Concentration of Cumulative Reward in Markov Decision Processes

Borna Sayedana borna.sayedana@mail.mcgill.ca
Peter E. Caines peterc@cim.mcgill.ca
Aditya Mahajan aditya.mahajan@mcgill.ca
Department of Electrical and Computer Engineering,
McGill University, Montreal, QC, H3A 0E9, Canada.

Editor:

Abstract

In this paper, we investigate the concentration properties of cumulative rewards 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 novel martingale decomposition of cumulative rewards, 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. In the MDP setup, the main focus is on maximizing the expected cumulative rewards (aka., return) (Sutton and Barto, 2018). However, in many applications, focusing only on the expected cumulative reward overlooks important aspects of its distribution, which may lead to undesirable outcomes. As a result, various methods have been developed to design policies that shape the distribution of cumulative rewards to have specific characteristics. These include frameworks such as risk-sensitive

MDPs (Ruszczyński (2010)), constrained MDPs (Beutler and Ross (1985); Altman (2021)), and distributional reinforcement learning (Bellemare et al. (2017, 2023)).

Another line of research focuses on characterizing the sample path and distributional behavior of cumulative rewards in the standard MDP framework. The variance of discounted cumulative rewards is investigated in Sobel (1982). Using Markov chain theory, asymptotic concentration of cumulative rewards, such as the Law of Large Numbers (LLN), the Central Limit Theorem (CLT), and the Law of Iterated Logarithms (LIL) are established in the average cost setting (Duflo (2013); Meyn and Tweedie (2012); Hernández-Lerma and Lasserre (2012)).

In this paper, we revisit this problem and 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. To the best of our knowledge, our results are the first non-asymptotic concentration characteristics of cumulative rewards for MDPs in finite-horizon, discounted reward and average reward setups.

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 rewards 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 rewards 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 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 law of iterated logarithm. Using the vanishing discount analysis, we show that under appropriate conditions, the concentration bounds for discounted reward MDPs approaches to the concentration bounds for the average reward MDPs as the discount factor approaches 1.

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. Our

concluding remarks are presented in Sec. 6. 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 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 $V^* : \mathcal{S} \to \mathbb{R}$ 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 $V^\pi : \mathcal{S} \to \mathbb{R}$ 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 $\Pi_{\mathrm{AR}} \subseteq \Pi_{\mathrm{SD}}$ 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 $\Pi_{\mathrm{AR}} \subseteq \Pi_{\mathrm{SD}}$. 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 $\Pi_{\mathrm{AR}} = \Pi_{\mathrm{SD}}$, as well as conditions guaranteeing that Π_{AR} is non-empty, showing the existence of an optimal policy $\pi^* \in \Pi_{\mathrm{AR}}$.

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

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

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

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 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 $T^\pi(s, s')$. The diameter of $\mathcal{M}$ is defined as*

$$D = \mathrm{diam}(\mathcal{M}) \coloneqq \max_{\substack{s,s' \in \mathcal{S} \\ s \neq s'}} \min_{\pi \in \Pi_{\mathrm{SD}}} T^\pi(s, s').$$

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*

$$\sigma^\pi(s) \leq K^\pi \leq H^\pi < \infty, \quad \forall s \in \mathcal{S}. \tag{6}$$

2. *If $\mathcal{M}$ is communicating, then for any policy $\pi \in \Pi_{\mathrm{AR}}$, we have $H^\pi \leq DR_{\max}$. Therefore,*

$$\sigma^\pi(s) \leq K^\pi \leq H^\pi \leq DR_{\max}, \quad \forall s \in \mathcal{S}. \tag{7}$$

3. *If $\mathcal{M}$ is weakly communicating, then for any optimal policy $\pi^* \in \Pi_{\mathrm{AR}}$, we have $H^* \leq DR_{\max}$. Therefore,*

$$\sigma^*(s) \leq K^* \leq H^* \leq DR_{\max}, \quad \forall s \in \mathcal{S}. \tag{8}$$

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 $\{\Sigma_t^\pi\}_{t \geq 0}$ 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.,*

$$\lim_{T\to\infty} \frac{R^\pi_T(\omega)}{T} = J^\pi, \quad a.s. \tag{9}$$

2. *(Central Limit Theorem) Assume that $\mathbb{P}(\Omega^\pi_0) = 1$. Let the stopping time ν_t be defined as $\nu_t \coloneqq \min\Big\{T \geq 1 \,:\, \Sigma^\pi_T \geq t\Big\}$. Then*

$$\lim_{T\to\infty} \frac{R^\pi_{\nu_T}(\omega) - \nu_T J^\pi}{\sqrt{T}} \xrightarrow{(d)} \mathcal{N}(0,1). \tag{10}$$

3. *(Law of Iterated Logarithm) For almost all $\omega \in \Omega^\pi_0$, we have*

$$\liminf_{T\to\infty} \frac{R^\pi_T(\omega) - TJ^\pi}{\sqrt{2\Sigma^\pi_T \log\log \Sigma^\pi_T}} = -1, \quad \limsup_{T\to\infty} \frac{R^\pi_T(\omega) - TJ^\pi}{\sqrt{2\Sigma^\pi_T \log\log \Sigma^\pi_T}} = 1. \tag{11}$$

The proof is presented in App. C.2.

Corollary 16 *For any optimal policy $\pi^* \in \Pi^*$, the cumulative reward $R^{\pi^*}_T(\omega)$ satisfies the asymptotic concentration rates in (9)–(11), 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 (9)–(11). ■

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*

$$\big|R^\pi_T - TJ^\pi - \big(V^\pi(S_0) - V^\pi(S_T)\big)\big| \leq K^\pi \sqrt{2T \log\frac{2}{\delta}}. \tag{12}$$

2. *For any $\delta \in (0,1)$, for all $T \geq T^\pi_0(\delta) \coloneqq \Big\lceil \frac{173}{K^\pi} \log\frac{4}{\delta} \Big\rceil$, with probability at least $1-\delta$, we have*

$$\big|R^\pi_T - TJ^\pi - \big(V^\pi(S_0) - V^\pi(S_T)\big)\big| \leq \max\Big\{K^\pi \sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{2}{\delta}\Big)}, (K^\pi)^2\Big\}. \tag{13}$$

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 $R^\pi_T(\omega)$ 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*

$$\left|R_T^\pi - TJ^\pi\right| \le K^\pi \sqrt{2T \log \frac{2}{\delta}} + H^\pi. \tag{14}$$

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

$$\left|R_T^\pi - TJ^\pi\right| \le \max\left\{K^\pi \sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{2}{\delta}\Big)}, (K^\pi)^2\right\} + H^\pi. \tag{15}$$

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 (14)–(15), 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 (14)–(15). ■

Corollary 20 *If $\mathcal{M}$ is unichain or recurrent, then any policy $\pi \in \Pi_{\mathrm{SD}}$ satisfies asymptotic concentration rates in (9)–(11) and non-asymptotic concentration rates in (14)–(15).*

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 (9)–(11) hold for the policy π and by Theorem 18, the non-asymptotic rates in (14)–(15) 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 (9)–(11) and non-asymptotic concentration rates in (14)–(15). (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 (9)–(11) hold for every optimal policy $\pi^* \in \Pi_{\mathrm{AR}}$. Furthermore, by Corollary 19, the non-asymptotic concentration rates in (14)–(15) 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. As a result, these bounds are policy-dependent upper-bounds. At the cost of loosening these bounds, we derive policy-independent upper-bounds. These bounds are in terms of the diameter of the MDP D and the maximum reward $R_{\max}$.

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*

$$\left|R_T^\pi - TJ^\pi\right| \le DR_{\max}\sqrt{2T\log\frac{2}{\delta}} + DR_{\max}. \tag{16}$$

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

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

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 (16)–(17), where in the LHS, J^π is replaced with J^*.*

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 $DR_{\max}$. This implies that if the diameter of $\mathcal{M}$ or maximum reward $R_{\max}$ increases, these upper-bounds loosen with a linear rate.

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 24 *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*

$$\Big|\left|R_T^{\pi_1} - R_T^{\pi_2}\right| - \left|TJ^{\pi_1} - TJ^{\pi_2}\right|\Big| \le K^{\pi_1}\sqrt{2T\log\frac{4}{\delta}} + H^{\pi_1} + K^{\pi_2}\sqrt{2T\log\frac{4}{\delta}} + H^{\pi_2}. \tag{18}$$

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

$$\begin{aligned}\Big|\left|R_T^{\pi_1} - R_T^{\pi_2}\right| - \left|TJ^{\pi_1} - TJ^{\pi_2}\right|\Big| \le{}& \max\left\{K^{\pi_1}\sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{4}{\delta}\Big)}, (K^{\pi_1})^2\right\} + H^{\pi_1} \\ &+ \max\left\{K^{\pi_2}\sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{4}{\delta}\Big)}, (K^{\pi_2})^2\right\} + H^{\pi_2}.\end{aligned} \tag{19}$$

The proof is presented in App. C.7.

Corollary 25 *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*

$$\big|R_T^{\pi_1^*} - R_T^{\pi_2^*}\big| \leq 2\Big(K^* \sqrt{2T \log \frac{4}{\delta}} + H^*\Big). \tag{20}$$

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

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

Proof Since both policies $\pi_1^*, \pi_2^* \in \Pi_{\text{AR}}$ 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 24, the difference $\big|R_T^{\pi_1^*} - R_T^{\pi_2^*}\big|$ satisfies the non-asymptotic concentration rates in Corollary 24 with the RHS of (18)–(19) being simplified to RHS of (20)–(21). ■

Remark 26 *Similar to the Corollary 22, by imposing the assumption that $\mathcal{M}$ is communicating or weakly communicating, we can derive the counterpart of* (18)–(19) *and* (20)–(21) *in terms of $DR_{\max}$ 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 27 *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 28 *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*

$$\liminf_{T\to\infty} \frac{\mathcal{D}_T(\omega)}{\sqrt{2\Sigma_T^* \log\log \Sigma_T^*}} = -1, \quad \limsup_{T\to\infty} \frac{\mathcal{D}_T(\omega)}{\sqrt{2\Sigma_T^* \log\log \Sigma_T^*}} = 1. \tag{22}$$

Proof is presented in App. C.8.

In addition to the asymptotic results presented in Theorem 28, 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 29 *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 29 do not depend on the choice of the learning policy either.

In Theorem 29, 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 30 *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 31 *Notice that conditions of Corollary 30 are weaker than the conditions of Corollary 22. As a result, Corollary 30 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 32 *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 $\bar{R}_T^\mu(\omega) \leq \tilde{\mathcal{O}}(\sqrt{T})$, 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 $R_T^\mu(\omega) \leq \tilde{\mathcal{O}}(\sqrt{T})$.*

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

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

Note that $R_T^{\pi,\gamma}(\omega)$ 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

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

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_γ^π 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

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

A policy π^* is called optimal if

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

Definition 33 *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 34 *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 35 (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 $\pi \in \Pi_{\text{SD}}$, 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

$$H^{\pi,\gamma} \coloneqq \text{sp}(V_\gamma^\pi) = \max_{s\in\mathcal{S}} V_\gamma^\pi(s) - \min_{s\in\mathcal{S}} V_\gamma^\pi(s). \tag{23}$$

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

$$K^{\pi,\gamma} \coloneqq \max_{s,s_+\in\mathcal{S}} \Big| V_\gamma^\pi(s_+) - \mathbb{E}\big[V_\gamma^\pi(S_+) \mid s, \pi(s)\big]\Big|. \tag{24}$$

For any optimal policy $\pi^* \in \Pi_{\text{SD}}$, 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 36 *For any policy $\pi \in \Pi_{\text{SD}}$ and any $s \in \mathcal{S}$, we have:*

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

$$\Big|R_T^{\pi,\gamma} - \big(V_\gamma^\pi(S_0) - \gamma^T V_\gamma^\pi(S_T)\big)\Big| \le K^{\pi,\gamma}\sqrt{2f^\gamma(T)\log\frac{2}{\delta}}. \tag{25}$$

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 $T \ge T_0(\delta) \coloneqq \min\Big\{T' \ge 1 : f^\gamma(T') > \dfrac{173}{K^{\pi,\gamma}}\log\dfrac{4}{\delta}\Big\}$, with probability at least $1-\delta$, we have*

$$\begin{aligned}&\Big|R_T^{\pi,\gamma} - \big(V_\gamma^\pi(S_0) - \gamma^T V_\gamma^\pi(S_T)\big)\Big| \\ &\qquad \le \max\left\{K^{\pi,\gamma}\sqrt{3f^\gamma(T)\Big(2\log\log\big(\frac{3}{2}f^\gamma(T)\big) + \log\frac{2}{\delta}\Big)}, (K^{\pi,\gamma})^2\right\}.\end{aligned} \tag{26}$$

The proof is presented in App. D.1.

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

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

$$\left| R_T^{\pi,\gamma} - V_\gamma^\pi(S_0) \right| \le K^{\pi,\gamma} \sqrt{2 f^\gamma(T) \log \frac{2}{\delta}} + \frac{\gamma^T}{1-\gamma} R_{\max}. \tag{27}$$

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 $T \ge T_0(\delta) := \min\left\{ T' \ge 1 : f^\gamma(T') > \dfrac{173}{K^{\pi,\gamma}} \log \dfrac{4}{\delta} \right\}$, with probability at least $1-\delta$, we have*

$$\begin{aligned} &\left| R_T^{\pi,\gamma} - V_\gamma^\pi(S_0) \right| \\ &\qquad \le \max\left\{ K^{\pi,\gamma} \sqrt{3 f^\gamma(T) \big(2 \log\log(\frac{3}{2} f^\gamma(T)) + \log \frac{2}{\delta} \big)}, (K^{\pi,\gamma})^2 \right\} + \frac{\gamma^T}{1-\gamma} R_{\max}. \end{aligned} \tag{28}$$

The proof is presented in App. D.2.

Corollary 38 *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 (25)–(28), 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 36 and Corollary 37, the optimal policy satisfies the non-asymptotic concentration rates in (25)–(28). ■

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 39 *Consider two policies $\pi_1, \pi_2 \in \Pi_{\mathrm{SD}}$. Let $\{S_t^{\pi_1}\}_{t\ge 0}$ and $\{S_t^{\pi_2}\}_{t\ge 0}$ 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*

$$\begin{aligned} &\Big| \big| R_T^{\pi_1,\gamma} - R_T^{\pi_2,\gamma} \big| - \big| \big[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1}) \big] - \big[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2}) \big] \big| \Big| \\ &\le K^{\pi_1,\gamma} \sqrt{2 f^\gamma(T) \log \frac{4}{\delta}} + K^{\pi_2,\gamma} \sqrt{2 f^\gamma(T) \log \frac{4}{\delta}}. \end{aligned} \tag{29}$$

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*

$$T_0^{\pi_i}(\frac{\delta}{2}) \coloneqq \min\Big\{T' \geq 1 : f^\gamma(T') > \frac{173}{K^{\pi_i,\gamma}} \log \frac{8}{\delta}\Big\}, \quad i \in \{1,2\}. \tag{30}$$

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

$$\begin{aligned}
&\Big|\big|R_T^{\pi_1,\gamma} - R_T^{\pi_2,\gamma}\big| - \big|\big[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\big] - \big[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\big]\big|\Big| \\
&\leq \max\Big\{K^{\pi_1,\gamma}\sqrt{3f^\gamma(T)\Big(2\log\log\frac{3}{2}f^\gamma(T) + \log\frac{4}{\delta}\Big)}, (K^{\pi_1,\gamma})^2\Big\} \\
&+ \max\Big\{K^{\pi_2,\gamma}\sqrt{3f^\gamma(T)\Big(2\log\log\frac{3}{2}f^\gamma(T) + \log\frac{4}{\delta}\Big)}, (K^{\pi_2,\gamma})^2\Big\}.
\end{aligned} \tag{31}$$

The proof if presented in App. D.3.

Corollary 40 *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*

$$\Big|\big|R_T^{\pi_1^*,\gamma} - R_T^{\pi_2^*,\gamma}\big| - \gamma^T\big|V_\gamma^*(S_T^{\pi_2^*}) - V_\gamma^*(S_T^{\pi_1^*})\big|\Big| \leq 2\Big(K^{*,\gamma}\sqrt{2f^\gamma(T)\log\frac{4}{\delta}}\Big). \tag{32}$$

2. *Consider $T_0^{\pi^*}(\frac{\delta}{2})$ defined in (30). For any $\delta \in (0,1)$, for all $T \geq T_0^{\pi^*}(\frac{\delta}{2})$, with probability at least $1-\delta$, we have*

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

Proof Since both policies $\pi_1^*, \pi_2^* \in \Pi_{\text{SD}}$ 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 39, the difference $\big|R_T^{\pi_1^*,\gamma} - R_T^{\pi_2^*,\gamma}\big|$ satisfies the non-asymptotic concentration rates in Corollary 39 with the RHS of (29) and (31) being simplified to RHS of (32)–(33). ■

4.4 Vanishing Discount Analysis

In order to observe the connection between the upper-bounds established in Theorem 17 and Theorem 36, 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 41 *For any policy $\pi \in \Pi_{\mathrm{AR}}$, we have the following asymptotic relations between the bounds in Theorem 17 and Theorem 36.*

1. *As γ goes to 1 from below, the quantity in the LHS of (25)–(26) converges to the LHS of (12), 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 (25) converges to the RHS in (12), i.e.,*

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

3. *As γ goes to 1 from below, the RHS in (26) converges to the RHS in (13), i.e.,*

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

Proof is presented in App. D.4.

Remark 42 *The non-asymptotic characterizations are established in Theorem 36. Since the discounted cumulative return $R_T^{\pi,\gamma}$ is finite for $\mathcal{M}$, we cannot provide any asymptotic characterization for this quantity. However, Corollary 41 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 43 *Given a model $\mathcal{M} = (P, r, h)$, define Π_{FD} to be the set of finite-horizon deterministic policies, i.e., for any $\pi = (\pi_0, \pi_1, \ldots, \pi_h) \in \Pi_{\mathrm{FD}}$, 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

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

where $\mathbb{E}^{\pi}$ is the expectation with respect to the joint distribution of all the system variables induced by π. 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 44 *The sequence of finite-horizon optimal value functions $\big\{V_t^{*,h}\big\}_{t=0}^{h+1} : \mathcal{S} \to \mathbb{R}$ is defined as follows*

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

and for $t \in \{h, h-1, \ldots, 0\}$, 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 45 *Given a policy $\pi \in \Pi_{\mathrm{FD}}$, the sequence of finite-horizon value functions $\big\{V_t^{\pi,h}\big\}_{t=0}^{h+1} : \mathcal{S} \to \mathbb{R}$ corresponding to the policy π is defined as follows*

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

and for $t \in \{h, h-1, \ldots, 0\}$, 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 46 (Bertsekas (2012b)) *Let $\pi^* = (\pi_0^*, \pi_1^*, \ldots, \pi_h^*) \in \Pi_{\mathrm{FD}}$ 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

$$H_t^{\pi,h} \coloneqq \mathrm{sp}(V_t^{\pi,h}), \quad \forall t \in \{0,1,\dots,h\}. \tag{34}$$

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

$$K_t^{\pi,h} \coloneqq \max_{s,s_+} \Big| V_t^{\pi,h}(s_+) - \mathbb{E}\big[V_t^{\pi,h}(S_+) \bigm| s, \pi_t(s)\big] \Big|, \quad \forall t \in \{0,1,\dots,h\}. \tag{35}$$

Similar to the results in Theorem 18 and Theorem 36 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

$$\bar{K}_T^{\pi,h} = \max_{0\le t\le T} K_t^{\pi,h}, \quad \bar{H}_T^{\pi,h} = \max_{0\le t\le T} H_t^{\pi,h}, \tag{36}$$

and let

$$g^{\pi,h}(T) \coloneqq \frac{\sum_{t=1}^{T} (K_t^{\pi,h})^2}{\big(\bar{K}_T^{\pi,h}\big)^2}. \tag{37}$$

For any optimal policy $\pi^* \in \Pi_{\mathrm{FD}}$, 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 36.

Theorem 47 *For any policy $\pi \in \Pi_{\mathrm{FD}}$, we have:*

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

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

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*

$$T_0^{\pi,h}(\delta) \coloneqq \min\Big\{ T' \ge 1 : g^{\pi,h}(T') \ge 173 \log \frac{4}{\delta} \Big\}. \tag{38}$$

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

$$\begin{aligned} &\Big| R_T^{\pi,h} - \big(V_0^{\pi,h}(S_0) - V_T^{\pi,h}(S_T)\big) \Big| \\ &\qquad \le \max\left\{ \bar{K}_T^{\pi,h} \sqrt{3 g^{\pi,h}(T) \Big(2 \log\log(\frac{3}{2} g^{\pi,h}(T)) + \log \frac{2}{\delta} \Big)}, (\bar{K}_T^{\pi,h})^2 \right\}. \end{aligned} \tag{39}$$

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 37.

Corollary 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*

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

2. *For any $\delta \in (0,1)$, if $g^{\pi,h}(h) \geq 173\log\frac{4}{\delta}$, define $T_0^{\pi,h}(\delta)$ as specified in (38). Then with probability at least $1-\delta$, for all $T_0^{\pi,h}(\delta) \leq T \leq h+1$, we have*

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

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 49 *Consider two policies $\pi_1, \pi_2 \in \Pi_{\text{FD}}$. 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 policy 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*

$$\begin{aligned}&\Big|\big|R_T^{\pi_1,h} - R_T^{\pi_2,h}\big| - \big|\big[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\big] - \big[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\big]\big|\Big| \\ &\leq \bar{K}_T^{\pi_1,h}\sqrt{2g^{\pi_1,h}(T)\log\frac{4}{\delta}} + \bar{K}_T^{\pi_2,h}\sqrt{2g^{\pi_2,h}(T)\log\frac{4}{\delta}}. \end{aligned} \tag{40}$$

2. *For any $\delta \in (0,1)$, if $\min\left\{g^{\pi_1,h}(h), g^{\pi_2,h}(h)\right\} \geq 173\log\frac{8}{\delta}$, define $T_0^{\pi,h}(\delta)$ as specified in (38) 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

$$\begin{aligned}
&\Big|\big|R_T^{\pi_1,h} - R_T^{\pi_2,h}\big| - \big|\big[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\big] - \big[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\big]\big|\Big| \\
&\le \max\left\{\bar{K}_T^{\pi_1,h}\sqrt{3g^{\pi_1,h}(T)\Big(2\log\log(\frac{3}{2}g^{\pi_1,h}(T)) + \log\frac{4}{\delta}\Big)}, (\bar{K}_T^{\pi_1,h})^2\right\} \\
&+ \max\left\{\bar{K}_T^{\pi_2,h}\sqrt{3g^{\pi_2,h}(T)\Big(2\log\log(\frac{3}{2}g^{\pi_2,h}(T)) + \log\frac{4}{\delta}\Big)}, (\bar{K}_T^{\pi_2,h})^2\right\}. \qquad (41)
\end{aligned}$$

The proof is presented in App. E.3.

Corollary 50 *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 $\big|R_T^{\pi_1^*,h} - R_T^{\pi_2^*,h}\big|$, we have:*

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

$$\Big|\big|R_T^{\pi_1^*,h} - R_T^{\pi_2^*,h}\big| - \big|V_T^{*,h}(S_T^{\pi_2^*}) - V_T^{*,h}(S_T^{\pi_1^*})\big|\Big| \le 2\Big(\bar{K}_T^{*,h}\sqrt{2g^{*,h}(T)\log\frac{4}{\delta}}\Big). \qquad (42)$$

2. *For any $\delta \in (0,1)$, if $g^{*,h}(h) \ge 173\log\frac{4}{\delta}$, define $T_0^{\pi^*,h}(\delta)$ as specified in (38). Then with probability at least $1-\delta$, for all $T_0^{\pi^*,h}(\delta) \le T \le h+1$, we have*

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

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

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

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 49, the difference $\big|R_T^{\pi_1^*,h} - R_T^{\pi_2^*,h}\big|$ satisfies the non-asymptotic concentration rates in Corollary 49 with the RHS of (40)–(41) being simplified to RHS of (42)–(43). ■

6 Conclusion

In this paper, we investigated the sample path behavior of cumulative rewards 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 $\tilde{\mathcal{O}}(\sqrt{T})$. Lastly, we extended the non-asymptotic concentration results to the case of discounted reward MDPs and finite-horizon setup. 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. (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. A natural future research direction is to establish similar results for MDPs with non-compact state and action spaces.

7 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), RGPIN-2021-03511 (A. Mahajan), and Alliance Grant 570764-2021 (A. Mahajan).

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.

Eitan Altman. *Constrained Markov Decision Processes*. Routledge, New York, 2021. ISBN 978-1-351-45082-3. doi: 10.1201/9781315140223.

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.

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, 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.

Frederick J. Beutler and Keith W. Ross. Optimal policies for controlled markov chains with a constraint. *Journal of Mathematical Analysis and Applications*, 112(1):236–252, 1985.

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.

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.

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.

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

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*. PhD thesis, Université de Lille 1, Sciences et Technologies; CRIStAL UMR 9189, 2019.

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.

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.

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

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

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.

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.

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

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 I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *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 H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *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.

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.

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

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

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 2nd edition, 2018. ISBN 978-0-262-03924-6.

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.

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 Table of 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 51 *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 52 *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 $\{Y_t\}_{t\geq 0}$ be the MDS corresponding to $\{X_t\}_{t\geq 0}$. Then, we can express $\{A_t\}_{t\geq 0}$ in terms of $\{Y_t^2\}_{t\geq 0}$. 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 $\{Y_t^2\}_{t\geq 0}$.

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 53 (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 54 (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 55 (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 54. 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 56 (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 56, we get following equivalent form of Azuma-Hoeffding inequality.

Corollary 57 *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 56 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 58 (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 $T \geq T_0(\delta) := \min\Big\{T \,:\, \sum_{t=1}^{T} c_t^2 \geq 173\log\frac{4}{\delta}\Big\}$, with probability at least $1-\delta$, we have*

$$\Big|\sum_{t=1}^{T} Y_t\Big| \leq \sqrt{3\Big(\sum_{t=1}^{T} c_t^2\Big)\Big(2\log\log\frac{3\sum_{t=1}^{T} c_t^2}{2\big|\sum_{t=1}^{T} Y_t\big|} + \log\frac{2}{\delta}\Big)}. \tag{44}$$

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

Corollary 59 *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 $T \geq T_0(\delta) := \Big\lceil \frac{173}{C}\log\frac{4}{\delta}\Big\rceil$, with probability at least $1-\delta$, we have*

$$\Big|\sum_{t=1}^{T} Y_t\Big| \leq C\max\Big\{\sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{2}{\delta}\Big)}, C\Big\}. \tag{45}$$

Proof This corollary follows from Theorem 58, by substituting the sequence $\{c_t\}_{t\geq 1}$ with the constant C on the RHS of (44). 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 58, 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

$$\Big|\sum_{t=1}^{T} Y_t\Big| \leq \max\Bigg\{C\sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{2}{\delta}\Big)}, C^2\Bigg\}. \tag{46}$$

■

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 60 *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*

$$M_t^\pi \coloneqq V^\pi(S_t) - \mathbb{E}\big[V^\pi(S_t) \;\big|\; S_{t-1}, \pi(S_{t-1})\big], \quad \forall t \geq 1, \tag{47}$$

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

Lemma 61 *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 62 *Given any policy $\pi \in \Pi_{\mathrm{AR}}$, we can rewrite the cumulative reward R_T^π as follows*

$$R_T^\pi = TJ^\pi + \sum_{t=1}^{T} M_t^\pi + V^\pi(S_0) - V^\pi(S_T). \tag{48}$$

Proof Since $\pi \in \Pi_{\mathrm{AR}}$, (ARPE) implies that along the trajectory of states $\{S_{1:t}\}$ 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 (47).

■

C.1.2 A Consequence of The Union Bound

Lemma 63 *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

$$\mathbb{P}\Big(|X_T| > h_1(T, \delta/2)\Big) < \frac{\delta}{2}. \tag{49}$$

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

$$\mathbb{P}\Big(|Y_T| > h_2(T, \delta/2)\Big) < \frac{\delta}{2}. \tag{50}$$

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 (49)–(50), 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

$$\sigma_\pi(s) \overset{(a)}{\leq} K^\pi \overset{(b)}{\leq} H^\pi \overset{(c)}{\leq} \infty, \quad \forall s \in \mathcal{S}. \tag{51}$$

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

$$V^\pi(s) - \mathbb{E}[V^\pi(S_+)|s,\pi(s)] \leq V^\pi(s) - \min_{s\in\mathcal{S}} V^\pi(s) \leq \max_{s\in\mathcal{S}} V^\pi(s) - \min_{s\in\mathcal{S}} V^\pi(s), \quad \forall s \in \mathcal{S}. \tag{52}$$

Similarly, we have

$$\mathbb{E}[V^\pi(S_+)|s,\pi(s)] - V^\pi(s) \leq \max_{s\in\mathcal{S}} V^\pi(s) - V^\pi(s) \leq \max_{s\in\mathcal{S}} V^\pi(s) - \min_{s\in\mathcal{S}} V^\pi(s), \quad \forall s \in \mathcal{S}. \tag{53}$$

Therefore (52)–(53) 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 $H^\pi \le DR_{\max}$. Consider $s, s' \in \mathcal{S}$ where $s \neq s'$. By (Puterman, 2014), we have:

$$V^\pi(s) = \mathbb{E}\Big[\sum_{t=0}^{\infty} [r(S_t, A_t) - J^\pi] \;\Big|\; S_0 = s \Big]. \tag{54}$$

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

$$V^\pi(s) - V^\pi(s') = \mathbb{E}\Big[\sum_{t=0}^{\tau_0 - 1} [r(S_t, A_t) - J^\pi] \Big] \le \mathbb{E}\Big[\sum_{t=0}^{\tau_0 - 1} [r(S_t, A_t)] \Big] \overset{(e)}{\le} T^\pi(s_1, s_2) R_{\max} \overset{(f)}{\le} DR_{\max},$$

where (e) follows from the definition of $T^\pi(s_1, s_2)$ and (f) 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 $H^\pi \le DR_{\max}$.

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 $H^* \le DR_{\max}$.

C.2 Proof of Theorem 15

C.2.1 Proof of Part 1

By Lemma 62, for any policy $\pi \in \Pi_{\mathrm{AR}}$, we can rewrite the cumulative return R_T^π as follows

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

By (5) and Lemma 13 we have

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

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 53, 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 54 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

$$\frac{\sum_{t=1}^{\nu_T} M_t^\pi}{\sqrt{T}} \xrightarrow{(d)} \mathcal{N}(0,1). \tag{55}$$

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

$$\frac{V^\pi(S_0) - V^\pi(S_T)}{\sqrt{T}} \to 0, \quad a.s. \tag{56}$$

By combining (55) and (56) and by using Theorem 71, 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 55 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 55, we get

$$\liminf_{T\to\infty} \frac{\sum_{t=1}^T M_t^\pi}{\sqrt{2\Sigma_t^\pi \log\log \Sigma_t^\pi}} = -1, \quad \limsup_{T\to\infty} \frac{\sum_{t=1}^T M_t^\pi}{\sqrt{2\Sigma_t^\pi \log\log \Sigma_t^\pi}} = 1. \tag{57}$$

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

$$\lim_{T\to\infty} \frac{V^\pi(S_0) - V^\pi(S_T)}{\sqrt{2\Sigma_t^\pi \log\log \Sigma_t^\pi}} = 0, \quad \text{a.s.} \tag{58}$$

By combining (57) and (58), 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 62, for any policy $\pi \in \Pi_{\text{AR}}$, we can rewrite the cumulative return $R_T^\pi(\omega)$ 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

$$\big|R_T^\pi(\omega) - TJ^\pi - \big(V^\pi(S_0) - V^\pi(S_T)\big)\big| = \big|\sum_{t=1}^T M_t^\pi\big|. \tag{59}$$

In order to upper-bound the term $\big|\sum_{t=1}^T M_t^\pi\big|$, we verify the conditions of Corollary 57. 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 57 implies that for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\Big|\sum_{t=1}^T M_t^\pi\Big| \leq \sqrt{2T(K^\pi)^2 \log(\frac{2}{\delta})}. \tag{60}$$

By combining (59) and (60), 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 62, we have

$$\left|R_T^\pi(\omega) - TJ^\pi - \left(V^\pi(S_0) - V^\pi(S_T)\right)\right| = \left|\sum_{t=1}^T M_t^\pi\right| \tag{61}$$

Moreover, MDS $\{M_t^\pi\}_{t\ge 0}$ is a uniformly bounded MDS with respect to the constant K^π. Therefore, Corollary 59 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

$$\left|\sum_{t=1}^T M_t^\pi\right| \le \max\left\{K^\pi \sqrt{3T\left(2\log\log\frac{3T}{2} + \log\frac{2}{\delta}\right)}, (K^\pi)^2\right\}. \tag{62}$$

By combining (61) and (62), 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 62, for any policy $\pi \in \Pi_{\mathrm{AR}}$, we can rewrite the cumulative return $R_T^\pi(\omega)$ 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

$$\begin{aligned}
\left|R_T^\pi(\omega) - TJ^\pi\right| &= \left|\sum_{t=1}^T M_t^\pi + V^\pi(S_0) - V^\pi(S_T)\right| \\
&\overset{(a)}{\le} \left|\sum_{t=1}^T M_t^\pi\right| + \left|V^\pi(S_0) - V^\pi(S_T)\right| \\
&\overset{(b)}{\le} \left|\sum_{t=1}^T M_t^\pi\right| + H^\pi,
\end{aligned} \tag{63}$$

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 57. 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 57 implies that for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\Big|\sum_{t=1}^{T} M_t^\pi\Big| \leq \sqrt{2T(K^\pi)^2 \log(\frac{2}{\delta})}. \tag{64}$$

By combining (63) and (64), 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 62, we have

$$\big|R_T^\pi(\omega) - TJ^\pi\big| \leq \Big|\sum_{t=1}^{T} M_t^\pi\Big| + H^\pi. \tag{65}$$

Moreover, MDS $\{M_t^\pi\}_{t\geq 0}$ is a uniformly bounded MDS with respect to the constant K^π. Therefore, Corollary 59 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

$$\Big|\sum_{t=1}^{T} M_t^\pi\Big| \leq \max\Big\{K^\pi \sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{2}{\delta}\Big)}, (K^\pi)^2\Big\}. \tag{66}$$

By combining (65) and (66), 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, for any policy $\pi \in \Pi_{\mathrm{AR}}$, we have

$$\big|M_t^\pi\big| \leq K^\pi \leq DR_{\max}, \quad \forall t \geq 1. \tag{67}$$

As a result, the MDS $\{M_t^\pi\}_{t\geq 1}$ is a uniformly bounded MDS with respect to the constant $DR_{\max}$. Therefore, by repeating the arguments of the proof of Theorem 18, Part 1, and substituting H^π with $DR_{\max}$ in the RHS of (63) and replacing K^π with $DR_{\max}$ in the RHS of (64), we get that with probability at least $1-\delta$, we have:

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

C.5.2 Proof of Part 2

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

$$\left|M_t^\pi\right| \le K^\pi \le DR_{\max}, \quad \forall t \ge 1. \tag{68}$$

As a result, the MDS $\{M_t^\pi\}_{t\ge1}$ is a uniformly bounded MDS with respect to the constant $DR_{\max}$. Therefore, by repeating the arguments of the proof of Theorem 18, Part 2, and substituting H^π with $DR_{\max}$ in the RHS of (65) and substituting K^π with $DR_{\max}$ in the RHS of (66), we prove the claim, i.e, for any $\delta \in (0,1)$, for all $T \ge T_0(\delta) \coloneqq \left\lceil \frac{173}{DR_{\max}} \log \frac{4}{\delta} \right\rceil$, with probability at least $1-\delta$, we have

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

C.6 Proof of Corollary 23

In the case of communicating $\mathcal{M}$, since $\pi^* \in \Pi_{\mathrm{AR}}$, by Corollary 22, we get that $R_T^{\pi^*}(\omega)$ satisfies the non-asymptotic concentration rates in (16)–(17).

In the case of weakly communicating $\mathcal{M}$, by Lemma 13, for any optimal policy $\pi^* \in \Pi_{\mathrm{AR}}$, we have

$$\left|M_t^{\pi^*}\right| = \left|V^*(S_t) - \mathbb{E}\big[V^*(S_t) \mid S_{t-1}, \pi(S_{t-1})\big]\right| \le K^* \le DR_{\max}, \quad \forall t \ge 1. \tag{69}$$

As a result, the MDS $\{M_t^{\pi^*}\}_{t\ge1}$ 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 (16)–(17).

C.7 Proof of Corollary 24

C.7.1 Proof of Part 1

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

$$\begin{aligned}\left|R_T^{\pi_1} - R_T^{\pi_2}\right| =& \left|R_T^{\pi_1} - TJ^{\pi_1} + TJ^{\pi_1} - TJ^{\pi_2} + TJ^{\pi_2} - R_T^{\pi_2}\right| \\ \overset{(a)}{\le}& \left|R_T^{\pi_1} - TJ^{\pi_1}\right| + \left|TJ^{\pi_1} - TJ^{\pi_2}\right| + \left|TJ^{\pi_2} - R_T^{\pi_2}\right|,\end{aligned} \tag{70}$$

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

$$\begin{aligned}\left|TJ^{\pi_1} - TJ^{\pi_2}\right| =& \left|TJ^{\pi_1} - R_T^{\pi_1} + R_T^{\pi_1} - R_T^{\pi_2} + R_T^{\pi_2} - TJ^{\pi_2}\right| \\ \overset{(b)}{\le}& \left|TJ^{\pi_1} - R_T^{\pi_1}\right| + \left|R_T^{\pi_1} - R_T^{\pi_2}\right| + \left|R_T^{\pi_2} - TJ^{\pi_2}\right|,\end{aligned} \tag{71}$$

where (b) follows from the triangle inequality. (70)–(71) imply that

$$\left|\left|R_T^{\pi_1} - R_T^{\pi_2}\right| - \left|TJ^{\pi_1} - TJ^{\pi_2}\right|\right| \le \left|R_T^{\pi_1} - TJ^{\pi_1}\right| + \left|R_T^{\pi_2} - TJ^{\pi_2}\right|. \tag{72}$$

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 63 and (72), we get that 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 63 and (72), we get that 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 28

By Corollary 16, for any optimal policy $\pi^* \in \Pi_{\mathrm{AR}}$, the quantity R^{π^*} satisfies the asymptotic concentration rates in (9)–(11). 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 (9)–(11), 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 29

By Corollary 19, for any optimal policy $\pi^* \in \Pi_{\mathrm{AR}}$, the quantity $|R_T^{\pi^*} - TJ^*|$ satisfies the asymptotic concentration rates in (14)–(15). 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 (14)–(15), 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 30

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 (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 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 32

C.11.1 Proof of Part 1

This part of the theorem is a consequence of Theorem 28. Recall that by definition, we have

$$\mathcal{D}_T(\omega) = \mathcal{R}_T^{\mu}(\omega) - \bar{\mathcal{R}}_T^{\mu}(\omega). \tag{73}$$

On the other hand, we can rewrite the law of iterated logarithm in Theorem 28 using the $\tilde{\mathcal{O}}(\cdot)$ notation as follows

$$\mathcal{D}_T(\omega) \leq \tilde{\mathcal{O}}(\sqrt{T}), \quad a.s. \tag{74}$$

As a result, for any learning policy μ that satisfies $R_T^{\mu}(\omega) \leq \tilde{\mathcal{O}}(\sqrt{T})$, almost surely, (73)–(74) 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, (73)–(74) 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 29. 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

$$R_T^\mu(\omega) \leq h_1(\delta_1, T), \tag{75}$$

where for a fixed δ_1, we have $h_1(\delta_1, T) = \tilde{\mathcal{O}}(\sqrt{T})$. Moreover, by Theorem 29, 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

$$D_T(\omega) \leq h_2(\delta_2, T), \tag{76}$$

where for a fixed δ_2, we have $h_2(\delta_2, T) = \tilde{\mathcal{O}}(\sqrt{T})$. As a result, by (73), (75)–(76), and Lemma 63, we get that for any $\delta \in (0,1)$, for all $T \geq \max\{T_1(\delta/2), T_2(\delta/2)\}$, 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 36

D.1.1 Preliminary Results

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

Definition 64 *Let filtration $\mathcal{F} = \{\mathcal{F}_t\}_{t\geq 0}$ be defined as $\mathcal{F}_t := \sigma(S_{0:t}, A_{0:t})$. For any policy $\pi \in \Pi_{\text{SD}}$, let V_γ^π denote the corresponding discounted value function. We define the sequence $\{N_t^{\pi,\gamma}\}_{t\geq 1}$ as follows*

$$N_t^{\pi,\gamma} := \Big[V_\gamma^\pi(S_t) - \mathbb{E}\big[V_\gamma^\pi(S_t) \;\big|\; S_{t-1}, \pi(S_{t-1})\big]\Big], \quad \forall t \geq 1, \tag{77}$$

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

Lemma 65 *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 $\pi \in \Pi_{\mathrm{SD}}$.

Lemma 66 *Given any policy $\pi \in \Pi_{\mathrm{SD}}$, we can rewrite the discounted cumulative return $R_T^{\pi,\gamma}$ as follows*

$$R_T^{\pi,\gamma}(\omega) = \sum_{t=1}^{T} \gamma^t N_t^{\pi,\gamma} + V_\gamma^\pi(S_0) - \gamma^T V_\gamma^\pi(S_T). \tag{78}$$

Proof Since $\pi \in \Pi_{\mathrm{SD}}$, (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 62, 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 $\{N_t^{\pi,\gamma}\}_{t\geq 0}$. ■

D.1.2 Proof of Theorem 36

Proof of this theorem follows from the martingale decomposition stated in Lemma 66 and the concentration bounds stated in Corollary 57 and Theorem 58.

Proof of Part 1 By Lemma 66, 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

$$\Big| R_T^{\pi,\gamma}(\omega) - \big(V_\gamma^\pi(S_0) - \gamma^T V_\gamma^\pi(S_T)\big)\Big| = \big|\sum_{t=1}^{T} \gamma^t N_t^{\pi,\gamma}\big|. \tag{79}$$

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 57. By (24) 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 57 implies that for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\begin{aligned}
\Big|\sum_{t=1}^{T} \gamma^t N_t^{\pi,\gamma}\Big| &\le \sqrt{2\sum_{t=1}^{T} (K^{\pi,\gamma})^2 \gamma^{2t} \log\frac{2}{\delta}} \\
&= K^{\pi,\gamma}\sqrt{2\sum_{t=1}^{T} \gamma^{2t} \log\frac{2}{\delta}} \\
&= K^{\pi,\gamma}\sqrt{2 f^\gamma(T) \log\frac{2}{\delta}}.
\end{aligned} \tag{80}$$

As a result, by combining (79) and (80), we get that with probability at least $1-\delta$, we have

$$\Big| R_T^{\pi,\gamma}(\omega) - \big(V_\gamma^\pi(S_0) - \gamma^T V_\gamma^\pi(S_T)\big)\Big| \le K^{\pi,\gamma}\sqrt{2 f^\gamma(T) \log\frac{2}{\delta}}. \tag{81}$$

Proof of Part 2: Similar to the proof of Part 1, by Lemma 66, we have

$$\Big| R_T^{\pi,\gamma}(\omega) - \big(V_\gamma^\pi(S_0) - \gamma^T V_\gamma^\pi(S_T)\big)\Big| = \big|\sum_{t=1}^{T} \gamma^t N_t^{\pi,\gamma}\big|. \tag{82}$$

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 58 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

$$\Big|\sum_{t=1}^{T}\gamma^t N_t^{\pi,\gamma}\Big| \le \max\Big\{K^{\pi,\gamma}\sqrt{3f^{\gamma}(T)\big(2\log\log\frac{3}{2}f^{\gamma}(T)+\log\frac{2}{\delta}\big)},(K^{\pi,\gamma})^2\Big\}. \tag{83}$$

By combining (82)–(83), with probability at least $1-\delta$, we have

$$\begin{aligned}&\Big|R_T^{\pi,\gamma}(\omega)-\big(V_{\gamma}^{\pi}(S_0)-\gamma^T V_{\gamma}^{\pi}(S_T)\big)\Big|\\ &\qquad\le \max\Big\{K^{\pi,\gamma}\sqrt{3f^{\gamma}(T)\big(2\log\log(\frac{3}{2}f^{\gamma}(T))+\log\frac{2}{\delta}\big)},(K^{\pi,\gamma})^2\Big\}.\end{aligned} \tag{84}$$

D.2 Proof of Corollary 37

Proof of Part 1: By Lemma 66, 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

$$\Big|R_T^{\pi,\gamma}(\omega)-V_{\gamma}^{\pi}(S_0)\Big| \overset{(a)}{\le} \Big|\sum_{t=1}^{T}\gamma^t N_t^{\pi,\gamma}\Big|+\Big|\gamma^T V_{\gamma}^{\pi}(S_T)\Big|, \tag{85}$$

where (a) follows from the triangle inequality. In the proof of Theorem 36, Part 1, we showed that with probability at least $1-\delta$, we have

$$\Big|\sum_{t=1}^{T}\gamma^t N_t^{\pi,\gamma}\Big| \le K^{\pi,\gamma}\sqrt{2f^{\gamma}(T)\log\frac{2}{\delta}}. \tag{86}$$

Moreover, we have

$$\begin{aligned}\gamma^T V_{\gamma}^{\pi}(S_T)&=\gamma^T\mathbb{E}^{\pi}\Big[\lim_{T\to\infty}\sum_{t=0}^{T-1}\gamma^t r(S_t,A_t)\;\Big|\;S_0=S_T\Big]\\ &=\gamma^T\mathbb{E}^{\pi}\Big[\lim_{T\to\infty}\sum_{t=0}^{T-1}\gamma^t R_{\max}\;\Big|\;S_0=S_T\Big]\le\frac{\gamma^T}{1-\gamma}R_{\max}.\end{aligned} \tag{87}$$

By combining (85)–(87), 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 66, we have

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

Moreover, we have

$$\Big|\gamma^T V_\gamma^\pi(S_T)\Big| \le \gamma^T\frac{R_{\max}}{1-\gamma}. \tag{89}$$

In addition, from proof of Theorem 36, 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

$$\Big|\sum_{t=1}^{T}\gamma^t N_t^{\pi,\gamma}\Big| \le \max\Big\{K^{\pi,\gamma}\sqrt{3f^\gamma(T)\big(2\log\log\frac{3}{2}f^\gamma(T) + \log\frac{2}{\delta}\big)}, (K^{\pi,\gamma})^2\Big\}. \tag{90}$$

By combining (88)–(90), with probability at least $1-\delta$, we have

$$\begin{aligned}&\Big|R_T^{\pi,\gamma}(\omega) - V_\gamma^\pi(S_0)\Big| \\ &\qquad\le \max\Big\{K^{\pi,\gamma}\sqrt{3f^\gamma(T)\big(2\log\log(\frac{3}{2}f^\gamma(T)) + \log\frac{2}{\delta}\big)}, (K^{\pi,\gamma})^2\Big\} + \frac{\gamma^T}{1-\gamma}R_{\max}.\end{aligned} \tag{91}$$

D.3 Proof of Corollary 39

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

$$\begin{aligned}\Big|R_T^{\pi_1,\gamma} - R_T^{\pi_2,\gamma}\Big| &\overset{(a)}{=} \Big|R_T^{\pi_1,\gamma} - \big[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\big] + \big[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\big] \\ &\quad - \big[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\big] + \big[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\big] - R_T^{\pi_2,\gamma}\Big| \\ &\overset{(b)}{\le} \Big|R_T^{\pi_1,\gamma} - \big[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\big]\Big| + \Big|\big[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\big] - R_T^{\pi_2,\gamma}\Big| \\ &\quad + \Big|\big[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\big] - \big[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\big]\Big|,\end{aligned} \tag{92}$$

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

$$\begin{aligned}
&\left|\left[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\right] - \left[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\right]\right| \overset{(a)}{=} \\
&\left|\left[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\right] - R_T^{\pi_1,\gamma} + R_T^{\pi_1,\gamma} - R_T^{\pi_2,\gamma} + R_T^{\pi_2,\gamma} - \left[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\right]\right| \\
\overset{(b)}{\leq}&\left|R_T^{\pi_1,\gamma} - \left[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\right]\right| + \left|R_T^{\pi_2,\gamma} - \left[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\right]\right| \\
&+\left|R_T^{\pi_1,\gamma} - R_T^{\pi_2,\gamma}\right|, \qquad (93)
\end{aligned}$$

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. (92)–(93) imply that

$$\begin{aligned}
&\left|\left|R_T^{\pi_1,\gamma} - R_T^{\pi_2,\gamma}\right| - \left|\left[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\right] - \left[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\right]\right|\right| \\
\leq&\left|R_T^{\pi_1,\gamma} - \left[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\right]\right| + \left|R_T^{\pi_2,\gamma} - \left[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\right]\right|. \qquad (94)
\end{aligned}$$

By Theorem 36, we know that for any $\delta_1 \in (0,1)$, with probability at least $1-\delta_1$, we have

$$\left|R_T^{\pi_1,\gamma} - \left[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\right]\right| \leq K^{\pi_1,\gamma}\sqrt{2f^\gamma(T)\log\frac{2}{\delta_1}}. \qquad (95)$$

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

$$\left|R_T^{\pi_2,\gamma} - \left[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\right]\right| \leq K^{\pi_2,\gamma}\sqrt{2f^\gamma(T)\log\frac{2}{\delta_2}}. \qquad (96)$$

As a result, by applying Lemma 63 and (94)–(96), we get that for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

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

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

$$\begin{aligned}
&\left|\left|R_T^{\pi_1,\gamma} - R_T^{\pi_2,\gamma}\right| - \left|\left[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\right] - \left[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\right]\right|\right| \\
\leq&\left|R_T^{\pi_1,\gamma} - \left[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\right]\right| + \left|R_T^{\pi_2,\gamma} - \left[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\right]\right|. \qquad (97)
\end{aligned}$$

By Theorem 36, 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 63, we get that 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 41

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 67 *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*

$$\lim_{\gamma\uparrow 1} f(T,\gamma) = T. \tag{98}$$

4. We have

$$\lim_{\gamma\uparrow 1} R_T^{\pi,\gamma} = R_T^{\pi}. \tag{99}$$

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 41

Proof of Part 1: By Lemma 67, Part 4, for all $T \geq 1$, we have

$$\lim_{\gamma \uparrow 1} \big[R_T^{\pi,\gamma}\big] = R_T^{\pi}. \tag{100}$$

Moreover, we have

$$\begin{aligned}\lim_{\gamma \uparrow 1} \big[V_\gamma^\pi(S_0) - \gamma^T V_\gamma^\pi(S_T)\big] &= \lim_{\gamma \uparrow 1} \big[V_\gamma^\pi(S_0) - V_\gamma^\pi(S_T) + V_\gamma^\pi(S_T) - \gamma^T V_\gamma^\pi(S_T)\big] \\ &\overset{(a)}{=} V^\pi(S_0) - V^\pi(S_T) + TJ^\pi + V^\pi(S_T) - V^\pi(S_T) \\ &= TJ^\pi + V^\pi(S_0) - V^\pi(S_T), \end{aligned} \tag{101}$$

where (a) follows from Lemma 67, Parts 1 and 2. The result of this part follows by substituting (100)–(101) on the LHS of (25).

Proof of Part 2: By Lemma 67, 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

$$\lim_{\gamma \uparrow 1} \big[K^{\pi,\gamma}\big] = K^\pi. \tag{102}$$

Moreover, by Lemma 67, Part 3, we have

$$\lim_{\gamma \uparrow 1} f^\gamma(T) = T. \tag{103}$$

The result of this part follows by substituting (102)–(103) on the RHS of (25).

Proof of Part 3: The result of this part follows by substituting (102)–(103) on the RHS of (26).

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

E.1 Proof of Theorem 47

E.1.1 Preliminary Results

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

Definition 68 *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*

$$W_t^{\pi,h} := \Big[V_t^{\pi,h}(S_t) - \mathbb{E}\big[V_t^{\pi,h}(S_t) \;\big|\; S_{t-1}, \pi_{t-1}(S_{t-1})\big]\Big], \quad \forall t \in \{1, \dots, h+1\}, \tag{104}$$

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

Lemma 69 *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 70 *Given any policy* $\pi \in \Pi_{\mathrm{FD}}$*, we can rewrite the cumulative reward* $R_T^{\pi,h}$ *as follows*

$$R_T^{\pi,h}(\omega) = \sum_{t=1}^{T} W_t^{\pi,h} + V_0^{\pi,h}(S_0) - V_T^{\pi,h}(S_T). \tag{105}$$

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 62, 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 (104).

E.1.2 Proof of Theorem 47

Proof of this theorem follows from the martingale decomposition stated in Lemma 70 and the concentration bounds stated in Theorem 56 and Theorem 58.

Proof of Part 1 By Lemma 70, 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

$$\Big| R_T^{\pi,h}(\omega) - \big(V_0^{\pi,h}(S_0) - V_T^{\pi,h}(S_T)\big)\Big| = \Big|\sum_{t=1}^{T} W_t^{\pi,h}\Big|. \tag{106}$$

In order to upper-bound the term $\big|\sum_{t=1}^{T} W_t^{\pi,h}\big|$, we verify the conditions of Corollary 57. By (35), we have

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

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 57 implies that for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\begin{aligned}\Big|\sum_{t=1}^{T} W_t^{\pi,h}\Big| &\le \sqrt{2\sum_{t=1}^{T}(K_t^{\pi,h})^2 \log\frac{2}{\delta}} \\ &\overset{(a)}{=} \bar{K}_T^{\pi,h}\sqrt{2g^{\pi,h}(T)\log\frac{2}{\delta}},\end{aligned} \tag{107}$$

where (a) follows from (37). By combining (106) and (107), we get that with probability at least $1-\delta$, we have

$$\Big| R_T^{\pi,h} - \big(V_0^{\pi,h}(S_0) - V_T^{\pi,h}(S_T)\big)\Big| \le \sqrt{2g^{\pi,h}(T)\log\frac{2}{\delta}}. \tag{108}$$

Proof of Part 2: Similar to the proof of Part 1, by Lemma 70, we have

$$\Big| R_T^{\pi,h} - \big(V_0^{\pi,h}(S_0) - V_T^{\pi,h}(S_T)\big)\Big| = \Big|\sum_{t=1}^{T} W_t^{\pi,h}\Big|. \tag{109}$$

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 58 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

$$\Big|\sum_{t=1}^{T} W_t^{\pi,h}\Big| \le \max\Big\{\bar{K}_T^{\pi,h}\sqrt{3g^{\pi,h}(T)\big(2\log\log\big(\frac{3}{2}g^{\pi,h}(T)\big)+\log\frac{2}{\delta}\big)}, (\bar{K}_T^{\pi,h})^2\Big\}. \tag{110}$$

By combining (109)–(110), with probability at least $1-\delta$, we have

$$\begin{aligned}
&\Big|R_T^{\pi,h}(\omega) - \big(V_0^{\pi,h}(S_0) - V_T^{\pi,h}(S_T)\big)\Big| \\
&\qquad \le \max\Big\{\bar{K}_T^{\pi,h}\sqrt{3g^{\pi,h}(T)\big(2\log\log(\frac{3}{2}g^{\pi,h}(T))+\log\frac{2}{\delta}\big)}, (\bar{K}_T^{\pi,h})^2\Big\}.
\end{aligned} \tag{111}$$

E.2 Proof of Corollary 48

Proof of Part 1 By Lemma 70, 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

$$\Big|R_T^{\pi,h}(\omega) - V_0^{\pi,h}(S_0)\Big| \overset{(a)}{\le} \Big|\sum_{t=1}^{T} W_t^{\pi,h}\Big| + \Big|V_T^{\pi,h}(S_T)\Big|, \tag{112}$$

where (a) follows from the triangle inequality. In the proof of Theorem 47, Part 1, we showed that with probability at least $1-\delta$, we have

$$\begin{aligned}
\Big|\sum_{t=1}^{T} W_t^{\pi,h}\Big| &\le \sqrt{2\sum_{t=1}^{T}(K_t^{\pi,h})^2\log\frac{2}{\delta}} \\
&\overset{(b)}{\le} \bar{K}_T^{\pi,h}\sqrt{2T\log\frac{2}{\delta}},
\end{aligned} \tag{113}$$

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

$$V_T^{\pi,h}(S_T) \le \bar{H}_T^{\pi,h}, \quad \forall t \le T. \tag{114}$$

By combining (112)–(114), 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 70, we have

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

Moreover, we have

$$V_T^{\pi,h}(S_T) \le \bar{H}_T^{\pi,h}. \tag{116}$$

In addition, from proof of Theorem 47, 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

$$\begin{aligned}\Big|\sum_{t=1}^{T} W_t^{\pi,h}\Big| &\le \max\left\{\bar{K}_T^{\pi,h}\sqrt{3g^{\pi,h}(T)\big(2\log\log\big(\frac{3}{2}g^{\pi,h}(T)\big)+\log\frac{2}{\delta}\big)}, (\bar{K}_T^{\pi,h})^2\right\}\\ &\overset{(c)}{\le} \max\left\{\bar{K}_T^{\pi,h}\sqrt{3T\big(2\log\log\big(\frac{3T}{2}\big)+\log\frac{2}{\delta}\big)}, (\bar{K}_T^{\pi,h})^2\right\},\end{aligned} \tag{117}$$

where (c) follows from the fact that $g^{\pi,h}(T) \le T$. By combining (115)–(117), 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 49

E.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 sequence of states encountered by following policies π_1 and π_2. We have

$$\begin{aligned}\Big|R_T^{\pi_1,h} - R_T^{\pi_2,h}\Big| &\overset{(a)}{=} \Big|R_T^{\pi_1,h} - \big[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\big] + \big[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\big]\\ &\quad - \big[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\big] + \big[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\big] - R_T^{\pi_2,h}\Big|\\ &\overset{(b)}{\le} \Big|R_T^{\pi_1,h} - \big[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\big]\Big| + \Big|\big[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\big] - R_T^{\pi_2,h}\Big|\\ &\quad + \Big|\big[V_0^{\pi_1}(S_0^{\pi_1}) - V_T^{\pi_1}(S_T^{\pi_1})\big] - \big[V_0^{\pi_2}(S_0^{\pi_2}) - V_T^{\pi_2}(S_T^{\pi_2})\big]\Big|,\end{aligned} \tag{118}$$

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

$$\begin{aligned}
&\left|\left[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\right] - \left[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\right]\right| \overset{(a)}{=} \\
&\left|\left[V_0^{\pi_1}(S_0^{\pi_1}) - V_T^{\pi_1}(S_T^{\pi_1})\right] - R_T^{\pi_1,h} + R_T^{\pi_1,h} - R_T^{\pi_2,h} + R_T^{\pi_2,T} - \left[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\right]\right| \\
&\overset{(b)}{\leq}\left|R_T^{\pi_1,h} - \left[V_0^{\pi_1}(S_0^{\pi_1}) - V_T^{\pi_1}(S_T^{\pi_1})\right]\right| + \left|R_T^{\pi_2,h} - \left[V_0^{\pi_2}(S_0^{\pi_2}) - V_T^{\pi_2}(S_T^{\pi_2})\right]\right| \\
&+\left|R_T^{\pi_1,h} - R_T^{\pi_2,h}\right|, \qquad (119)
\end{aligned}$$

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. (118)–(119) imply that

$$\begin{aligned}
&\left|\left|R_T^{\pi_1,h} - R_T^{\pi_2,h}\right| - \left|\left[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\right] - \left[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\right]\right|\right| \\
\leq&\left|R_T^{\pi_1,h} - \left[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\right]\right| + \left|R_T^{\pi_2,h} - \left[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_\gamma^{\pi_2,h}(S_T^{\pi_2})\right]\right|. \qquad (120)
\end{aligned}$$

By Theorem 47, we know that for any $\delta_1 \in (0,1)$, with probability at least $1-\delta_1$, we have

$$\left|R_T^{\pi_1,h} - \left(V_0^{\pi_1,h}(S_0) - V_T^{\pi_1,h}(S_T)\right)\right| \leq \bar{K}_T^{\pi_1,h}\sqrt{2g^{\pi_1,h}(T)\log\frac{2}{\delta_1}}. \qquad (121)$$

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

$$\left|R_T^{\pi_2,h} - \left(V_0^{\pi_2,h}(S_0) - V_T^{\pi_2,h}(S_T)\right)\right| \leq \bar{K}_T^{\pi_2,h}\sqrt{2g^{\pi_2,h}(T)\log\frac{2}{\delta_2}}. \qquad (122)$$

As a result, by applying Lemma 63 and (120)–(122), we get that 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 $\pi_1, \pi_2 \in \Pi_{\mathrm{FD}}$, we have

$$\begin{aligned}
&\left|\left|R_T^{\pi_1,h} - R_T^{\pi_2,h}\right| - \left|\left[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\right] - \left[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\right]\right|\right| \\
\leq&\left|R_T^{\pi_1,h} - \left[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\right]\right| + \left|R_T^{\pi_2,h} - \left[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_\gamma^{\pi_2,h}(S_T^{\pi_2})\right]\right|. \qquad (123)
\end{aligned}$$

By Corollary 49, for any $\delta_1 \in (0,1)$, if $g^{\pi_1,h}(h) \geq 173\log\frac{4}{\delta_1}$, let

$$T_0^{\pi,h}(\delta_1) \coloneqq \min\left\{T' \geq 1 : g^{\pi,h}(T') \geq 173\log\frac{4}{\delta_1}\right\}. \qquad (124)$$

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

$$\Big|R_T^{\pi_1,h} - \big(V_0^{\pi_1,h}(S_0) - V_T^{\pi_1,h}(S_T)\big)\Big| \\ \le \max\left\{\bar{K}_T^{\pi_1,h}\sqrt{3g^{\pi_1,h}(T)\big(2\log\log(\frac{3}{2}g^{\pi_1,h}(T)) + \log\frac{2}{\delta_1}\big)}, (\bar{K}_T^{\pi_1,h})^2\right\}. \tag{125}$$

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

$$\Big|R_T^{\pi_2,h} - \big(V_0^{\pi_2,h}(S_0) - V_T^{\pi_2,h}(S_T)\big)\Big| \\ \le \max\left\{\bar{K}_T^{\pi_2,h}\sqrt{3g^{\pi_2,h}(T)\big(2\log\log(\frac{3}{2}g^{\pi_2,h}(T)) + \log\frac{2}{\delta_2}\big)}, (\bar{K}_T^{\pi_2,h})^2\right\}. \tag{126}$$

As a result, by applying Lemma 63, for any $\delta \in (0,1)$, if $\min\big\{g^{\pi_1,h}(h), g^{\pi_2,h}(h) \ge 173\log\frac{8}{\delta}\big\}$, 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

$$\begin{aligned}
&\Big|\big|R_T^{\pi_1,h} - R_T^{\pi_2,h}\big| - \big|\big[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\big] - \big[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\big]\big|\Big| \\
\le&\Big|R_T^{\pi_1,h} - \big[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\big]\Big| + \Big|R_T^{\pi_2,h} - \big[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\big]\Big| \\
\le&\max\left\{\bar{K}_T^{\pi_1,h}\sqrt{3g^{\pi_1,h}(T)\big(2\log\log(\frac{3}{2}g^{\pi_1,h}(T)) + \log\frac{4}{\delta}\big)}, (\bar{K}_T^{\pi_1,h})^2\right\} \\
&+\max\left\{\bar{K}_T^{\pi_2,h}\sqrt{3g^{\pi_2,h}(T)\big(2\log\log(\frac{3}{2}g^{\pi_2,h}(T)) + \log\frac{4}{\delta}\big)}, (\bar{K}_T^{\pi_2,h})^2\right\}.
\end{aligned} \tag{127}$$

Appendix F. Miscellaneous Theorems

F.1 Slutsky's Theorem

Theorem 71 (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. $\dfrac{X_t}{Y_t} \xrightarrow{(d)} \dfrac{X}{c}$, *if* $c \neq 0$.

Remark 72 *Since convergence in the almost-sure sense implies convergence in probability, same results hold when $Y_t \xrightarrow{(a.s.)} c$.*