal{O}}(\sqrt{T})$, almost surely, (83)–(84) imply that $\bar{R}_T^\mu(\omega) \leq \tilde{\mathcal{O}}(\sqrt{T})$. Similarly, for any learning policy μ that satisfies $\bar{R}_T^\mu(\omega) \leq$

$\tilde{\mathcal{O}}(\sqrt{T})$, almost surely, (83)–(84) imply that $R_T^\mu(\omega) \leq \tilde{\mathcal{O}}(\sqrt{T})$. Therefore, statements 1 and 2 are equivalent.

C.11.2 Proof of Part 2

Proof of this part is a consequence of Theorem 30. By the theorem's hypothesis, for any $\delta_1 \in (0,1)$, there exists a pair of functions $(T_1(\delta_1), h_1(\delta_1, T))$, such that for all $T \geq T_1(\delta_1)$, with probability at least $1-\delta_1$, we have

$$R_T^\mu(\omega) \leq h_1(\delta_1, T), \tag{85}$$

where for a fixed δ_1, we have $h_1(\delta_1, T) = \tilde{\mathcal{O}}(\sqrt{T})$. Moreover, by Theorem 30, we have that for any $\delta_2 \in (0,1)$, there exists a pair of functions $(T_2(\delta_2), h_2(\delta_2, T))$, such that for all $T \geq T_2(\delta_2)$, with probability at least $1-\delta_2$, we have

$$\mathcal{D}_T(\omega) \leq h_2(\delta_2, T), \tag{86}$$

where for a fixed δ_2, we have $h_2(\delta_2, T) = \tilde{\mathcal{O}}(\sqrt{T})$. As a result, by (83), (85)–(86), and Lemma 68, for any $\delta \in (0,1)$, for all $T \geq \max\big\{T_1(\delta/2), T_2(\delta/2)\big\}$, with probability at least $1-\delta$, we have

$$\bar{R}_T^\mu(\omega) \leq h_1(\delta/2) + h_2(\delta/2).$$

At last since for a fixed δ, both $h_1(\delta/2)$ and $h_2(\delta/2)$ satisfy

$$h_1(\delta/2) \leq \tilde{\mathcal{O}}(\sqrt{T}), \quad \text{and,} \quad h_2(\delta/2) \leq \tilde{\mathcal{O}}(\sqrt{T}),$$

we get that $\bar{R}_T^\mu(\omega) \leq \tilde{\mathcal{O}}(\sqrt{T})$. With repeating the similar arguments we can prove the 2^{nd} statement.

Appendix D. Proof of Main Results for Discounted Reward Setup

D.1 Proof of Theorem 37

D.1.1 Preliminary Results

We first present a few preliminary lemmas. To simplify the notation, we define following martingale difference sequence.

Definition 69 *Let filtration $\mathcal{F} = \{\mathcal{F}_t\}_{t\geq 0}$ be defined as $\mathcal{F}_t \coloneqq \sigma(S_{0:t}, A_{0:t})$. For any policy $\pi \in \Pi_{\mathrm{SD}}$, let V_γ^π denote the corresponding discounted value function. We define the sequence $\{N_t^{\pi,\gamma}\}_{t\geq 1}$ as follows*

$$N_t^{\pi,\gamma} \coloneqq \Big[V_\gamma^\pi(S_t) - \mathbb{E}\big[V_\gamma^\pi(S_t) \;\big|\; S_{t-1}, \pi(S_{t-1})\big]\Big], \quad \forall t \geq 1, \tag{87}$$

where $\{S_t\}_{t\geq 1}$ denotes the random sequence of states encountered along the current sample path.

Lemma 70 *Sequence $\{\gamma^t N_t^{\pi,\gamma}\}_{t\geq 1}$ is an MDS.*

Proof By the definition of $\{\mathcal{F}_t\}_{t\geq 0}$, we have that S_{t-1} is $\mathcal{F}_{t-1}$-measurable. As a result, we have

$$\begin{aligned}\mathbb{E}\Big[\gamma^t N_t^{\pi,\gamma} \;\Big|\; \mathcal{F}_{t-1}\Big] &= \mathbb{E}\Big[\gamma^t\big(V_\gamma^\pi(S_t) - \mathbb{E}\big[V_\gamma^\pi(S_t) \;\big|\; S_{t-1}, \pi(S_{t-1})\big]\big) \;\Big|\; \mathcal{F}_{t-1}\Big] \\ &= \gamma^t \mathbb{E}\Big[V_\gamma^\pi(S_t) \;\Big|\; \mathcal{F}_{t-1}\Big] - \gamma^t \mathbb{E}\Big[V_\gamma^\pi(S_t) \;\Big|\; S_{t-1}, \pi(S_{t-1})\Big] = 0,\end{aligned}$$

which shows that $\{\gamma_t N_t^{\pi,\gamma}\}_{t\geq 0}$ is an MDS with respect to the filtration $\{\mathcal{F}_t\}_{t\geq 0}$. ■

We now present a martingale decomposition for the discounted cumulative reward $R_T^{\pi,\gamma}(\omega)$ for any policy $\pi \in \Pi_{\text{SD}}$.

Lemma 71 *Given any policy $\pi \in \Pi_{\text{SD}}$, we can rewrite the discounted cumulative reward $R_T^{\pi,\gamma}$ as follows*

$$R_T^{\pi,\gamma}(\omega) = \sum_{t=1}^{T} \gamma^t N_t^{\pi,\gamma} + V_\gamma^\pi(S_0) - \gamma^T V_\gamma^\pi(S_T). \tag{88}$$

Proof Since $\pi \in \Pi_{\text{SD}}$, (DRPE) implies that along the trajectory of states $\{S_t\}_{t=0}^T$ induced by the policy π, we have

$$r\big(S_t, \pi(S_t)\big) = V_\gamma^\pi(S_t) - \gamma \mathbb{E}\Big[V_\gamma^\pi(S_{t+1}) \;\Big|\; S_t, \pi(S_t)\Big].$$

Repeating similar steps as in the proof of Lemma 67, we have

$$\begin{aligned}R_T^{\pi,\gamma}(\omega) &= \sum_{t=0}^{T-1} \gamma^t r(S_t, \pi(S_t)) \\ &= \sum_{t=0}^{T-1} \gamma^t \Big[V_\gamma^\pi(S_t) - \gamma \mathbb{E}\big[V_\gamma^\pi(S_{t+1}) | S_t, \pi(S_t)\big]\Big] \\ &\overset{(a)}{=} \sum_{t=0}^{T-1} \gamma^t \Big[V_\gamma^\pi(S_t) - \gamma \mathbb{E}\big[V_\gamma^\pi(S_{t+1}) | S_t, \pi(S_t)\big]\Big] + \gamma^T V_\gamma^\pi(S_T) - \gamma^T V_\gamma^\pi(S_T) \\ &\overset{(b)}{=} \sum_{t=0}^{T-1} \gamma^{t+1} \Big[V_\gamma^\pi(S_{t+1}) - \mathbb{E}\big[V_\gamma^\pi(S_{t+1}) | S_t, \pi(S_t)\big]\Big] + V_\gamma^\pi(S_0) - \gamma^T V_\gamma^\pi(S_T) \\ &\overset{(c)}{=} \sum_{t=0}^{T-1} \gamma^{t+1} N_{t+1}^{\pi,\gamma} + V_\gamma^\pi(S_0) - \gamma^T V_\gamma^\pi(S_T) \\ &= \sum_{t=1}^{T} \gamma^t N_t^{\pi,\gamma} + V_\gamma^\pi(S_0) - \gamma^T V_\gamma^\pi(S_T),\end{aligned}$$

where (a) follows from adding and subtracting the term $\gamma^T V_\gamma^\pi(S_T)$, (b) follows from re-arranging the terms in the summation, and (c) follows from the definition of $\{N_t^{\pi,\gamma}\}_{t\geq 1}$. ■

D.1.2 Proof of Theorem 37

Proof of this theorem follows from the martingale decomposition stated in Lemma 71 and the concentration bounds stated in Corollary 62 and Theorem 63.

Proof of Part 1 By Lemma 71, we have

$$R_T^{\pi,\gamma}(\omega) = \sum_{t=1}^{T} \gamma^t N_t^{\pi,\gamma} + V_\gamma^\pi(S_0) - \gamma^T V_\gamma^\pi(S_T).$$

As a result, we have

$$\Big| R_T^{\pi,\gamma}(\omega) - \big(V_\gamma^\pi(S_0) - \gamma^T V_\gamma^\pi(S_T)\big)\Big| = \big|\sum_{t=1}^{T} \gamma^t N_t^{\pi,\gamma}\big|. \tag{89}$$

In order to upper-bound the term $\big|\sum_{t=1}^{T} \gamma^t N_t^{\pi,\gamma}\big|$, we verify the conditions of Corollary 62. By (26) and Lemma 13, we have

$$\big|\gamma^t N_t^{\pi,\gamma}\big| \le \gamma^t K^{\pi,\gamma} < \infty, \quad \forall t \ge 1.$$

As a result, MDS $\{\gamma^t N_t^{\pi,\gamma}\}_{t\ge1}$ is a sequentially bounded MDS with respect to the sequence $\{\gamma^t K^{\pi,\gamma}\}_{t\ge1}$. Therefore, Corollary 62 implies that for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\begin{aligned}
\Big|\sum_{t=1}^{T} \gamma^t N_t^{\pi,\gamma}\Big| &\le \sqrt{2\sum_{t=1}^{T} (K^{\pi,\gamma})^2 \gamma^{2t} \log\frac{2}{\delta}} \\
&= K^{\pi,\gamma}\sqrt{2\sum_{t=1}^{T} \gamma^{2t} \log\frac{2}{\delta}} \\
&= K^{\pi,\gamma}\sqrt{2 f^\gamma(T) \log\frac{2}{\delta}}.
\end{aligned} \tag{90}$$

As a result, by combining (89) and (90), with probability at least $1-\delta$, we have

$$\Big| R_T^{\pi,\gamma}(\omega) - \big(V_\gamma^\pi(S_0) - \gamma^T V_\gamma^\pi(S_T)\big)\Big| \le K^{\pi,\gamma}\sqrt{2 f^\gamma(T) \log\frac{2}{\delta}}. \tag{91}$$

Proof of Part 2: Similar to the proof of Part 1, by Lemma 71, we have

$$\Big| R_T^{\pi,\gamma}(\omega) - \big(V_\gamma^\pi(S_0) - \gamma^T V_\gamma^\pi(S_T)\big)\Big| = \big|\sum_{t=1}^{T} \gamma^t N_t^{\pi,\gamma}\big|. \tag{92}$$

Moreover, MDS $\{\gamma^t N_t^{\pi,\gamma}\}_{t\ge1}$ is a sequentially bounded MDS with respect to the sequence $\{\gamma^t K^{\pi,\gamma}\}_{t\ge1}$. Therefore, Theorem 63 implies that for any $\delta \in (0,1)$, for all $T \ge T_0(\delta) := \min\big\{T \ge 1 : f^\gamma(T) > \dfrac{173}{K^{\pi,\gamma}} \log\dfrac{4}{\delta}\big\}$, with probability at least $1-\delta$, we have

$$\Big|\sum_{t=1}^{T} \gamma^t N_t^{\pi,\gamma}\Big| \le \sqrt{3\Big(\sum_{t=1}^{T} (K^{\pi,\gamma})^2 (\gamma^t)^2\Big)\Big(2\log\log\big(\frac{3\sum_{t=1}^{T}(K^{\pi,\gamma})^2(\gamma^t)^2}{2\big|\sum_{t=1}^{T}\gamma^t N_t^{\pi,\gamma}\big|}\big) + \log\frac{2}{\delta}\Big)}.$$

Now there are two cases: either $|\sum_{t=1}^{T}\gamma^t N_t^{\pi,\gamma}| \le (K^{\pi,\gamma})^2$ or $|\sum_{t=1}^{T}\gamma^t N_t^{\pi,\gamma}| \ge (K^{\pi,\gamma})^2$. If $|\sum_{t=1}^{T}\gamma^t N_t^{\pi,\gamma}| \ge (K^{\pi,\gamma})^2$, we get:

$$\begin{aligned}\Big|\sum_{t=1}^{T}\gamma^t N_t^{\pi,\gamma}\Big| &\le \sqrt{3\Big(\sum_{t=1}^{T}(K^{\pi,\gamma})^2(\gamma^t)^2\Big)\Big(2\log\log\big(\frac{3\sum_{t=1}^{T}(K^{\pi,\gamma})^2(\gamma^t)^2}{2\big|\sum_{t=1}^{T}\gamma^t N_t^{\pi,\gamma}\big|}\big)+\log\frac{2}{\delta}\Big)}\\ &\le \sqrt{3\Big(\sum_{t=1}^{T}(K^{\pi,\gamma})^2(\gamma^t)^2\Big)\Big(2\log\log\big(\frac{3\sum_{t=1}^{T}(K^{\pi,\gamma})^2(\gamma^t)^2}{2(K^{\pi,\gamma})^2}\big)+\log\frac{2}{\delta}\Big)}\\ &\overset{(a)}{=} K^{\pi,\gamma}\sqrt{3f^{\gamma}(T)\Big(2\log\log\big(\frac{3}{2}f^{\gamma}(T)\big)+\log\frac{2}{\delta}\Big)},\end{aligned}$$

where (a) follows from the geometric series formula and the definition of $f^{\gamma}(T)$. Otherwise, we have $|\sum_{t=1}^{T}\gamma^t N_t^{\pi,\gamma}| \le (K^{\pi,\gamma})^2$. As a result, we can summarize these two cases as follows

$$\Big|\sum_{t=1}^{T}\gamma^t N_t^{\pi,\gamma}\Big| \le \max\Big\{K^{\pi,\gamma}\sqrt{3f^{\gamma}(T)\big(2\log\log\frac{3}{2}f^{\gamma}(T)+\log\frac{2}{\delta}\big)},(K^{\pi,\gamma})^2\Big\}. \tag{93}$$

By combining (92)–(93), with probability at least $1-\delta$, we have

$$\begin{aligned}&\Big|R_T^{\pi,\gamma}(\omega)-\big(V_{\gamma}^{\pi}(S_0)-\gamma^T V_{\gamma}^{\pi}(S_T)\big)\Big|\\ &\qquad\le \max\Big\{K^{\pi,\gamma}\sqrt{3f^{\gamma}(T)\big(2\log\log(\frac{3}{2}f^{\gamma}(T))+\log\frac{2}{\delta}\big)},(K^{\pi,\gamma})^2\Big\}.\end{aligned} \tag{94}$$

D.2 Proof of Corollary 38

Proof of Part 1: By Lemma 71, we have

$$R_T^{\pi,\gamma}(\omega)=\sum_{t=1}^{T}\gamma^t N_t^{\pi,\gamma}+V_{\gamma}^{\pi}(S_0)-\gamma^T V_{\gamma}^{\pi}(S_T).$$

As a result, we have

$$\Big|R_T^{\pi,\gamma}(\omega)-V_{\gamma}^{\pi}(S_0)\Big| \overset{(a)}{\le} \Big|\sum_{t=1}^{T}\gamma^t N_t^{\pi,\gamma}\Big|+\Big|\gamma^T V_{\gamma}^{\pi}(S_T)\Big|, \tag{95}$$

where (a) follows from the triangle inequality. In the proof of Theorem 37, Part 1, we showed that with probability at least $1-\delta$, we have

$$\Big|\sum_{t=1}^{T}\gamma^t N_t^{\pi,\gamma}\Big| \le K^{\pi,\gamma}\sqrt{2f^{\gamma}(T)\log\frac{2}{\delta}}. \tag{96}$$

Moreover, we have

$$\begin{aligned}\gamma^T V_{\gamma}^{\pi}(S_T)&=\gamma^T\mathbb{E}^{\pi}\Big[\lim_{T\to\infty}\sum_{t=0}^{T-1}\gamma^t r(S_t,A_t)\;\Big|\;S_0=S_T\Big]\\ &=\gamma^T\mathbb{E}^{\pi}\Big[\lim_{T\to\infty}\sum_{t=0}^{T-1}\gamma^t R_{\max}\;\Big|\;S_0=S_T\Big]\le\frac{\gamma^T}{1-\gamma}R_{\max}.\end{aligned} \tag{97}$$

By combining (95)–(97), with probability $1-\delta$, we have

$$\Big|R_T^{\pi,\gamma}(\omega) - V_\gamma^\pi(S_0)\Big| \le K^{\pi,\gamma}\sqrt{2f^\gamma(T)\log\frac{2}{\delta}} + \frac{\gamma^T}{1-\gamma}R_{\max}.$$

Proof of Part 2: Similar to the proof of Part 1, by Lemma 71, we have

$$\Big|R_T^{\pi,\gamma}(\omega) - V_\gamma^\pi(S_0)\Big| \le \Big|\sum_{t=1}^{T}\gamma^t N_t^{\pi,\gamma}\Big| + \Big|\gamma^T V_\gamma^\pi(S_T)\Big|. \tag{98}$$

Moreover, we have

$$\Big|\gamma^T V_\gamma^\pi(S_T)\Big| \le \gamma^T\frac{R_{\max}}{1-\gamma}. \tag{99}$$

In addition, from proof of Theorem 37, Part 2, we have for any $\delta\in(0,1)$, for all $T \ge T_0(\delta) \coloneqq \min\big\{T'\ge 1 : f^\gamma(T') > \frac{173}{K^{\pi,\gamma}}\log\frac{4}{\delta}\big\}$, with probability at least $1-\delta$, we have

$$\Big|\sum_{t=1}^{T}\gamma^t N_t^{\pi,\gamma}\Big| \le \max\Big\{K^{\pi,\gamma}\sqrt{3f^\gamma(T)\big(2\log\log\frac{3}{2}f^\gamma(T)+\log\frac{2}{\delta}\big)}, (K^{\pi,\gamma})^2\Big\}. \tag{100}$$

By combining (98)–(100), with probability at least $1-\delta$, we have

$$\begin{aligned}&\Big|R_T^{\pi,\gamma}(\omega) - V_\gamma^\pi(S_0)\Big| \\ &\qquad\le \max\Big\{K^{\pi,\gamma}\sqrt{3f^\gamma(T)\big(2\log\log(\frac{3}{2}f^\gamma(T))+\log\frac{2}{\delta}\big)}, (K^{\pi,\gamma})^2\Big\} + \frac{\gamma^T}{1-\gamma}R_{\max}.\end{aligned} \tag{101}$$

D.3 Proof of Corollary 40

D.3.1 Proof of Part 1

Consider two policies $\pi_1,\pi_2\in\Pi_{\mathrm{SD}}$. Let $\{S_t^{\pi_1}\}_{t\ge0}$ and $\{S_t^{\pi_2}\}_{t\ge0}$ denote the random sequences of states encountered by following policies π_1 and π_2. We have

$$\begin{aligned}\Big|R_T^{\pi_1,\gamma} - R_T^{\pi_2,\gamma}\Big| &\overset{(a)}{=} \Big|R_T^{\pi_1,\gamma} - \big[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\big] + \big[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\big] \\ &\quad - \big[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\big] + \big[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\big] - R_T^{\pi_2,\gamma}\Big| \\ &\overset{(b)}{\le} \Big|R_T^{\pi_1,\gamma} - \big[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\big]\Big| + \Big|\big[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\big] - R_T^{\pi_2,\gamma}\Big| \\ &\quad + \Big|\big[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\big] - \big[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\big]\Big|,\end{aligned} \tag{102}$$

where (a) follows by adding and subtracting $\left[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\right]$ and $\left[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\right]$ and (b) follows from the triangle inequality. Similarly, we have

$$\begin{aligned}
&\Big|\left[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\right] - \left[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\right]\Big| \overset{(a)}{=} \\
&\Big|\left[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\right] - R_T^{\pi_1,\gamma} + R_T^{\pi_1,\gamma} - R_T^{\pi_2,\gamma} + R_T^{\pi_2,\gamma} - \left[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\right]\Big| \\
\overset{(b)}{\leq}&\Big|R_T^{\pi_1,\gamma} - \left[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\right]\Big| + \Big|R_T^{\pi_2,\gamma} - \left[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\right]\Big| \\
&+\Big|R_T^{\pi_1,\gamma} - R_T^{\pi_2,\gamma}\Big|, \qquad (103)
\end{aligned}$$

where (a) follows by adding and subtracting $R_T^{\pi_1,\gamma}$ and $R_T^{\pi_2,\gamma}$ and (b) follows from the triangle inequality. (102)–(103) imply that

$$\begin{aligned}
&\Big|\big|R_T^{\pi_1,\gamma} - R_T^{\pi_2,\gamma}\big| - \big|\left[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\right] - \left[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\right]\big|\Big| \\
\leq&\Big|R_T^{\pi_1,\gamma} - \left[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\right]\Big| + \Big|R_T^{\pi_2,\gamma} - \left[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\right]\Big|. \qquad (104)
\end{aligned}$$

By Theorem 37, we know that for any $\delta_1 \in (0,1)$, with probability at least $1-\delta_1$, we have

$$\Big|R_T^{\pi_1,\gamma} - \left[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\right]\Big| \leq K^{\pi_1,\gamma}\sqrt{2f^\gamma(T)\log\frac{2}{\delta_1}}. \qquad (105)$$

Similarly, we have that for any $\delta_2 \in (0,1)$, with probability at least $1-\delta_2$, we have

$$\Big|R_T^{\pi_2,\gamma} - \left[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\right]\Big| \leq K^{\pi_2,\gamma}\sqrt{2f^\gamma(T)\log\frac{2}{\delta_2}}. \qquad (106)$$

As a result, by applying Lemma 68 and (104)–(106), for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\Big|\big|R_T^{\pi_1,\gamma} - R_T^{\pi_2,\gamma}\big| - \big|\left[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\right] - \left[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\right]\big|\Big| \\
&\leq \Big|R_T^{\pi_1,\gamma} - \left[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\right]\Big| + \Big|R_T^{\pi_2,\gamma} - \left[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\right]\Big| \\
&\leq K^{\pi_1,\gamma}\sqrt{2f^\gamma(T)\log\frac{4}{\delta}} + K^{\pi_2,\gamma}\sqrt{2f^\gamma(T)\log\frac{4}{\delta}}.
\end{aligned}$$

D.3.2 Proof of Part 2

As we showed in the proof of part 1, for any two policies $\pi_1, \pi_2 \in \Pi_{\text{SD}}$, we have

$$\begin{aligned}
&\Big|\big|R_T^{\pi_1,\gamma} - R_T^{\pi_2,\gamma}\big| - \big|\left[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\right] - \left[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\right]\big|\Big| \\
\leq&\Big|R_T^{\pi_1,\gamma} - \left[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\right]\Big| + \Big|R_T^{\pi_2,\gamma} - \left[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\right]\Big|. \qquad (107)
\end{aligned}$$

By Theorem 37, we have that for any $\delta_1 \in (0,1)$, for all $T \geq T_0^{\pi_1}(\delta_1) \coloneqq \min\big\{T' \geq 1 : f^\gamma(T') > \frac{173}{K^{\pi_1,\gamma}} \log \frac{4}{\delta_1}\big\}$, with probability at least $1-\delta_1$, we have:

$$\Big|R_T^{\pi_1,\gamma} - \big[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\big]\Big| \leq \max\Big\{K^{\pi_1,\gamma}\sqrt{3f^\gamma(T)\Big(2\log\log\frac{3}{2}f^\gamma(T) + \log\frac{2}{\delta_1}\Big)}, (K^{\pi_1,\gamma})^2\Big\}.$$

Similarly, we have that for any $\delta_2 \in (0,1)$, for all $T \geq T_0^{\pi_2}(\delta_2) \coloneqq \min\big\{T' \geq 1 : f^\gamma(T') > \frac{173}{K^{\pi_2,\gamma}} \log \frac{4}{\delta_2}\big\}$, with probability at least $1-\delta_2$, we have:

$$\Big|R_T^{\pi_2,\gamma} - \big[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\big]\Big| \leq \max\Big\{K^{\pi_2,\gamma}\sqrt{3f^\gamma(T)\Big(2\log\log\frac{3}{2}f^\gamma(T) + \log\frac{2}{\delta_2}\Big)}, (K^{\pi_2,\gamma})^2\Big\}.$$

As a result, by applying Lemma 68, for all $T \geq T_0^{\pi}(\delta) \coloneqq \max\big\{T_0^{\pi_1}(\frac{\delta}{2}), T_0^{\pi_2}(\frac{\delta}{2})\big\}$, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\Big|\big|R_T^{\pi_1,\gamma} - R_T^{\pi_2,\gamma}\big| - \big|\big[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\big] - \big[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\big]\big|\Big| \\
\leq &\Big|R_T^{\pi_1,\gamma} - \big[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\big]\Big| + \Big|R_T^{\pi_2,\gamma} - \big[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\big]\Big| \\
\leq &\max\Big\{K^{\pi_1,\gamma}\sqrt{3f^\gamma(T)\Big(2\log\log\frac{3}{2}f^\gamma(T) + \log\frac{4}{\delta}\Big)}, (K^{\pi_1,\gamma})^2\Big\} \\
&+ \max\Big\{K^{\pi_2,\gamma}\sqrt{3f^\gamma(T)\Big(2\log\log\frac{3}{2}f^\gamma(T) + \log\frac{4}{\delta}\Big)}, (K^{\pi_2,\gamma})^2\Big\}.
\end{aligned}$$

D.4 Proof of Corollary 42

Since policy $\pi \in \Pi_{\mathrm{AR}}$, we know the pair (J^π, V^π) exists and J^π is constant for all $s \in \mathcal{S}$. We first prove following preliminary lemma.

D.4.1 Preliminary Lemma

Lemma 72 *For any policy $\pi \in \Pi_{\mathrm{AR}}$, as γ goes to 1 from below, following statements hold.*

1. *For any two states $s_1, s_2 \in \mathcal{S}$, we have*

$$\lim_{\gamma\uparrow 1}\Big[V_\gamma^\pi(s_1) - V_\gamma^\pi(s_2)\Big] = V^\pi(s_1) - V^\pi(s_2).$$

2. *For any two states $s_1, s_2 \in \mathcal{S}$, we have*

$$\lim_{\gamma\uparrow 1}\Big[V_\gamma^\pi(s_1) - \gamma^T V_\gamma^\pi(s_2)\Big] = TJ^\pi + V^\pi(s_1) - V^\pi(s_2).$$

3. *We have*

$$\lim_{\gamma\uparrow 1} f(T,\gamma) = T. \tag{108}$$

4. We have

$$\lim_{\gamma\uparrow 1} R_T^{\pi,\gamma} = R_T^{\pi}. \tag{109}$$

Proof of Part 1: From the Laurent series expansion ((Bertsekas, 2012a, Proposition 5.1.2), for any policy $\pi \in \Pi_{\text{SD}}$, we have

$$V_\gamma^\pi(s) = \frac{J^\pi}{1-\gamma} + V^\pi(s) + O(|1-\gamma|), \quad \forall s \in \mathcal{S}.$$

As a result, we have

$$\begin{aligned}
&\lim_{\gamma\uparrow 1}\Big[V_\gamma^\pi(s_1) - V_\gamma^\pi(s_2)\Big] \\
=&\lim_{\gamma\uparrow 1}\left[\frac{J^\pi}{1-\gamma} + V^\pi(s_1) + O(|1-\gamma|) - \Big[\frac{J^\pi}{1-\gamma} + V^\pi(s_2) + O(|1-\gamma|)\Big]\right] \\
=&\lim_{\gamma\uparrow 1}\Big[V^\pi(s_1) - V^\pi(s_2)\Big] = V^\pi(s_1) - V^\pi(s_2).
\end{aligned}$$

Proof of Part 2: Again from the Laurent series expansion ((Bertsekas, 2012a, Proposition 5.1.2), for any policy $\pi \in \Pi_{\text{SD}}$, we have

$$V_\gamma^\pi(s) = \frac{J^\pi}{1-\gamma} + V^\pi(s) + O(|1-\gamma|), \quad \forall s \in \mathcal{S}.$$

As a result, we have

$$\begin{aligned}
&\lim_{\gamma\uparrow 1}\Big[V_\gamma^\pi(s_1) - \gamma^T V_\gamma^\pi(s_2)\Big] \\
=&\lim_{\gamma\uparrow 1}\left[\frac{J^\pi}{1-\gamma} + V^\pi(s_1) + O(|1-\gamma|) - \Big[\frac{\gamma^T J^\pi}{1-\gamma} + \gamma^T V^\pi(s_2) + O(\gamma^T|1-\gamma|)\Big]\right] \\
=&\lim_{\gamma\uparrow 1}\left[\frac{(1-\gamma^T)}{1-\gamma}J^\pi + V^\pi(s_1) - \gamma^T V^\pi(s_2)\right] \\
=& TJ^\pi + V^\pi(s_1) - V^\pi(s_2).
\end{aligned}$$

Proof of Part 3: From the definition, we have

$$\lim_{\gamma\uparrow 1} f(T,\gamma) = \lim_{\gamma\uparrow 1}\big[\frac{\gamma^2 - \gamma^{2T+2}}{1-\gamma^2}\big] = \lim_{\gamma\uparrow 1}\big[\sum_{t=1}^{T}\gamma^{2t}\big] = T.$$

Proof of Part 4: From the definition, for any finite $T \geq 1$, we have

$$\lim_{\gamma\uparrow 1}\big[R_T^{\pi,\gamma}\big] = \lim_{\gamma\uparrow 1}\Big[\sum_{t=0}^{T-1}\gamma^t r(S_t,A_t)\Big] = \sum_{t=0}^{T-1} r(S_t,A_t) = R_T^\pi.$$

■

D.4.2 Proof of Corollary 42

Proof of Part 1: By Lemma 72, Part 4, for all $T \geq 1$, we have

$$\lim_{\gamma \uparrow 1} \big[R_T^{\pi,\gamma}\big] = R_T^{\pi}. \tag{110}$$

Moreover, we have

$$\begin{aligned}\lim_{\gamma \uparrow 1} \big[V_\gamma^\pi(S_0) - \gamma^T V_\gamma^\pi(S_T)\big] &= \lim_{\gamma \uparrow 1} \big[V_\gamma^\pi(S_0) - V_\gamma^\pi(S_T) + V_\gamma^\pi(S_T) - \gamma^T V_\gamma^\pi(S_T)\big] \\ &\overset{(a)}{=} V^\pi(S_0) - V^\pi(S_T) + TJ^\pi + V^\pi(S_T) - V^\pi(S_T) \\ &= TJ^\pi + V^\pi(S_0) - V^\pi(S_T), \end{aligned} \tag{111}$$

where (a) follows from Lemma 72, Parts 1 and 2. The result of this part follows by substituting (110)–(111) on the LHS of (27).

Proof of Part 2: By Lemma 72, Part 2, for all $s_1, s_2 \in \mathcal{S}$, we have

$$\lim_{\gamma \uparrow 1} \Big[V_\gamma^\pi(s_1) - V_\gamma^\pi(s_2)\Big] = V^\pi(s_1) - V^\pi(s_2).$$

This implies that

$$\lim_{\gamma \uparrow 1} \big[K^{\pi,\gamma}\big] = K^\pi. \tag{112}$$

Moreover, by Lemma 72, Part 3, we have

$$\lim_{\gamma \uparrow 1} f^\gamma(T) = T. \tag{113}$$

The result of this part follows by substituting (112)–(113) on the RHS of (27).

Proof of Part 3: The result of this part follows by substituting (112)–(113) on the RHS of (28).

Appendix E. Proof of Main Results for Finite-Horizon Setup

E.1 Proof of Theorem 48

E.1.1 Preliminary Results

We first present a few preliminary lemmas. To simplify the notation, we define following martingale difference sequence.

Definition 73 *Let filtration $\mathcal{F} = \{\mathcal{F}_t\}_{t=0}^h$ be defined as $\mathcal{F}_t := \sigma(S_{0:t}, A_{0:t})$. For any policy $\pi \in \Pi_{\mathrm{FD}}$, let $\{V_t^{\pi,h}\}_{t=0}^{h+1}$ denote the corresponding finite-horizon value function. We define the sequence $\{W_t^{\pi,h}\}_{t=0}^{h+1}$ as follows*

$$W_t^{\pi,h} := \Big[V_t^{\pi,h}(S_t) - \mathbb{E}\big[V_t^{\pi,h}(S_t) \;\big|\; S_{t-1}, \pi_{t-1}(S_{t-1})\big]\Big], \quad \forall t \in \{1, \dots, h+1\}, \tag{114}$$

where $\{S_t\}_{t=0}^h$ denotes the random sequence of states encountered along the current sample path.

Lemma 74 *Sequence* $\{W_t^{\pi,h}\}_{t=0}^{h+1}$ *is an MDS.*

Proof By the definition of $\{\mathcal{F}_t\}_{t=0}^{h}$, we have that S_{t-1} is $\mathcal{F}_{t-1}$-measurable. As a result, we have

$$\begin{aligned}\mathbb{E}\Big[W_t^{\pi,h} \;\Big|\; \mathcal{F}_{t-1}\Big] &= \mathbb{E}\Big[V_t^{\pi,h}(S_t) - \mathbb{E}\big[V_t^{\pi,h}(S_t) \;\big|\; S_{t-1}, \pi_{t-1}(S_{t-1})\big] \;\Big|\; \mathcal{F}_{t-1}\Big] \\ &= \mathbb{E}\Big[V_t^{\pi,h}(S_t) \;\Big|\; \mathcal{F}_{t-1}\Big] - \mathbb{E}\Big[V_t^{\pi,h}(S_t) \;\Big|\; S_{t-1}, \pi_{t-1}(S_{t-1})\Big] = 0,\end{aligned}$$

which shows that $\{W_t^{\pi,h}\}_{t=0}^{h+1}$ is an MDS with respect to the filtration $\{\mathcal{F}_t\}_{t=0}^{h}$. ■

We now present a martingale decomposition for the cumulative reward $R_T^{\pi,h}(\omega)$ for any policy $\pi \in \Pi_{\mathrm{FD}}$.

Lemma 75 *Given any policy* $\pi \in \Pi_{\mathrm{FD}}$*, we can rewrite the cumulative reward* $R_T^{\pi,h}$ *as follows*

$$R_T^{\pi,h}(\omega) = \sum_{t=1}^{T} W_t^{\pi,h} + V_0^{\pi,h}(S_0) - V_T^{\pi,h}(S_T). \tag{115}$$

Proof (FHPE) implies that along the trajectory of states $\{S_t\}_{t=0}^{T}$ induced by the policy π, we have

$$r(S_t, \pi(S_t)) = V_t^{\pi,h}(S_t) - \mathbb{E}\Big[V_{t+1}^{\pi,h}(S_{t+1}) \;\Big|\; S_t, \pi(S_t)\Big].$$

■

For any $1 \le T \le h+1$, by repeating similar steps as in the proof of Lemma 67, we have

$$\begin{aligned}R_T^{\pi,h} &= \sum_{t=0}^{T-1} \Big[V_t^{\pi,h}(S_t) - \mathbb{E}\big[V_{t+1}^{\pi,h}(S_{t+1}) \;\big|\; S_t, \pi_t(S_t)\big]\Big] \\ &\overset{(a)}{=} \sum_{t=0}^{T-1} \Big[V_t^{\pi,h}(S_t) - \mathbb{E}\big[V_{t+1}^{\pi,h}(S_{t+1}) \;\big|\; S_t, \pi_t(S_t)\big]\Big] + V_T^{\pi,h}(S_T) - V_T^{\pi,h}(S_T) \\ &\overset{(b)}{=} \sum_{t=0}^{T-1} \Big[V_{t+1}^{\pi,h}(S_{t+1}) - \mathbb{E}\big[V_{t+1}^{\pi,h}(S_{t+1}) \;\big|\; S_t, \pi_t(S_t)\big]\Big] + V_0^{\pi,h}(S_0) - V_T^{\pi,h}(S_T) \\ &\overset{(c)}{=} \sum_{t=0}^{T-1} W_{t+1}^{\pi,h} + V_0^{\pi,h}(S_0) - V_T^{\pi,h}(S_T) \\ &= \sum_{t=1}^{T} W_t^{\pi,h} + V_0^{\pi,h}(S_0) - V_T^{\pi,h}(S_T),\end{aligned}$$

where (a) follows from adding and subtracting $V_T^{\pi,h}(S_T)$, (b) follows from re-arranging the terms in the summation, and (c) follows from the definition of $\{W_t^{\pi,h}\}_{t=0}^{h+1}$ in (114).

E.1.2 Proof of Theorem 48

Proof of this theorem follows from the martingale decomposition stated in Lemma 75 and the concentration bounds stated in Theorem 61 and Theorem 63.

Proof of Part 1 By Lemma 75, we have

$$R_T^{\pi,h}(\omega) = \sum_{t=1}^{T} W_t^{\pi,h} + V_0^{\pi,h}(S_0) - V_T^{\pi,h}(S_T).$$

As a result, we have

$$\Big| R_T^{\pi,h}(\omega) - \big(V_0^{\pi,h}(S_0) - V_T^{\pi,h}(S_T)\big)\Big| = \Big|\sum_{t=1}^{T} W_t^{\pi,h}\Big|. \tag{116}$$

In order to upper-bound the term $\big|\sum_{t=1}^{T} W_t^{\pi,h}\big|$, we verify the conditions of Corollary 62. By (37), we have

$$\big|W_t^{\pi,h}\big| = \big|V_t^{\pi,h}(S_t) - \mathbb{E}\big[V_t^{\pi,h}(S_t) \;\big|\; S_{t-1}, \pi_{t-1}(S_{t-1})\big]\big| \le K_t^{\pi,h} < \infty, \quad \forall t \in \{1,\dots,T\}.$$

As a result, MDS $\{W_t^{\pi,h}\}_{t=1}^{h+1}$ is a sequentially bounded MDS with respect to the sequence $\{K_t^{\pi,h}\}_{t=1}^{h+1}$. Therefore, Corollary 62 implies that for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\begin{aligned}\Big|\sum_{t=1}^{T} W_t^{\pi,h}\Big| &\le \sqrt{2\sum_{t=1}^{T}(K_t^{\pi,h})^2 \log\frac{2}{\delta}} \\ &\overset{(a)}{=} \bar{K}_T^{\pi,h}\sqrt{2g^{\pi,h}(T)\log\frac{2}{\delta}},\end{aligned} \tag{117}$$

where (a) follows from (39). By combining (116) and (117), with probability at least $1-\delta$, we have

$$\Big| R_T^{\pi,h} - \big(V_0^{\pi,h}(S_0) - V_T^{\pi,h}(S_T)\big)\Big| \le \sqrt{2g^{\pi,h}(T)\log\frac{2}{\delta}}. \tag{118}$$

Proof of Part 2: Similar to the proof of Part 1, by Lemma 75, we have

$$\Big| R_T^{\pi,h} - \big(V_0^{\pi,h}(S_0) - V_T^{\pi,h}(S_T)\big)\Big| = \Big|\sum_{t=1}^{T} W_t^{\pi,h}\Big|. \tag{119}$$

Moreover, MDS $\{W_t^{\pi,h}\}_{t=1}^{h+1}$ is a sequentially bounded MDS with respect to the sequence $\{K_t^{\pi,h}\}_{t=1}^{h+1}$. Therefore, Theorem 63 implies that for any $\delta \in (0,1)$, if $g^{\pi,h}(h+1) \ge 173\log\frac{4}{\delta}$, define $T_0^{\pi,h}(\delta)$ to be

$$T_0^{\pi,h}(\delta) \coloneqq \min\{T' \ge 1 : g^{\pi,h}(T') \ge 173\log\frac{4}{\delta}\}.$$

Then with probability at least $1-\delta$, for all $T_0^{\pi,h}(\delta) \le T \le h+1$, we have

$$\Big|\sum_{t=1}^{T} W_t^{\pi,h}\Big| \le \sqrt{3\Big(\sum_{t=1}^{T}(K_t^{\pi,\gamma})^2\Big)\Big(2\log\log\big(\frac{3\sum_{t=1}^{T}(K_t^{\pi,\gamma})^2}{2\big|\sum_{t=1}^{T} W_t^{\pi,h}\big|}\big)+\log\frac{2}{\delta}\Big)}.$$

Now there are two cases: either $|\sum_{t=1}^{T} W_t^{\pi,h}| \le (\bar{K}_T^{\pi,h})^2$ or $|\sum_{t=1}^{T} W_t^{\pi,h}| \ge (\bar{K}_T^{\pi,\gamma})^2$. If $|\sum_{t=1}^{T} W_t^{\pi,h}| \ge (\bar{K}_T^{\pi,\gamma})^2$, we get:

$$\begin{aligned}
\Big|\sum_{t=1}^{T} W_t^{\pi,h}\Big| &\le \sqrt{3\Big(\sum_{t=1}^{T}(K_t^{\pi,h})^2\Big)\Big(2\log\log\big(\frac{3\sum_{t=1}^{T}(K_t^{\pi,h})^2}{2\big|\sum_{t=1}^{T} W_t^{\pi,h}\big|}\big)+\log\frac{2}{\delta}\Big)} \\
&\le \sqrt{3\Big(\sum_{t=1}^{T}(K_t^{\pi,h})^2\Big)\Big(2\log\log\big(\frac{3\sum_{t=1}^{T}(K_t^{\pi,h})^2}{2(\bar{K}_T^{\pi,h})^2}\big)+\log\frac{2}{\delta}\Big)} \\
&\overset{(a)}{=} \bar{K}_T^{\pi,h}\sqrt{3g^{\pi,h}(T)\Big(2\log\log\big(\frac{3}{2}g^{\pi,h}(T)\big)+\log\frac{2}{\delta}\Big)},
\end{aligned}$$

where (a) follows from the definition of $g^{\pi,h}(T)$. Otherwise, we have $|\sum_{t=1}^{T} W_t^{\pi,h}| \le (\bar{K}_T^{\pi,\gamma})^2$. As a result, we can summarize these two cases as follows

$$\Big|\sum_{t=1}^{T} W_t^{\pi,h}\Big| \le \max\Big\{\bar{K}_T^{\pi,h}\sqrt{3g^{\pi,h}(T)\big(2\log\log\big(\frac{3}{2}g^{\pi,h}(T)\big)+\log\frac{2}{\delta}\big)}, (\bar{K}_T^{\pi,h})^2\Big\}. \tag{120}$$

By combining (119)–(120), with probability at least $1-\delta$, we have

$$\begin{aligned}
&\Big|R_T^{\pi,h}(\omega) - \big(V_0^{\pi,h}(S_0) - V_T^{\pi,h}(S_T)\big)\Big| \\
&\qquad \le \max\Big\{\bar{K}_T^{\pi,h}\sqrt{3g^{\pi,h}(T)\big(2\log\log(\frac{3}{2}g^{\pi,h}(T))+\log\frac{2}{\delta}\big)}, (\bar{K}_T^{\pi,h})^2\Big\}.
\end{aligned} \tag{121}$$

E.2 Proof of Corollary 49

Proof of Part 1 By Lemma 75, we have

$$R_T^{\pi,h}(\omega) = \sum_{t=1}^{T} W_t^{\pi,h} + V_0^{\pi,h}(S_0) - V_T^{\pi,h}(S_T).$$

As a result, we have

$$\Big|R_T^{\pi,h}(\omega) - V_0^{\pi,h}(S_0)\Big| \overset{(a)}{\le} \Big|\sum_{t=1}^{T} W_t^{\pi,h}\Big| + \Big|V_T^{\pi,h}(S_T)\Big|, \tag{122}$$

where (a) follows from the triangle inequality. In the proof of Theorem 48, Part 1, we showed that with probability at least $1-\delta$, we have

$$\begin{aligned}
\Big|\sum_{t=1}^{T} W_t^{\pi,h}\Big| &\le \sqrt{2\sum_{t=1}^{T}(K_t^{\pi,h})^2\log\frac{2}{\delta}} \\
&\overset{(b)}{\le} \bar{K}_T^{\pi,h}\sqrt{2T\log\frac{2}{\delta}},
\end{aligned} \tag{123}$$

where (b) follows by $K_t^{\pi,h} \le \bar{K}_T^{\pi,h}$, for all $t \le T$. Moreover, by definition, we have

$$V_T^{\pi,h}(S_T) \le \bar{H}_T^{\pi,h}, \quad \forall t \le T. \tag{124}$$

By combining (122)–(124), with probability at least $1-\delta$, we have

$$\Big|R_T^{\pi,h}(\omega) - V_0^{\pi,h}(S_0)\Big| \le \bar{K}_T^{\pi,h}\sqrt{2T\log\frac{2}{\delta}} + \bar{H}_T^{\pi,h}.$$

Proof of Part 2: Similar to the proof of Part 1, by Lemma 75, we have

$$\Big|R_T^{\pi,h}(\omega) - V_0^{\pi,h}(S_0)\Big| \le \Big|\sum_{t=1}^{T} W_t^{\pi,h}\Big| + \Big|V_T^{\pi,h}(S_T)\Big|, \tag{125}$$

Moreover, we have

$$V_T^{\pi,h}(S_T) \le \bar{H}_T^{\pi,h}. \tag{126}$$

In addition, from proof of Theorem 48, Part 2, we have for any $\delta \in (0,1)$, for all $T \ge T_0(\delta) \coloneqq \min\{T \ge 1 : g^{\pi,h}(T) \ge 173\log\frac{4}{\delta}\}$, with probability at least $1-\delta$, we have

$$\begin{aligned}\Big|\sum_{t=1}^{T} W_t^{\pi,h}\Big| &\le \max\left\{\bar{K}_T^{\pi,h}\sqrt{3g^{\pi,h}(T)\big(2\log\log\big(\frac{3}{2}g^{\pi,h}(T)\big)+\log\frac{2}{\delta}\big)}, (\bar{K}_T^{\pi,h})^2\right\}\\ &\overset{(c)}{\le} \max\left\{\bar{K}_T^{\pi,h}\sqrt{3T\big(2\log\log\big(\frac{3T}{2}\big)+\log\frac{2}{\delta}\big)}, (\bar{K}_T^{\pi,h})^2\right\},\end{aligned} \tag{127}$$

where (c) follows from the fact that $g^{\pi,h}(T) \le T$. By combining (125)–(127), with probability at least $1-\delta$, we have

$$\Big|R_T^{\pi,h}(\omega) - V_0^{\pi,h}(S_0)\Big| \le \max\left\{\bar{K}_T^{\pi,h}\sqrt{3T\big(2\log\log\big(\frac{3T}{2}\big)+\log\frac{2}{\delta}\big)}, (\bar{K}_T^{\pi,h})^2\right\} + \bar{H}_T^{\pi,h}.$$

E.3 Proof of Corollary 50

E.3.1 Proof of Part 1

Consider two policies $\pi_1, \pi_2 \in \Pi_{\text{SD}}$. Let $\{S_t^{\pi_1}\}_{t\ge0}$ and $\{S_t^{\pi_2}\}_{t\ge0}$ denote the random sequence of states encountered by following policies π_1 and π_2. We have

$$\begin{aligned}\Big|R_T^{\pi_1,h} - R_T^{\pi_2,h}\Big| &\overset{(a)}{=} \Big|R_T^{\pi_1,h} - \big[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\big] + \big[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\big]\\ &\quad - \big[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\big] + \big[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\big] - R_T^{\pi_2,h}\Big|\\ &\overset{(b)}{\le} \Big|R_T^{\pi_1,h} - \big[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\big]\Big| + \Big|\big[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\big] - R_T^{\pi_2,h}\Big|\\ &\quad + \Big|\big[V_0^{\pi_1}(S_0^{\pi_1}) - V_T^{\pi_1}(S_T^{\pi_1})\big] - \big[V_0^{\pi_2}(S_0^{\pi_2}) - V_T^{\pi_2}(S_T^{\pi_2})\big]\Big|,\end{aligned} \tag{128}$$

where (a) follows by adding and subtracting $\left[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\right]$ and $\left[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\right]$ and (b) follows from the triangle inequality. Similarly, we have

$$\begin{aligned}
&\left|\left[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\right] - \left[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\right]\right| \overset{(a)}{=} \\
&\left|\left[V_0^{\pi_1}(S_0^{\pi_1}) - V_T^{\pi_1}(S_T^{\pi_1})\right] - R_T^{\pi_1,h} + R_T^{\pi_1,h} - R_T^{\pi_2,h} + R_T^{\pi_2,T} - \left[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\right]\right| \\
&\overset{(b)}{\leq}\left|R_T^{\pi_1,h} - \left[V_0^{\pi_1}(S_0^{\pi_1}) - V_T^{\pi_1}(S_T^{\pi_1})\right]\right| + \left|R_T^{\pi_2,h} - \left[V_0^{\pi_2}(S_0^{\pi_2}) - V_T^{\pi_2}(S_T^{\pi_2})\right]\right| \\
&+\left|R_T^{\pi_1,h} - R_T^{\pi_2,h}\right|, \qquad (129)
\end{aligned}$$

where (a) follows by adding and subtracting $R_T^{\pi_1,h}$ and $R_T^{\pi_2,h}$ and (b) follows from the triangle inequality. (128)–(129) imply that

$$\begin{aligned}
&\left|\left|R_T^{\pi_1,h} - R_T^{\pi_2,h}\right| - \left|\left[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\right] - \left[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\right]\right|\right| \\
\leq&\left|R_T^{\pi_1,h} - \left[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\right]\right| + \left|R_T^{\pi_2,h} - \left[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_\gamma^{\pi_2,h}(S_T^{\pi_2})\right]\right|. \qquad (130)
\end{aligned}$$

By Theorem 48, we know that for any $\delta_1 \in (0,1)$, with probability at least $1-\delta_1$, we have

$$\left|R_T^{\pi_1,h} - \left(V_0^{\pi_1,h}(S_0) - V_T^{\pi_1,h}(S_T)\right)\right| \leq \bar{K}_T^{\pi_1,h}\sqrt{2g^{\pi_1,h}(T)\log\frac{2}{\delta_1}}. \qquad (131)$$

Similarly, we have that for any $\delta_2 \in (0,1)$, with probability at least $1-\delta_2$, we have

$$\left|R_T^{\pi_2,h} - \left(V_0^{\pi_2,h}(S_0) - V_T^{\pi_2,h}(S_T)\right)\right| \leq \bar{K}_T^{\pi_2,h}\sqrt{2g^{\pi_2,h}(T)\log\frac{2}{\delta_2}}. \qquad (132)$$

As a result, by applying Lemma 68 and (130)–(132), for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\left|\left|R_T^{\pi_1,h} - R_T^{\pi_2,h}\right| - \left|\left[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\right] - \left[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\right]\right|\right| \\
&\leq \left|R_T^{\pi_1,h} - \left[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\right]\right| + \left|R_T^{\pi_2,h} - \left[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\right]\right| \\
&\leq \bar{K}_T^{\pi_1,h}\sqrt{2g^{\pi_1,h}(T)\log\frac{4}{\delta}} + \bar{K}_T^{\pi_2,h}\sqrt{2g^{\pi_2,h}(T)\log\frac{4}{\delta}}.
\end{aligned}$$

E.4 Proof of Part 2

As we showed in the proof of part 1, for any two policies $\pi_1, \pi_2 \in \Pi_{\text{FD}}$, we have

$$\begin{aligned}
&\left|\left|R_T^{\pi_1,h} - R_T^{\pi_2,h}\right| - \left|\left[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\right] - \left[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\right]\right|\right| \\
\leq&\left|R_T^{\pi_1,h} - \left[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\right]\right| + \left|R_T^{\pi_2,h} - \left[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_\gamma^{\pi_2,h}(S_T^{\pi_2})\right]\right|. \qquad (133)
\end{aligned}$$

By Corollary 50, for any $\delta_1 \in (0,1)$, if $g^{\pi_1,h}(h) \geq 173\log\frac{4}{\delta_1}$, let

$$T_0^{\pi,h}(\delta_1) \coloneqq \min\left\{T' \geq 1 : g^{\pi,h}(T') \geq 173\log\frac{4}{\delta_1}\right\}. \qquad (134)$$

Then with probability at least $1-\delta_1$, for all $T_0^{\pi_1,h}(\delta_1) \le T \le h+1$, we have

$$\Big|R_T^{\pi_1,h} - \big(V_0^{\pi_1,h}(S_0) - V_T^{\pi_1,h}(S_T)\big)\Big| \le \max\left\{\bar{K}_T^{\pi_1,h}\sqrt{3g^{\pi_1,h}(T)\big(2\log\log(\frac{3}{2}g^{\pi_1,h}(T)) + \log\frac{2}{\delta_1}\big)}, (\bar{K}_T^{\pi_1,h})^2\right\}. \tag{135}$$

Similarly, for any $\delta_2 \in (0,1)$, if $g^{\pi_2,h}(h) \ge 173\log\frac{4}{\delta_2}$, with probability at least $1-\delta_2$, for all $T_0^{\pi,h}(\delta_2) \le T \le h+1$, we have

$$\Big|R_T^{\pi_2,h} - \big(V_0^{\pi_2,h}(S_0) - V_T^{\pi_2,h}(S_T)\big)\Big| \le \max\left\{\bar{K}_T^{\pi_2,h}\sqrt{3g^{\pi_2,h}(T)\big(2\log\log(\frac{3}{2}g^{\pi_2,h}(T)) + \log\frac{2}{\delta_2}\big)}, (\bar{K}_T^{\pi_2,h})^2\right\}. \tag{136}$$

As a result, by applying Lemma 68, for any $\delta \in (0,1)$, if $\min\big\{g^{\pi_1,h}(h), g^{\pi_2,h}(h)\big\} \ge 173\log\frac{8}{\delta}$, let

$$T_0(\delta) \coloneqq \max\{T_0^{\pi_1,h}(\frac{8}{\delta}), T_0^{\pi_1,h}(\frac{8}{\delta})\}.$$

Then, with probability at least $1-\delta$, for all $T_0(\delta) \le T \le h+1$, we have

$$\begin{aligned}
&\Big|\big|R_T^{\pi_1,h} - R_T^{\pi_2,h}\big| - \big|\big[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\big] - \big[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\big]\big|\Big| \\
\le&\Big|R_T^{\pi_1,h} - \big[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\big]\Big| + \Big|R_T^{\pi_2,h} - \big[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\big]\Big| \\
\le&\max\left\{\bar{K}_T^{\pi_1,h}\sqrt{3g^{\pi_1,h}(T)\big(2\log\log(\frac{3}{2}g^{\pi_1,h}(T)) + \log\frac{4}{\delta}\big)}, (\bar{K}_T^{\pi_1,h})^2\right\} \\
&+\max\left\{\bar{K}_T^{\pi_2,h}\sqrt{3g^{\pi_2,h}(T)\big(2\log\log(\frac{3}{2}g^{\pi_2,h}(T)) + \log\frac{4}{\delta}\big)}, (\bar{K}_T^{\pi_2,h})^2\right\}.
\end{aligned} \tag{137}$$

Appendix F. Proof of Main Results for Random Reward Setup

F.1 Preliminary Results

F.1.1 Reward Martingale Decomposition

To simplify the notation, we define following martingale difference sequence.

Definition 76 *Let filtration $\mathcal{F} = \{\mathcal{F}_t\}_{t\ge0}$ be defined as $\mathcal{F}_t \coloneqq \sigma(S_{0:t}, A_{0:t})$. For any policy $\pi \in \tilde{\Pi}_{\mathrm{AR}}$, we define the sequence $\{\tilde{W}_t^\pi\}_{t\ge1}$ as follows*

$$\tilde{W}_t^\pi \coloneqq \tilde{r}(S_{t-1}, A_{t-1}, E_{t-1}) - \mathbb{E}\big[\tilde{r}(S_{t-1}, A_{t-1}, E_{t-1})\big|S_{t-1}, A_{t-1}{=}\pi(S_{t-1})\big], \quad \forall t \ge 1, \tag{138}$$

where $\{S_t\}_{t\ge0}$ denotes the random sequence of states encountered along the current sample path.

The sequence $\{\tilde{W}_t^\pi\}_{t\ge1}$ is an MDS with respect to the filtration $\mathcal{F} = \{\mathcal{F}_t\}_{t\ge0}$ since

$$\begin{aligned}
\mathbb{E}\Big[\tilde{W}_t^\pi\Big|\mathcal{F}_{t-1}\Big] &= \mathbb{E}\Big[\tilde{r}(S_{t-1}, A_{t-1}, E_{t-1}) - \mathbb{E}\big[\tilde{r}(S_{t-1}, A_{t-1}, E_{t-1})\big|\mathcal{F}_{t-1}\big]\Big|\mathcal{F}_{t-1}\Big] \\
&= \mathbb{E}\Big[\tilde{r}(S_{t-1}, A_{t-1}, E_{t-1})\big|S_{t-1}, A_{t-1}{=}\pi(S_{t-1})\big] \\
&\quad - \mathbb{E}\Big[\tilde{r}(S_{t-1}, A_{t-1}, E_{t-1})\big|S_{t-1}, A_{t-1}{=}\pi(S_{t-1})\big] = 0.
\end{aligned} \tag{139}$$

Recall the definitions of processes $\tilde{R}_T^\pi$ and R_T^π

$$\tilde{R}_T^\pi = \sum_{t=0}^{T-1} \tilde{r}(S_t, A_t, E_t), \tag{140}$$

$$R_T^\pi = \sum_{t=0}^{T-1} \mathbb{E}\big[\tilde{r}(S_t, A_t, E_t)\big|S_t, A_t = \pi(S_t)\big]. \tag{141}$$

By the definition of $\{\tilde{W}_t^\pi\}_{t\geq 1}$ in (138) we have

$$\tilde{R}_T^\pi = R_T^\pi + \sum_{t=1}^{T} \tilde{W}_t^\pi, \tag{142}$$

where $\{\tilde{W}_t^\pi\}_{t\geq 1}$ is an MDS with respect to the filtration $\mathcal{F} = \{\mathcal{F}_t\}_{t\geq 0}$ by (139). The following theorem establishes the concentration of the process R_T^π around the quantity $T\tilde{J}^\pi - \big(\tilde{V}^\pi(S_T) - \tilde{V}^\pi(S_0)\big)$.

Theorem 77 *For any policy $\pi \in \tilde{\Pi}_{\mathrm{AR}}$, the following upper-bounds hold:*

1. *For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have:*

$$\Big|R_T^\pi - T\tilde{J}^\pi - \big(\tilde{V}^\pi(S_0) - \tilde{V}^\pi(S_T)\big)\Big| \leq \tilde{K}^\pi \sqrt{2T \log \frac{2}{\delta}}. \tag{143}$$

2. *For any $\delta \in (0,1)$, for all $T \geq T_0(\delta) := \Big\lceil \dfrac{173}{\tilde{K}^\pi} \log \dfrac{4}{\delta} \Big\rceil$, with probability at least $1-\delta$, we have*

$$\Big|R_T^\pi - T\tilde{J}^\pi - \big(\tilde{V}^\pi(S_0) - \tilde{V}^\pi(S_T)\big)\Big| \leq \max\Big\{\tilde{K}^\pi \sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{2}{\delta}\Big)}, \big(\tilde{K}^\pi\big)^2\Big\}. \tag{144}$$

Proof This theorem follows by applying the result of Theorem 17 to the reduced model of $\tilde{\mathcal{M}}$ defined in Def. 53. Let $\mathcal{M} = (P, r)$ denote the reduced model of $\tilde{\mathcal{M}} = (P, \tilde{r})$ with $r(s,a)$ defined in (47). We can rewrite the process R_T^π as the cumulative reward process associated with the reduced model $\mathcal{M} = (P, r)$, i.e.,

$$R_T^\pi = \sum_{t=0}^{T-1} \mathbb{E}\big[\tilde{r}(S_t, A_t, E_t)\big|S_t, A_t = \pi(S_t)\big] = \sum_{t=0}^{T-1} r(S_t, A_t).$$

The result of this theorem follows by applying the result of Theorem 17 on the cumulative reward process R_T^π and recalling that quantities $\tilde{J}^\pi$, $\tilde{V}^\pi$, and $\tilde{K}^\pi$ are defined based on the reduced model M. ■

F.2 Proof of Theorem 54

F.2.1 Proof of Part 1

For any policy $\pi \in \tilde{\Pi}_{\mathrm{AR}}$, we have

$$\big|\tilde{R}_T^\pi - T\tilde{J}^\pi - \big(\tilde{V}^\pi(S_0) - \tilde{V}^\pi(S_T)\big)\big| = \big|R_T^\pi - T\tilde{J}^\pi - \big(\tilde{V}^\pi(S_0) - \tilde{V}^\pi(S_T)\big) + \tilde{R}_T^\pi - R_T^\pi\big|. \tag{145}$$

By (142), we have

$$\Big|\tilde{R}_T^\pi - R_T^\pi\Big| = \Big|\sum_{t=1}^T \tilde{W}_t^\pi\Big|.$$

Moreover by (49), $\{\tilde{W}_t^\pi\}_{t\geq 0}$ is a uniformly bounded MDS with respect to the constant $\tilde{K}_r^\pi$. Therefore, Corollary 62 implies that for any $\delta_1 \in (0,1)$, with probability at least $1-\delta_1$, we have

$$\Big|\sum_{t=1}^T \tilde{W}_t^\pi\Big| \leq \tilde{K}_r^\pi \sqrt{2T \log \frac{2}{\delta_1}}. \tag{146}$$

Moreover, Theorem 77, Part 1, implies that for any $\delta_2 \in (0,1)$, with probability at least $1-\delta_2$, we have

$$\Big|R_T^\pi - T\tilde{J}^\pi - \Big(\tilde{V}^\pi(S_0) - \tilde{V}^\pi(S_T)\Big)\Big| \leq \tilde{K}^\pi \sqrt{2T \log \frac{2}{\delta_2}}. \tag{147}$$

As a result, by combining (145), (146), and (147) and applying Lemma 68, for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\Big|R_T^\pi - T\tilde{J}^\pi - \Big(\tilde{V}^\pi(S_0) - \tilde{V}^\pi(S_T)\Big)\Big| \tag{148}$$

$$\leq \tilde{K}^\pi \sqrt{2T \log \frac{4}{\delta}} + \tilde{K}_r^\pi \sqrt{2T \log \frac{4}{\delta}}. \tag{149}$$

F.2.2 Proof of Part 2

Similar to the proof of Part 1, for any policy $\pi \in \tilde{\Pi}_{\mathrm{AR}}$, we have

$$\big|\tilde{R}_T^\pi - T\tilde{J}^\pi - \big(\tilde{V}^\pi(S_0) - \tilde{V}^\pi(S_T)\big)\big| = \big|R_T^\pi - T\tilde{J}^\pi - \big(\tilde{V}^\pi(S_0) - \tilde{V}^\pi(S_T)\big) + \tilde{R}_T^\pi - R_T^\pi\big|. \tag{150}$$

By (142), we have

$$\Big|\tilde{R}_T^\pi - R_T^\pi\Big| = \Big|\sum_{t=1}^T \tilde{W}_t^\pi\Big|.$$

Moreover by (49), $\{\tilde{W}_t^\pi\}_{t\geq 0}$ is a uniformly bounded MDS with respect to the constant $\tilde{K}_r^\pi$. Therefore, Corollary 64 implies that for any $\delta_1 \in (0,1)$, for all $T \geq T_0(\delta_1) \coloneqq \Big\lceil \frac{173}{\tilde{K}_r^\pi} \log \frac{4}{\delta_1} \Big\rceil$, with probability at least $1-\delta_1$, we have

$$\Big|\sum_{t=1}^T \tilde{W}_t^\pi\Big| \leq \max\Big\{\tilde{K}_r^\pi \sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{2}{\delta_1}\Big)}, \big(\tilde{K}_r^\pi\big)^2\Big\}. \tag{151}$$

Moreover, Theorem 77, Part 2, implies that for any $\delta_2 \in (0,1)$, for all $T \geq T_0(\delta_2) \coloneqq \left\lceil \frac{173}{\tilde{K}^\pi} \log \frac{4}{\delta_2} \right\rceil$, with probability at least $1-\delta_2$, we have

$$\left| R_T^\pi - T\tilde{J}^\pi - \left(\tilde{V}^\pi(S_0) - \tilde{V}^\pi(S_T)\right) \right| \leq \max\left\{ \tilde{K}^\pi \sqrt{3T\left(2\log\log\frac{3T}{2} + \log\frac{2}{\delta_2}\right)}, \left(\tilde{K}^\pi\right)^2 \right\}. \quad (152)$$

As a result, by combining (150), (151), and (152) and applying Lemma 68, for any $\delta \in (0,1)$, for all $T \geq T_0(\delta) \coloneqq \max\left\{ \left\lceil \frac{173}{\tilde{K}^\pi} \log \frac{8}{\delta} \right\rceil, \left\lceil \frac{173}{\tilde{K}_r^\pi} \log \frac{8}{\delta} \right\rceil \right\}$, with probability at least $1-\delta$, we have

$$\begin{aligned} &\left| \tilde{R}_T^\pi - T\tilde{J}^\pi - \left(\tilde{V}^\pi(S_0) - \tilde{V}^\pi(S_T)\right) \right| && (153) \\ &\quad \leq \max\left\{ \tilde{K}^\pi \sqrt{3T\left(2\log\log\frac{3T}{2} + \log\frac{4}{\delta}\right)}, (\tilde{K}^\pi)^2 \right\} && (154) \\ &\quad + \max\left\{ \tilde{K}_r^\pi \sqrt{3T\left(2\log\log\frac{3T}{2} + \log\frac{4}{\delta}\right)}, (\tilde{K}_r^\pi)^2 \right\}. && (155) \end{aligned}$$

Appendix G. Miscellaneous Theorems

G.1 Slutsky's Theorem

Theorem 78 (see (Ash and Doléans-Dade, 2000, Theorem 7.7.1)) *If $X_t \xrightarrow{(d)} X$ and $Y_t \xrightarrow{(d)} c$, where $c \in \mathbb{R}$ (equivalently $Y_t \xrightarrow{(P)} c$) then we have*

1. $X_t + Y_t \xrightarrow{(d)} X + c$.
2. $X_t Y_t \xrightarrow{(d)} cX$.
3. $\frac{X_t}{Y_t} \xrightarrow{(d)} \frac{X}{c}$, *if* $c \neq 0$.

Remark 79 *Since convergence in the almost-sure sense implies convergence in probability, same results hold when $Y_t \xrightarrow{(a.s.)} c$.*