Learning, Control and Concentration of Cumulative Rewards in MDPs and Markov Jump Systems

Borna Sayedana

Department of Electrical and Computer Engineering
McGill University, Montreal

December 2024

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Abstract

This thesis examines various aspects of integrating learning and control across different families of stochastic systems. This work includes three key aspects: (i) learning the unknown system parameters from input-output data sequences (i.e., system identification), (ii) integrating learning within the control of dynamical systems (i.e., adaptive control), and (iii) providing probabilistic guarantees for the developed methodologies, including regret upper bounds and concentration bounds. The analysis is conducted within three frameworks of stochastic systems: Markov jump linear systems, finite-state and finite-action Markov Decision Processes (MDPs), and linear stochastic systems.

In the framework of Markov jump linear systems, two primary problems are addressed. First, we focus on the full-state observation system identification problem, i.e., learning system parameters from observed sequences of discrete and continuous states. We propose a variant of least squares algorithm called switched least squares. By leveraging classical regression theory, we establish the algorithm's strong consistency and derive its convergence rate. Furthermore, we integrate this algorithm into a certainty-equivalence framework tailored for controlling Markov jump linear systems. By leveraging the convergence rate of switched least squares, a novel regret decomposition, and the concentration properties of martingale difference sequences, we derive a sub-linear regret upper bound for the proposed algorithm.

In the second part of this thesis, 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. The asymptotic results include the law of large numbers, the central limit theorem, and the law of iterated logarithm, while the non-asymptotic results include Azuma-Hoeffding-type inequalities and a non-asymptotic version of the law of iterated logarithm. Using these results, we show that two alternative definitions of regret for learning policies in the literature are rate-equivalent. The proofs rely on a novel martingale decomposition of cumulative reward, properties of the solutions of the policy-evaluation fixed-point equation, and asymptotic and non-asymptotic concentration of martingales.

Finally, the analysis is extended to the case of linear systems, where we establish the asymptotic normality of the cumulative cost induced by the optimal policies in linear quadratic regulators (LQRs). These results address some of the key theoretical questions in integrating learning and control in stochastic systems.

Résumé

Cette thèse examine divers aspects de l'intégration de l'apprentissage et du contrôle dans différentes familles de systèmes stochastiques. Ce travail comprend trois aspects clés : (i) l'apprentissage des paramètres du système à partir de séquences de données d'entrée-sortie (c'est-à-dire, l'identification du système), (ii) l'intégration de l'apprentissage dans le contrôle des systèmes dynamiques (c'est-à-dire, le contrôle adaptatif), et (iii) la fourniture de garanties probabilistes pour les méthodologies développées, y compris les bornes supérieures de regret et les bornes de concentration. L'analyse est réalisée dans trois cadres de systèmes stochastiques : les systèmes linéaires à sauts de Markov, les processus de décision markoviens à états et actions finis (MDPs), et les systèmes linéaires stochastiques.

Dans le cadre des systèmes linéaires à sauts de Markov, deux problèmes principaux sont abordés. Tout d'abord, nous nous concentrons sur le problème d'identification du système avec observation de l'état complet, c'est-à-dire l'apprentissage des paramètres du système à partir de séquences observées d'états discrets et continus. Nous proposons une variante de l'algorithme des moindres carrés appelée "moindres carrés commutés". En nous appuyant sur la théorie classique de la régression, nous établissons la forte consistance de l'algorithme et dérivons son taux de convergence. De plus, nous intégrons cet algorithme dans un cadre d'équivalence de certitude spécifiquement conçu pour le contrôle des systèmes linéaires à sauts de Markov. En utilisant le taux de convergence des moindres carrés commutés, une nouvelle décomposition du regret, et les propriétés de concentration des suites de différences de martingales, nous dérivons une borne supérieure de regret sub-linéaire pour l'algorithme proposé.

Dans la deuxième partie de la thèse, nous investiguons les propriétés de concentration des récompenses cumulées dans les processus de décision markoviens (MDPs), en nous concentrant sur les cadres asymptotiques et non asymptotiques. Nous introduisons une approche unifiée pour caractériser la concentration des récompenses dans les MDPs, couvrant à la fois les cadres à horizon infini (c'est-à-dire les cadres de récompenses moyennes et remises) et à horizon fini. Les résultats asymptotiques incluent la loi des grands nombres, le théorème central limite et la loi des logarithmes itérés, tandis que les résultats non asymptotiques incluent les inégalités de type Azuma-Hoeffding et une version non asymptotique de la loi

des logarithmes itérés. En utilisant ces résultats, nous montrons que deux définitions alternatives du regret pour les politiques d'apprentissage dans la littérature sont équivalentes en termes de taux. Les démonstrations reposent sur une nouvelle décomposition martingale des récompenses cumulées, les propriétés des solutions de l'équation de point fixe d'évaluation de politique, et la concentration asymptotique et non asymptotique des martingales.

Enfin, l'analyse est étendue au cas des systèmes linéaires, où nous établissons la normalité asymptotique des récompenses cumulées induites par les politiques optimales dans les régulateurs quadratiques linéaires (LQRs). Ces résultats répondent à certaines des principales questions théoriques liées à l'intégration de l'apprentissage et du contrôle dans les systèmes stochastiques.

To my parents ...

Acknowledgements

I would like to begin by expressing my sincere gratitude to my advisor, Prof. Aditya Mahajan, and my co-advisor, Prof. Peter Caines, for their outstanding guidance and support throughout my PhD studies. Learning from them, both in their courses and while working on my dissertation, has been an incredible privilege. Their expertise and insights were influential in achieving my research results, and their professionalism and support during challenging times have been incredibly meaningful to me.

I also wish to thank my PhD committee members, Prof. Ioannis Psaromiligkos and Prof. Roland Malhame, for their insightful feedback and suggestions on my work. I would also like to express my gratitude to Prof. Masoud Asgharian, whose lectures had a significant influence on my academic journey. They were key in developing my interest in mathematical statistics, which has since become a central theme of my research.

I would like to express my heartfelt thanks to my dear friend in Montreal, Prof. MohammadHadi Shateri, whose generosity, support, and kindness have been a great source of encouragement throughout my academic journey. I also extend my appreciation to my friend and former collaborator, Dr. Mohammad Afshari, for his guidance and help in the works done during my PhD studies. I am truly thankful to my friend Dr. Alex Dunyak. The time we spent together throughout my studies was invaluable, and his companionship made my PhD journey significantly more enjoyable.

I am deeply grateful to the FRQNT (Fonds de recherche du Québec) for their generous funding through the FRQNT PhD scholarship. This financial support has been crucial in enabling me to focus fully on my research and academic development.

I would also like to express my gratitude to my long-time friends abroad—Pedram Hosseini, Hafez Damanpak, Soheil Haghighat, Parham Shams, and Mehdi Memarzadeh—for their motivation and encouragement.

I would like to express my gratitude to all my current and former lab mates—Nima, Berk, Raihan, Jayakumar, Reihaneh, Amit, Samin Yeasar, Samin, Edwin, and Erfan— who made my research experience enjoyable.

Finally, I would like to express my deepest gratitude to my parents for their unwavering support and encouragement throughout my journey. My father's strength and patience, cou-

pled with my mother's kindness and dedication, have been my primary source of motivation. None of my achievements would have been possible without their constant belief in me. I am profoundly grateful for all the sacrifices they made to ensure my success and for their unending love and support.

Borna Sayedana
Montreal, March 2025

Table of Contents

List of Figures

Figure

List of Abbreviations

a.s.	Almost Surely
AROE	Average Reward Optimality Equation
ARPE	Average Reward Policy Evaluation equation
CLT	Central Limit Theorem
DROE	Discounted Reward Optimality Equation
DRPE	Discounted Reward Policy Evaluation equation
FHDP	Finite-Horizon Dynamic Programming equation
FHPE	Finite-Horizon Policy Evaluation equation
i.o.	Infinitely Often
LHS	Left Hand Side
LIL	Law of Iterated Logarithm
LLN	Law of Large Numbers
LQR	Linear Quadratic Regulators
MDP	Markov Decision Process
MDS	Martingale Difference Sequence
MJS	Markov Jump System
MJLS	Markov Jump Linear Systems
MSS	Mean Square Stable

PDF Probability Density Function

PMF Probability Mass Function

RHS Right Hand Side

RL Reinforcement Learning

SLLN Strong Law of Large Numbers

SLS Switched Linear Systems

Chapter 1

Introduction

1.1 Motivation

Achieving fully autonomous agents has been a long-standing objective in engineering and computer science. However, realizing full autonomy requires the integration of learning algorithms into the control of dynamical systems. These systems must be capable of learning from their evolving environments and adapting accordingly. A key challenge is integrating adaptive learning algorithms into the control framework along with providing theoretical guarantees. This issue has been widely explored in fields such as system identification, adaptive control, and reinforcement learning (RL) [1–3]. Despite significant progress across these areas, integrating adaptive learning mechanisms into control of dynamical systems remains a challenging problem both theoretically and practically.

Physical engineering systems operate under various constraints. Stability is a fundamental requirement for most control systems, while other constraints include energy consumption and hard limits on physical parameters [4]. As a result, it is often necessary for controllers to provide certifiable guarantees that satisfy these constraints. While classical control literature offers a robust set of methodologies with guarantees for stability, robustness, and sensitivity, deriving similar guarantees becomes significantly more challenging when adaptive learning mechanisms are integrated into control systems. Moreover, the classical literature on adaptive control has primarily focused on ensuring asymptotic stability and asymptotic optimality of the resulting controllers (see [1] for an overview of these results); however, these guarantees often prove insufficient to meet the requirements of many practical applications. To better evaluate the performance of RL and adaptive controllers, new metrics have been introduced in the literature.

1.2 Modern Approaches

In RL and adaptive control literature, regret has been introduced as a key metric to evaluate the quality of an adaptive algorithm. Regret quantifies the cumulative difference between the cost or reward yielded by a given algorithm and the performance of the optimal policy. An adaptive algorithm shows a better performance if it incurs smaller regret. Treating regret as a primary objective, substantial effort has been devoted in the literature to developing reinforcement learning algorithms and proving that they achieve sub-linear regret (e.g. [5–25]). The majority of these results are established for finite-state and finite-action Markov Decision Processes (MDPs). While finite-state and finite-action MDP framework is widely applicable across different domains, it is not suitable for modeling all control systems. In many practical systems, states and actions are continuous and do not belong to a compact set. As a result, learning strategies designed for discrete spaces are not applicable in these setups. Furthermore, the non-compactness of these spaces makes the notion of stability critical, as the resulting policy could potentially destabilize the system. This has driven significant research toward RL in a different modeling framework that captures these features, known as Linear Quadratic Regulators (LQRs). Recently there has been a keen interest in the problem of adaptive learning and control of these systems and providing regret upper-bounds (e.g. [26–35]). In this model, both states and actions are assumed to be continuous variables, the system dynamics are assumed to be linear, and the controller's goal is to minimize a cost function quadratic in states and actions.

1.3 Categorizing the Literature

The literature on RL with providing regret guarantees can be categorized in various ways. One key distinction between different works lies in the probabilistic setup used to define regret. Regret is either computed in the Bayesian setup or frequentist setup. In the Bayesian setup, a prior distribution is assumed over the unknown system parameters, and the expected regret is calculated with respect to this prior (e.g. [10, 13, 33]). In contrast, in the frequentist setup it is assumed that there exists a fixed, true, unknown system parameter, and regret is measured with respect to the optimal policy derived from that parameter (e.g. [26]). Furthermore, the definition of regret itself varies across studies—some compare the performance of learning policies to the optimal policy on individual sample trajectories (e.g. [31, 36]), while others evaluate the performance against the optimal policy averaged over all trajectories (e.g. [6, 26, 37]). Depending on the definitions, regret may be a deterministic function of time or a random process. In the latter case, two types of guaran-

tees are commonly studied: (i) almost-sure asymptotic guarantees and (ii) non-asymptotic high-probability guarantees. The choice between these two often depends on the application. Asymptotic bounds (e.g. [31]) provide stronger guarantees for long-term performance but lack characterization in finite-time, which can be critical for certain applications. Non-asymptotic bounds (e.g. [6, 8, 26]) characterize the finite-time behavior of the policy with high probability but may fail to provide long-term guarantees. These different approaches and results show the richness of the field and the need for further research to unify these approaches and better address practical requirements.

In both RL and adaptive control, algorithms are generally divided into two categories: model-free and model-based. Model-free algorithms aim to directly learn a policy through sequential interactions with the environment (e.g. [32]). In contrast, model-based algorithms focus on estimating the parameters of the underlying system and subsequently using these estimated parameters to compute an efficient policy. In the latter case, accurately estimating system parameters from input-output data becomes a critical task, commonly referred to as system identification. The system identification problem has been extensively studied in the control literature for both linear and nonlinear systems (see [1, 2] for an overview of classical results). Often, there are two requirements for any system identification method. First, the method should be consistent, i.e., the estimator's value converges to the true parameter's value under an appropriate probabilistic notion of convergence. Second, the estimator should exhibit a fast rate of convergence, ensuring that the estimated value approaches to the true parameter efficiently. Similar to the regret analysis, the guarantees provided for system identification methods can be asymptotic (e.g. [38]) or non-asymptotic (e.g. [39]), and the choice between these guarantees depends on practical requirements.

1.4 Performance Guarantees

The literature on providing regret guarantees derive sample-path or distributional guarantees for adaptive algorithms in terms of the regret. However, providing similar guarantees for a fixed policy in the planning setup remains a remarkably under-explored problem. In many practical applications, safety and robustness requirements necessitate a focus not only on the mean behavior but also on the concentration behavior of the cumulative cost or reward. In these applications, concentration bounds can serve as certifiable guarantees for hard constraints on the controller. Furthermore, the concentration behavior of policies in the planning setup is closely related to that of policies in the learning setup. This connection highlights the importance of understanding concentration behaviors for integrating adaptive learning methods into control systems while maintaining performance guarantees.

1.5 Investigated Problems

A practical feature not captured by both MDP and LQR frameworks is time-varying dynamics. In many applications, system dynamics evolve either gradually or abruptly. To model abrupt changes, several mathematical frameworks for switching systems have been proposed in the literature. These switches may depend on states, actions, or occur randomly. Examples of such frameworks include hybrid systems, switched linear systems, Markov Jump Systems (MJS), and control affine systems. In these models, the learning and control problems become significantly more challenging due to the increased complexity.

In this thesis, we aim to address some of the challenges mentioned earlier in integrating learning into control systems. The problems explored in this work are divided into two main sections.

In the first section, we examine two key problems: system identification and reinforcement learning in the framework of Markov Jump Systems. We think that the MJS framework serves as a suitable transitional step for extending RL algorithms and analyses from linear systems to more general time-varying systems. For the system identification problem, we propose a variant of the least squares algorithm and establish its strong consistency and almost-sure rate of convergence. Building on this, we use the proposed algorithm within a model-based framework, known as the certainty-equivalence scheme, to design an RL algorithm for MJS along with almost-sure regret guarantees.

In the second section, we focus on the concentration properties of cumulative rewards in infinite-horizon and finite-horizon MDP frameworks. This includes deriving both asymptotic and non-asymptotic concentration results. While this problem is of independent interest in the planning setup, we also explore their implications in the learning setup. Specifically, we demonstrate that two distinct expressions for regret are rate-equivalent. Finally, we extend the methods developed in this section to linear systems, deriving concentration results for the cumulative cost induced by the optimal policy within the LQR framework.

A detailed description of each chapter is provided in the subsequent sub-sections.

1.5.1 Chapter 2

In this chapter, we investigate the problem of system identification in the Markov jump linear systems. Markov jump linear system (MJS) is an extension of the linear-time invariant framework to systems with abrupt changes in their dynamics. These systems find applications in various domains such as networked control systems [40] and cyber-physical systems [41, 42]. There is a rich literature on the stability analysis (e.g., [43–45]) and optimal control (e.g., [46]) of MJS. However, most of the literature assumes that the system model is

known. The question of system identification, i.e., identifying the dynamics from data, has not received much attention in this setup.

The problem of identifying the system model from data is a key component for control synthesis for both offline control methods and online control methods including adaptive control and reinforcement learning [2, 47]. There are four main approaches for system identification of linear systems: (i) maximum likelihood estimation which maximizes the likelihood function of the unknown parameter given the observation (e.g. see [48]); (ii) minimum prediction error methods which minimize the estimation error (residual process) according to some loss function (e.g. see [38, 49]); (iii) subspace methods, which find a minimum state space realization given the input, output data (e.g. see [50, 51]); (iv) least squares method which estimates the unknown parameter by considering the model as a regression problem (e.g. see [1, 52]).

These methods differ in terms of structural assumptions on the model (e.g., system order), hypotheses on the stochastic process, and convergence properties and guarantees.

Structural assumptions require the system to be stable in some sense (e.g., mean square stable, exponentially stable, etc.), and stochastic hypotheses restrict the noise processes to be of a certain type, (e.g., Gaussian, sub-Gaussian, or martingale difference sequences).

Convergence properties characterize the asymptotic behavior of system identification methods. The basic requirements for any system identification method is its consistency, asymptotic normality and rates of convergence, that is to establish that estimates converge asymptotically to the true unknown parameter and characterize the rate of convergence. System identification methods can be *weakly* consistent (i.e., estimates converge in probability) or *strongly* consistent (i.e., estimates convergence almost surely). For linear systems, there is a vast literature that establishes the consistency and rates of convergence for a variety of methods (e.g. see [1, 2] for a unified overview). Another characterization of the convergence is finite-time guarantees which provide lower-bounds on the number of samples required so that estimates have a specified degree of accuracy with a specified high probability (e.g [39]). As the number of samples grow to infinity, these results establish weak consistency of the proposed methods.

A modeling framework closely related to Markov jump systems is the general hybrid systems framework. There is extensive literature on the system identification of various subfamilies of hybrid systems. These models include switched auto-regressive models, piecewise affine systems, switched affine systems, etc. The problem of system identification in these systems under various modeling assumptions has been investigated in the literature (e.g., [53–63]). For a comprehensive review of these results, see [64].

System identification of MJS and switched linear systems (SLS) has received less attention in the literature. There is some work on designing asymptotically stable controllers for unknown SLS [65–67] but these papers do not establish rates of convergence for system identification. There are some recent papers which provide finite-time guarantees and rate of convergence for SLS [68–70], MJS [71], and bilinear systems [72]. System identification of a globally asymptotically stable SLS with controlled switching signal is investigated in [69], while the system identification of an unknown order SLS using subspace methods is investigated in [70]. Both these methods are developed for SLS and are not directly applicable to MJS. The model analyzed in [71] is an MJS system. Under the assumption that the system is mean square stable, the switching distribution is ergodic and the noise is i.i.d. subgaussian, it is established that the convergence rate is $\mathcal{O}(\sqrt{\log T/T})$ with high probability, where T denotes the number of samples. Then a certainty equivalence control algorithm is proposed and its regret is analyzed. Note that if we let the number of samples go to infinity, these results imply *weak* consistency of the proposed methods for MJS systems. As far as we are aware, there is no existing result which establishes *strong* consistency of a method for system identification of MJS.

The contributions of Chapter 2 are summarized as following:

- We propose *switched* least squares method for system identification of an unknown (autonomous) MJS and provide data-dependent and data-independent rates of convergence for this method.

- Our assumptions on the noise and stability of the system are weaker than those imposed in parallel works. We assume noise is a martingale difference process with finite $\alpha > 2$ moment. For the stability, we introduce the notion of stability in the average sense for the MJS systems, and assume the system is stable in the average sense. Under these assumptions, we prove *strong* consistency of the switched least squares method along with $\mathcal{O}(\sqrt{\log(T)/T})$ rate of convergence. In contrast to the existing high-probability convergence guarantees in the literature, our results show that the estimates converge to the true parameters *almost surely*.

- We highlight the technical difficulties that arise in system identification of MJS systems (compared to non-switched systems) and their interplay with stability of the systems. We show how the notion of stability in the average sense circumvents these difficulties.

- We establish that stability in the average sense is a weaker notion of stability compared to the commonly imposed assumptions in the literature. In particular, we show that if a system is mean square stable, then the system is stable in the average sense.

Furthermore, we show that the spectral conditions imposed in literature as a sufficient condition for almost sure stability also imply stability in the average sense. As a consequence, our results are applicable to broader families of the MJS systems investigated in the literature including mean square stable systems.

1.5.2 Chapter 3

In this chapter, we investigate the problem of reinforcement learning in the Markov jump linear systems framework. The main goal of reinforcement learning and adaptive control is simultaneous learning and control of unknown dynamical systems. Due to continuity and unboundedness of the state and action spaces in control setups, classical reinforcement learning algorithms do not achieve good performance. Recently, there has been a surge of interest in designing reinforcement learning algorithms for linear quadratic regulators (LQR) and analyzing the performance of these algorithms [26, 27, 29, 31, 34, 35, 73]. These results exploit the linearity, time-invariancy, and structure of the cost function in the proposed algorithms and analysis.

Markov jump systems are a mathematical formulation which model time-varying dynamical systems with abrupt and stochastic changes in the dynamics. These systems find application in cyber-physical system [40], networked control systems [41, 42], etc. In this chapter, we investigate the problem of simultaneous learning and controlling an unknown Markov jump linear system (MJLS). We use the switched least squares method proposed in Chapter 2 in the closed-loop setup for the system identification and use the system estimates in a certainty equivalence controller.

The problem of learning and controlling MJLS systems has recently received some attention in the literature. The sensitivity analysis of certainty equivalence controller to the system parameter is investigated in [74]. Based on the results of [74], a system identification algorithm and a certainty equivalence controller is proposed in [71] where it is shown that the proposed method achieves the regret of $\tilde{\mathcal{O}}(\sqrt{T})$ with high probability, where T denotes the time horizon, and notation $\tilde{O}$ hides logarithmic factors of T. It is shown in [75] that the policy gradient method converges to the optimal policy for MJLS systems. The performance of Thompson sampling algorithm in controlling networked control systems as a special case of switched linear systems is investigated in [68].

The contributions of Chapter 3 are summarized as following:

- We characterize the almost sure (relative to a certain subset of the noise process and the algorithm randomization) regret bounds for general class of linear adaptive policies.
- We use switched least squares method for closed-loop system identification of MJLS

systems, show that this method is strongly consistent, and establish that its rate of convergence is $\mathcal{O}(\sqrt{\log(T)/T})$.

- We propose a version of certainty equivalence controller based on the switched least squares system identification method and show that this algorithm achieves a regret of $\mathcal{O}(\sqrt{T}\log(T))$ relative to a certain subset of the sample space.

- We show that there exists a finite identification horizon T_0 for which this algorithm achieves the almost sure regret of $\mathcal{O}(\sqrt{T}\log(T))$ on the entire sample space.

1.5.3 Chapter 4

In this chapter, we investigate the concentration properties of cumulative reward in Markov decision processes. 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 dynamics that depends on the current state and action. In the MDP setup, the main focus is on maximizing the expected cumulative rewards (aka., return) [3]. 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 ([76]), constrained MDPs ([77, 78]), and distributional reinforcement learning ([79, 80]).

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 [81]. 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 ([82–84]).

In this chapter, 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 $\tilde{\mathcal{O}}(D|\mathcal{S}|\sqrt{|\mathcal{A}|T})$ [5], 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 the MDP. Various refinements of these results have been considered in the literature [6–25].

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 [16, 36]. 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 [16].

The contributions of Chapter 4 are summarized as follows:

- 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.

- 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 station-

ary 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.

- 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.

- 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 the 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.

1.5.4 Chapter 5

In this chapter, we investigate the asymptotic concentration of cumulative cost induced by the optimal policy in linear quadratic regulators. In Chapter 4, our results are restricted to the case of finite-state and finite-action MDPs. In this chapter, we extend the results of Chapter 4 to the case of Linear Quadratic Regulators in which the state and action do not belong to a compact set. In this chapter, we prove a central limit theorem for the cumulative cost. The Central Limit Theorem (CLT), is one of the most important asymptotic concentration results in probability theory and mathematical statistics. It establishes that the distribution of deviation from the mean in the law of large numbers asymptotically converges to a normal distribution. Similar asymptotic normality for the deviations emerges in other processes as well. For example, in the parameter estimation framework, the asymptotic normality is established for maximum likelihood estimation (see e.g. [85–87]). In regression models, asymptotic normality is established for various estimation and prediction methods (see e.g. [88–93], for a list of such results, see [1]). This property is also established in the stochastic approximation framework (see e.g. [94, 95]). The importance of asymptotic nor-

mality results become evident when they are used to derive confidence bounds for different frameworks.

In the systems and controls literature, there are various characterization of the law of large numbers (e.g. [83, 84, 96–100]) but the distribution of the deviation from the mean is less explored. There are some results on CLT for Markov cost/reward process (e.g. [83, 84, 99, 100]) which are derived using advanced tools in Markov chain theory including weighted geometric ergodicity and weighted uniform ergodicity. These results imply a CLT for the LQR setting. In this chapter, we revisit the distribution of the deviation from the mean for LQR setting and establish asymptotic normality using an elementary proof based on first principles. Our result is different from the existing characterizations in the literature and uses different and much simpler proof techniques.

The sample path behavior of the cumulative cost has recently also been studied in the context of regret analysis for adaptive controllers. These analyses are either in the Bayesian framework (e.g., in [13, 68]) or in terms of high probability guarantees for the frequentist regret (e.g., in [26, 28, 29, 34, 73, 101–103]) or almost sure guarantees for the frequentist regret (e.g., in [31, 104]). However, these bounds are not sharp enough to characterize the distribution of the cumulative cost.

The main contribution of Chapter 5 is to establish asymptotic normality of the cumulative cost in the LQR framework using an elementary argument. Under a mild technical assumption on the noise distribution, we show the cumulative cost incurred by the optimal policy converges weakly to a Gaussian distribution. Our analysis uses a completion of square argument to decompose the cumulative cost to bounded terms plus a Martingale Difference Sequence (MDS). The convergence argument follows from this decomposition, properties of the noise sequence with even density, and a version of the CLT for MDS.

1.5.5 Chapter 6

This chapter includes the concluding remarks and future research directions related to this thesis.

1.6 List of Publications

The following articles are resulting from the work presented in this thesis.

Published Journal Papers:

- B. Sayedana, M. Afshari, P. E. Caines, and A. Mahajan, "Strong consistency and

rate of convergence of switched least squares system identification for autonomous Markov jump linear systems", *IEEE Transactions on Automatic Control*, vol. 69, no. 6, pp. 3952–3959, 2024. DOI: 10.1109/TAC.2024.3351806

Submitted Journal Papers:

- B. Sayedana, P. Caines, A. Mahajan "Concentration of Cumulative rewards in Markov Decision Process", arXiv preprint arXiv:2411.18551, 2024 (Submitted to the Journal of Machine Learning Research (JMLR))

Published Peer-Reviewed Conference Papers

- B. Sayedana, P. E. Caines, and A. Mahajan, "Asymptotic normality of cumulative cost in linear quadratic regulators", in *2024 IEEE 63rd Conference on Decision and Control (CDC)*, 2024, pp. 1856–1862. DOI: 10.1109/CDC56724.2024.10886506

- B. Sayedana, M. Afshari, P. E. Caines, and A. Mahajan, "Relative almost sure regret bounds for certainty equivalence control of Markov jump systems", in *2023 IEEE 62nd Conference on Decision and Control (CDC)*, 2023, pp. 6629–6634. DOI: 10.1109/CDC49753.2023.10383246

- B. Sayedana, M. Afshari, P. E. Caines, and A. Mahajan, "Consistency and rate of convergence of switched least squares system identification for autonomous Markov jump linear systems", in *2022 IEEE 61st Conference on Decision and Control (CDC)*, 2022, pp. 6678–6685. DOI: 10.1109/CDC51059.2022.9993169

- B. Sayedana, M. Afshari, P. E. Caines, and A. Mahajan, "Thompson-sampling based reinforcement learning for networked control of unknown linear systems", in *2022 IEEE 61st Conference on Decision and Control (CDC)*, 2022, pp. 723–730. DOI: 10.1109/CDC51059.2022.9992565

Published Abstracts:

- B. Sayedana, M. Afshari, P. Caines, A. Mahajan "Consistency and Rate of Convergence of Switched Least Squares System Identification for Autonomous Switched Linear Systems", Multidisciplinary Conference on Reinforcement Learning and Decision Making, Brown, Providence, RI, USA, June 8-11th, 2022.

- B. Sayedana, M. Afshari, P. Caines, A. Mahajan "Relative Almost Sure Regret Bounds for Certainty Equivalence Control of Markov Jump Linear Systems", Canadian Mathematical Society Winter Meeting, Montreal, QC, CA, Dec 2023.

- B. Sayedana, M. Afshari, P. Caines, A. Mahajan "Almost Sure Regret Bounds for Certainty Equivalence Control of Markov Jump Linear Systems", 9th Meeting on Systems and Control Theory (MSCT), Waterloo, ON, May 2023.

- B. Sayedana, M. Afshari, P. Caines, A. Mahajan "Relative Almost Sure Regret Bounds for Certainty Equivalence Control of Markov Jump Linear Systems", Workshop on Confluence of Learning and Control Approaches in Multi-Agent Systems, American Control Conference (ACC), Toronto, ON, July 2024.

- B. Sayedana, P. Caines, A. Mahajan "A Tale of Two Regrets: Regret Equivalence via Sample Path Properties of the Optimal Policy". Montreal AI Symposium, Montreal, QC, CA, October 2024.

Technical Reports:

- B. Sayedana, M. Afshari, P. Caines, A. Mahajan, "Strong Consistency and Rate of Convergence of Switched Least Squares System Identification for Autonomous Markov Jump Linear Systems". Les Cahiers du Gerad. G-2023-05, 2022

- B. Sayedana, P. Caines, A. Mahajan, "Asymptotic Normality of Cumulative Cost in Linear Quadratic Regulators". Les Cahiers du Gerad. G-2025-07, 2025

1.7 Contributions of co-authors

In all chapters the problem formulation is written by Borna Sayedana with inputs from Aditya Mahajan and Peter Caines. The technical derivations and writing of Chapters 2 and 3 were carried out by Borna Sayedana, with constructive inputs from Mohammad Afshari, Aditya Mahajan, and Peter Caines. The technical derivations and writing of Chapter 4 and 5 were carried out by Borna Sayedana, with constructive inputs from Aditya Mahajan, and Peter Caines.

1.8 Notation

Given a vector v, $v(i)$ denotes its i-th component. Given a matrix A, $A(i,j)$ denotes its (i,j)-th element, $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote the largest and smallest magnitudes of right eigenvalues, $\sigma_{\max}(A) = \sqrt{\lambda_{\max}(A^\intercal A)}$ denotes the spectral norm. For a square matrix Q, $\mathrm{Tr}(Q)$ denotes the trace. When Q is symmetric, $Q \succeq 0$ and $Q \succ 0$ denote that Q is positive semi-definite and positive definite, respectively. For two square matrices, Q_1 and Q_2 of the same dimension, $Q_1 \succeq Q_2$ means $Q_1 - Q_2 \succeq 0$. Given two matrices A and B, $A \otimes B$ denotes the Kronecker product of the two matrices. Given a sequence of vectors $\{x_t\}_{t\in\mathcal{T}}$, $\mathrm{vec}(x_t)_{t\in\mathcal{T}}$ denotes the vector formed by vertically stacking $\{x_t\}_{t\in\mathcal{T}}$.

$\mathbb{R}$ and $\mathbb{N}$ denote the sets of real and natural numbers and $\mathbb{R}_+$ denotes the set of positive real numbers. For a set $\mathcal{T}$, $|\mathcal{T}|$ denotes its cardinality. For a vector x, $\|x\|$ denotes the Euclidean norm. For a matrix A, $\|A\|$ denotes the spectral norm and $\|A\|_\infty$ denotes the element with the largest absolute value. Notation $\mathrm{diag}(A_1, A_2, \ldots, A_n)$ denotes the block diagonal matrix, where the blocks are matrices $A_1, A_2, \ldots, A_n$. $\mathbf{0}$ denotes the zero-vector in the appropriate Euclidean space.

The notation $\lim_{\gamma\uparrow 1}$ means the limit as γ approaches 1 from below. Given a sequence of positive numbers $\{a_t\}_{t\geq 0}$, $a_T = \mathcal{O}(T)$ means that $\limsup_{T\to\infty} a_T/T < \infty$, and $a_T = o(T)$ means that $\limsup_{T\to\infty} a_T/T = 0$. Given a sequence of positive numbers $\{a_t\}_{t\geq 0}$, $a_T \asymp T$ means that $\limsup_{T\to\infty} a_T/T < \infty$, and $\liminf_{T\to\infty} a_T/T > 0$. When describing values that are taken by consecutive variables, for example s_t and s_{t+1}, we use s to denote a generic value of s_t and s_+ to denote a generic values of s_{t+1}.

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 $\alpha > 0$ such that $a_T = \mathcal{O}(\log(T)^\alpha f(T))$. 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 sequence of random variables $\{x_t\}_{t \geq 0}$, $x_{0:t}$ is a short hand for $(x_0, \cdots, x_t)$ and $\sigma(x_{0:t})$ denotes the sigma field generated by random variables $x_{0:t}$. Given a probability space $\{\Omega, \mathcal{F}, \mathbb{P}\}$, Ω denotes the sample space, $\omega \in \Omega$ denotes a generic elementary event, $\mathbb{P}(\cdot)$ denotes the probability measure, $\mathbb{E}[\cdot]$ denotes the expectation operator, and $\mathbb{1}\{\cdot\}$ denotes the indicator of an event. Given a finite set $\mathcal{S}$, $\Delta(\mathcal{S})$ denotes the space of probability measures defined on $\mathcal{S}$. 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 expression almost surely is abbreviated as $a.s.$ and the expression infinitely often is abbreviated as $i.o.$

Remark 1.1. *There are notational inconsistencies across different chapters of the thesis. For example in chapters 2, 3, and 5, x_t denotes the state of the system, consistent with the notation used in the systems and control sub-community. However, in Chapter 4, S_t denotes the state, aligning with the reinforcement learning sub-community. As far as the notation is concerned, each chapter should be read independently.*

Chapter 2

Strong Consistency and Rate of Convergence of Switched Least Squares System Identification for Autonomous Markov Jump Linear Systems

2.1 Overview

In this chapter, we investigate the problem of system identification in Markov Jump Linear Systems (MJLS). The results of this chapter are published in [105, 108].

2.1.1 Organization

This chapter is organized as follows. In Sec 2.2, we present the system model, assumptions, and the main results. In Sec. 2.3, we prove the main results. In Sec. 2.4, we explain the connection of stability in the average sense with mean square stability and almost sure stability. We present an illustrative example in Sec. 2.5. We conclude in Sec. 2.6.

2.2 System model and problem formulation

Consider a discrete-time (autonomous) MJS. The state of the system has two components: a discrete component $s_t \in \mathcal{S} \coloneqq \{1, \ldots, k\}$ and a continuous component $x_t \in \mathbb{R}^n$. There is a finite set $\mathcal{A} = \{A_1, \ldots, A_k\}$ of system matrices, where $A_i \in \mathbb{R}^{n \times n}$. The continuous component x_t of the state starts at a fixed value x_0 and the initial discrete state s_0 starts

according to a prior distribution π_0. The continuous state evolves according to:

$$x_{t+1} = A_{s_t} x_t + w_{t+1}, \quad t \geq 0, \tag{2.1}$$

where $\{w_t\}_{t\geq 0}$, $w_t \in \mathbb{R}^n$, is a noise process. The discrete component evolves in a Markovian manner according to a time-homogeneous irreducible and aperiodic transition matrix P, i.e. $\mathbb{P}(s_{t+1} = j | s_t = i) = P_{ij}$.

Let $\pi_t = (\pi_t(1), \ldots, \pi_t(k))$ denote the probability distribution of the discrete state at time t and π_∞ denote the stationary distribution. We assume $\pi_\infty(i) \neq 0$ for all i. Let $\mathcal{F}_t = \sigma(x_{0:t}, s_{0:t})$ denote the sigma-algebra generated by the history of the complete state. It is assumed that the noise process satisfies the following:

Assumption 2.1. *The noise process $\{w_t\}_{t\geq 0}$ is a martingale difference sequence with respect to $\{\mathcal{F}_t\}_{t\geq 0}$, i.e., $\mathbb{E}[\|w_t\|] < \infty$ and $\mathbb{E}[w_{t+1} \mid \mathcal{F}_t] = 0$. Furthermore, there exists a constant $\alpha > 2$ such that $\sup_{t\geq 0} \mathbb{E}[\|w_{t+1}\|^\alpha \mid \mathcal{F}_t] < \infty$ a.s. and there exists a symmetric and positive definite matrix $C \in \mathbb{R}^{n\times n}$ such that $\liminf_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} w_t w_t^\intercal = C$ a.s.*

Assumption 2.1 is a standard assumption in the asymptotic analysis of system identification of linear systems [1, 52, 89, 109, 110] and allows the noise process to be non-stationary and have heavy tails (as long as moment condition is satisfied). We use the following notion of stability for the MJS system (2.1).

Definition 2.1. *The MJS system* (2.1) *is called stable in the* average sense *if almost surely:*

$$\sum_{t=1}^{T} \|x_t\|^2 = \mathcal{O}(T) \quad \textit{i.e.} \quad \limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \|x_t\|^2 < \infty.$$

Assumption 2.2. *The MJS system* (2.1) *is stable in the average sense.*

The notion of stability in the average sense has been used in a few papers in the literature of linear systems [31],[111]; however, in the MJS literature, the commonly used notions of stability are mean square stability and almost sure stability of *noise-free* system. We compare stability in the average sense with both of these notions in Sec. 2.4. Specifically, we show that mean square stability implies stability in the average sense. Moreover, we show a common sufficient condition for almost sure stability of noise-free system implies stability in the average sense for MJS system (2.1). Therefore, the assumption of stability in the average sense is weaker than the commonly imposed stability assumptions in the literature.

2.2.1 System identification and switched least squares estimates

We are interested in the setting where the system dynamics $\mathcal{A}$ and the switching transition matrix P are unknown. Let $\theta^\intercal = [A_1, \dots, A_k] \in \mathbb{R}^{n \times nk}$ denote the unknown parameters of the system dynamics matrices. We consider an agent that observes the complete state (x_t, s_t) of the system at each time and generates an estimate $\hat{\theta}_T$ of θ as a function of the observation history $(x_{0:T}, s_{0:T})$. A commonly used estimate in such settings is the least squares estimate:

$$\hat{\theta}_T^\intercal = \underset{\theta^\intercal = [A_1, \dots, A_k]}{\arg\min} \sum_{t=0}^{T-1} \|x_{t+1} - A_{s_t} x_t\|^2. \tag{2.2}$$

The components $[\hat{A}_{1,T}, \dots, \hat{A}_{k,T}] = \hat{\theta}_T^\intercal$ of the least squares estimate can be computed in a switched manner. Let $\mathcal{T}_{i,T} = \{t \le T \mid s_t = i\}$ denote the time indices until time T when the discrete state of the system equals i. Note that for each $t \in \mathcal{T}_{i,T}$, $A_{s_t} = A_i$. Therefore, we have

$$\hat{A}_{i,T} \coloneqq \underset{A_i \in \mathbb{R}^{n \times n}}{\arg\min} \sum_{t \in \mathcal{T}_{i,T}} \|x_{t+1} - A_i x_t\|^2, \quad \forall i \in \{1, \cdots, k\}. \tag{2.3}$$

Let $X_{i,T}$ denote $\sum_{t \in \mathcal{T}_{i,T}} x_t x_t^\intercal$, which we call the unnormalized empirical covariance of the continuous component of the state at time T when the discrete component equals i. Then, $\hat{A}_{i,T}$ can be computed recursively as follows:

$$\hat{A}_{i,T+1} = \hat{A}_{i,T} + \left[\frac{X_{i,T}^{-1} x_T (x_{T+1} - \hat{A}_{i,T} x_T)^\intercal}{1 + x_T^\intercal X_{i,T}^{-1} x_T} \right] \mathbb{1}\{s_{T+1} = i\}, \tag{2.4}$$

where $X_{i,T}$ may be updated as $X_{i,T+1} = X_{i,T} + \left[x_{T+1} x_{T+1}^\intercal\right] \mathbb{1}\{s_{T+1} = i\}$. Due to the switched nature of the least squares estimate, we refer to above estimation procedure as *switched least squares* system identification.

A common way of estimating the transition matrix P is to use empirical counts, i.e.,

$$\hat{P}_{ij,T} = \frac{\sum_{t=1}^{T} \mathbb{1}(s_{t-1} = i, s_t = j)}{\sum_{t=1}^{T} \mathbb{1}(s_{t-1} = i)}, \quad \forall i, j \in \mathcal{S}.$$

Using [112, Lemma 7] and Borel-Cantelli Lemma, it is straight-forward to show that the empirical estimator $\hat{P}_{ij,T}$ converges almost surely. In particular,

$$\|\hat{P}_{ij,T} - P_{ij}\| \le \mathcal{O}\big(\sqrt{\log^2(T)/T}\big) \quad \text{a.s.}, \quad \forall i, j \in \mathcal{S}.$$

So, in the rest of the chapter, we focus on the convergence of the switched least squares

estimator.

2.2.2 The main results

A fundamental property of any sequential parameter estimation method is strong consistency, which we define below.

Definition 2.2. *An estimator $\hat{\theta}_T$ of parameter θ is called strongly consistent if*

$$\lim_{T\to\infty} \hat{\theta}_T = \theta, \quad a.s.$$

Our main result is to establish that the switched least squares estimator is strongly consistent. We do so by providing two different characterization of the rate of convergence. We first provide a data-dependent rate of convergence which depends on the spectral properties of the unnormalized empirical covariance. We then present a data-independent characterization of rate of convergence which only depends on T. All the proofs are presented in Sec. 2.3.

Theorem 2.1. *Under Assumptions 2.1 and 2.2, the switched least squares estimates $\{\hat{A}_{i,T}\}_{i=1}^{k}$ are strongly consistent, i.e., for each $i \in \mathcal{S}$, we have:*

$$\lim_{T\to\infty} \left\|\hat{A}_{i,T} - A_i\right\|_\infty = 0, \quad a.s.$$

Furthermore, the rate of convergence is upper bounded by:

$$\left\|\hat{A}_{i,T} - A_i\right\|_\infty \leq \mathcal{O}\left(\sqrt{\frac{\log\left[\lambda_{\max}(X_{i,T})\right]}{\lambda_{\min}(X_{i,T})}}\right), \quad a.s.$$

The proof is presented in Sec. 2.3.3.

Remark 2.1. *Theorem 2.1 is not a direct consequence of the decoupling procedure in the switched least squares method. The k least squares problems have a common covariate process $\{x_t\}_{t\geq 1}$. Therefore, the convergence of the switched least squares method and the stability of the MJS are interconnected problems. Our proof techniques carefully use the stability properties of the system to establish the consistency of the system identification method.*

We simplify the result of Theorem 2.1 and characterize the data dependent result of Theorem 2.1 in terms of horizon T and the cardinality of the set $\mathcal{T}_{i,T}$.

Corollary 2.1. *Under Assumptions 2.1 and 2.2, for each $i \in \mathcal{S}$, we have:*

$$\left\|\hat{A}_{i,T} - A_i\right\|_\infty \leq \mathcal{O}\left(\sqrt{\log(T)/|\mathcal{T}_{i,T}|}\right), \quad a.s.$$

Remark 2.2. *The assumption that $\pi_\infty(i) \neq 0$ implies that for sufficiently large T, $|\mathcal{T}_{i,T}| \neq 0$ almost surely, therefore the expressions in above bounds are well defined.*

The result of Corollary 2.1 still depends on data. When system identification results are used for adaptive control or reinforcement learning, it is useful to have a data-independent characterization of the rate of convergence. We present this characterization in the next theorem.

Theorem 2.2. *Under Assumptions 2.1 and 2.2, the rate of convergence of the switched least squares estimator $\hat{A}_{i,T}$, $i \in \mathcal{S}$ is upper-bounded by:*

$$\left\|\hat{A}_{i,T} - A_i\right\|_\infty \leq \mathcal{O}\big(\sqrt{\log(T)/\pi_\infty(i)T}\big), \quad a.s.$$

where the constants in the $\mathcal{O}(\cdot)$ notation do not depend on Markov chain $\{s_t\}_{t\geq 0}$ and horizon T. Therefore, the estimation process $\{\hat{\theta}_T\}_{T\geq 1}$ is strongly consistent, i.e., $\lim_{T\to\infty}\left\|\hat{\theta}_T - \theta\right\|_\infty = 0$ a.s. Furthermore, the rate of convergence is upper bounded by:

$$\left\|\hat{\theta}_T - \theta\right\|_\infty \leq \mathcal{O}\big(\sqrt{\log(T)/\pi^* T}\big), \quad a.s.$$

where $\pi^* = \min_{j\in\mathcal{S}} \pi_\infty(j)$.

The proof is presented in Section 2.3.5.

Theorem 2.2 shows that Assumptions 2.1 and 2.2 guarantee that the switched least squares estimator for MJS has the same rate of convergence of $\mathcal{O}(\sqrt{\log(T)/T})$ as non-switched case established in [52]. Moreover, the upper bound in Theorem 2.2 shows that the estimation error of $\hat{A}_{i,T}$ is proportional to $1/\sqrt{\pi_\infty(i)}$; therefore, the rate of convergence of $\hat{\theta}_T$ is proportional to $1/\sqrt{\pi^*}$, where π^* is the smallest probability in the stationary distribution π_∞.

Remark 2.3. *Switched Linear System (SLS) is a special case of MJS in which the discrete state evolves in an i.i.d. manner. The results presented in this section are valid for the SLS after substituting stationary distribution π_∞ with the i.i.d. Probability Mass Function (PMF) of switching probabilities defined over discrete state.*

2.3 Proofs of the main results

2.3.1 Preliminary results

We first state the Strong Law of Large Numbers (SLLN) for Martingale Difference Sequences (MDS) .

Theorem 2.3. *(see [113, Theorem 3.3.1]) Suppose $\{X_\tau\}_{\tau\geq 1}$is a martingale difference sequence with respect to the filtration $\{\mathcal{F}_\tau\}_{\tau\geq 1}$. Let a_τ be $\mathcal{F}_{\tau-1}$ measurable for each $\tau \geq 1$ such that*

$$0 < a_\tau \to \infty \quad as \quad \tau \to \infty, \quad a.s.$$

If for some $p \in (0, 2]$, we have:

$$\sum_{\tau=1}^{\infty} \mathbb{E}[|X_\tau|^p | \mathcal{F}_{\tau-1}]/a_\tau^p < \infty,$$

then:

$$\lim_{T\to\infty} \sum_{\tau=1}^{T} X_\tau/a_T = 0 \quad a.s.$$

Lemma 2.1. *The assumptions on the process $\{s_t\}_{t\geq 0}$ imply that $\lim_{T\to\infty} |\mathcal{T}_{i,T}|/T = \pi_\infty(i)$, a.s.*

Proof. $\{s_t\}_{t\geq 0}$ is an aperiodic and irreducible Markov chain, hence, by the Ergodic Theorem (Theorem 4.1, [114]), $\{s_t\}_{t\geq 0}$ is ergodic and therefore $\lim_{T\to\infty} |\mathcal{T}_{i,T}|/T = \pi_\infty(i)$ a.s. □

Lemma 2.2. *Assumption 2.1 and 2.2 imply:*

$$\sum_{\tau=1}^{\infty} \|x_\tau\|^2/\tau^2 < \infty \quad a.s.$$

Proof. The result is a direct consequence of Abel's lemma. Let $S_T \coloneqq \sum_{\tau=1}^{T} \|x_\tau\|^2$, then we have:

$$\begin{aligned}
\sum_{\tau=1}^{T} \frac{\|x_\tau\|^2}{\tau^2} &= \sum_{\tau=1}^{T} \frac{S_\tau - S_{\tau-1}}{\tau^2} \\
&= \frac{S_T}{T^2} - \frac{S_0}{1} + \sum_{\tau=2}^{T} S_{\tau-1}\Big(\frac{1}{(\tau-1)^2} - \frac{1}{\tau^2}\Big) \\
&\overset{(a)}{=} \frac{S_T}{T^2} - \frac{S_0}{1} + \sum_{\tau=2}^{T} \mathcal{O}(\tau-1)\Big(\frac{2\tau-1}{\tau^2(\tau-1)^2}\Big) \\
&= \frac{S_T}{T^2} - \frac{S_0}{1} + \sum_{\tau=2}^{T} \mathcal{O}\Big(\frac{1}{\tau^2}\Big) < \infty,
\end{aligned}$$

where (a) follows from Assumption 2.2. □

Lemma 2.3. *We have the following:*

$$\Big\| \sum_{\tau=1}^{T} A_{s_\tau} x_\tau w_{\tau+1}^\intercal + w_{\tau+1} x_\tau^\intercal A_{s_\tau}^\intercal \Big\| = o(T) \quad a.s.$$

Proof. We prove the limit element-wise. The (l,p)-th element of the matrix $A_{s_\tau} x_\tau w_{\tau+1}^\intercal$ is $\Big[\sum_{j=1}^{n} A_{s_\tau}(l,j) x_\tau(j)\Big] w_{\tau+1}(p)$.

We calculate the term:

$$\mathbb{E}\Big[\Big(\sum_{j=1}^{n} A_{s_\tau}(l,j) x_\tau(j) w_{\tau+1}(p)\Big)^2 \Big| \mathcal{F}_\tau\Big]. \tag{2.5}$$

Let $A_* = \max_{i\in\mathcal{S}} \|A_i\|_\infty$, then

$$\begin{aligned}
&\mathbb{E}\Big[\Big(\sum_{j=1}^{n} A_{s_\tau}(l,j) x_\tau(j)\Big)^2 w_{\tau+1}^2(p) \Big| \mathcal{F}_\tau\Big] \\
&\quad \overset{(a)}{\leq} A_*^2 \sup_\tau \mathbb{E}[w_{\tau+1}^2(p)\big|\mathcal{F}_\tau] \Big(\sum_{j=1}^{n} x_\tau(j)\Big)^2 \\
&\quad \overset{(b)}{\leq} n A_*^2 \sup_\tau \mathbb{E}\big[w_{\tau+1}^2(p)\big|\mathcal{F}_\tau\big] \|x_\tau\|^2,
\end{aligned}$$

where (a) uses the fact that s_τ and x_τ are $\mathcal{F}_\tau-$ measurable and that $|A_{s_\tau}(l,j)| \leq A_*$ and (b) is by Cauchy-Schwarz's inequality. Therefore:

$$\begin{aligned}
&\sum_{\tau=1}^{T} \frac{\mathbb{E}\Big[\Big(\big[\sum_{j=1}^{n} A_{s_\tau}(l,j) x_\tau(j)\big] w_{\tau+1}(p)\Big)^2 \Big| \mathcal{F}_\tau\Big]}{\tau^2} \\
&\qquad \leq n A_*^2 \sup_\tau \Big\{\mathbb{E}[w_{\tau+1}^2(p)|\mathcal{F}_\tau]\Big\} \sum_{\tau=1}^{T} \frac{\|x_\tau\|^2}{\tau^2} \overset{(c)}{\leq} \infty.
\end{aligned}$$

Since $\alpha > 2$ in Assumption 2.1, and finiteness of higher order moments imply finiteness of lower order moments, we get $\mathbb{E}\big[w_{\tau+1}^2(p)\big|\mathcal{F}_\tau\big]$ is uniformly bounded. This fact along with Lemma 2.2 imply (c). The result then follows by applying Theorem 2.3 by setting $a_t = t$ and $p = 2$. □

We characterize the asymptotic behavior of the matrix $X_{i,T}$.

Proposition 2.1. *Under Assumptions 2.1 and 2.2, the following hold a.s. for each $i \in \mathcal{S}$:*
(**P1**) $\lambda_{\max}(X_{i,T}) = \mathcal{O}(T)$, *a.s.*
(**P2**) $\liminf_{T\to\infty} \lambda_{\min}(X_{i,T})/|\mathcal{T}_{i,T}| > 0$, *a.s.*

Remark 2.4. *Property (P1) shows that when the system is stable in the average sense, $\lambda_{\max}(X_{i,T})$ cannot grow faster than linearly with time. Therefore, the stability of the system controls the rate at which $X_{i,T}$ can grow. Property (P2) shows that when the noise has a minimum covariance, $\lambda_{\min}(X_{i,T})$ cannot grow slower than linearly with time.*

Proof of (P1). The maximum eigenvalue of a matrix can be upper bounded as follows:

$$\lambda_{\max}\Big(\sum_{t\in\mathcal{T}_{i,T}} x_t x_t^\intercal\Big) \overset{(a)}{\leq} \mathrm{Tr}\Big(\sum_{t\in\mathcal{T}_{i,T}} x_t x_t^\intercal\Big) = \sum_{t\in\mathcal{T}_{i,T}} \|x_t\|^2 \leq \sum_{t=1}^{T} \|x_t\|^2 = \mathcal{O}(T),$$

where (a) follows from the fact that trace of a matrix is sum of its eigenvalues and all eigenvalues of $x_t x_t^\intercal$ are non-negative. □

Proof of (P2). For $\tau \geq 1$, we have:

$$\begin{aligned} x_\tau x_\tau^\intercal =& (A_{s_{\tau-1}} x_{\tau-1} + w_\tau)(A_{s_{\tau-1}} x_{\tau-1} + w_\tau)^\intercal \\ =& A_{s_{\tau-1}} x_{\tau-1} x_{\tau-1}^\intercal A_{s_{\tau-1}}^\intercal \\ &+ A_{s_{\tau-1}} x_{\tau-1} w_\tau^\intercal + w_\tau x_{\tau-1}^\intercal A_{s_{\tau-1}}^\intercal + w_\tau w_\tau^\intercal. \end{aligned}$$

Since $A_{s_{\tau-1}} x_{\tau-1} x_{\tau-1}^\intercal A_{s_{\tau-1}}^\intercal$ is positive semi-definite, we have:

$$x_\tau x_\tau^\intercal \succeq A_{s_{\tau-1}} x_{\tau-1} w_\tau^\intercal + w_\tau x_{\tau-1}^\intercal A_{s_{\tau-1}}^\intercal + w_\tau w_\tau^\intercal.$$

By summing over $\tau \in \mathcal{T}_{i,T}$, we get:

$$\begin{aligned} \sum_{\tau\in\mathcal{T}_{i,T}} x_\tau x_\tau^\intercal \succeq & \sum_{\tau\in\mathcal{T}_{i,T}} w_\tau w_\tau^\intercal + x_0 x_0^\intercal \mathbb{1}\{s_0 = i\} \\ &+ \sum_{\tau\in\mathcal{T}_{i,T}} \left[A_{s_{\tau-1}} x_{\tau-1} w_\tau^\intercal + w_\tau x_{\tau-1}^\intercal A_{s_{\tau-1}}^\intercal\right] \\ &\overset{(a)}{\succeq} \sum_{\tau\in\mathcal{T}_{i,T}} w_\tau w_\tau^\intercal + o(T) \quad \text{a.s.}, \end{aligned}$$

where (a) follows from Lemma 2.3 and $x_0 x_0^\intercal \mathbb{1}\{s_0 = i\} \succeq 0$. Furthermore, since $\lim_{T\to\infty} |\mathcal{T}_{i,T}|/T =$

$\pi_\infty(i)$ a.s. by Lemma 2.1 and $\pi_\infty(i) \neq 0$ by assumptions on $\{s_\tau\}_{\tau\geq 0}$, we have:

$$\liminf_{|\mathcal{T}_{i,T}|\to\infty} \frac{\sum_{\tau\in\mathcal{T}_{i,T}} x_\tau x_\tau^\intercal}{|\mathcal{T}_{i,T}|} \succeq \liminf_{|\mathcal{T}_{i,T}|\to\infty} \frac{\sum_{\tau\in\mathcal{T}_{i,T}} w_\tau w_\tau^\intercal}{|\mathcal{T}_{i,T}|} \overset{(b)}{=} C \succ 0 \quad \text{a.s.},$$

where (b) holds by Assumption 2.1 and independence of $\{w_\tau\}_{\tau\geq 0}$ and $\{s_\tau\}_{\tau\geq 0}$ processes. Therefore

$$\liminf_{|\mathcal{T}_{i,T}|\to\infty} \lambda_{\min}\Big(\frac{\sum_{\tau\in\mathcal{T}_{i,T}} x_\tau x_\tau^\intercal}{|\mathcal{T}_{i,T}|}\Big) \succ 0.$$

□

2.3.2 Background on least square estimator

Given a filtration $\{\mathcal{G}_t\}_{t\geq 0}$, consider the following regression model:

$$y_t = \beta^\intercal z_t + w_t, \quad t \geq 0, \tag{2.6}$$

where $\beta \in \mathbb{R}^n$ is an unknown parameter, $z_t \in \mathbb{R}^n$ is $\mathcal{G}_{t-1}$-measurable covariate process, y_t is the observation process, and $w_t \in \mathbb{R}$ is a noise process satisfying Assumption 2.1 with $\mathcal{F}_t$ replaced by $\mathcal{G}_t$. Then the least squares estimate $\hat{\beta}_T$ of β is given by:

$$\hat{\beta}_T = \arg\min_{\beta^\intercal} \sum_{\tau=0}^{T} \|y_\tau - \beta^\intercal z_\tau\|^2. \tag{2.7}$$

The following result by [89] characterizes the rate of convergence of $\hat{\beta}_T$ to β in terms of unnormalized covariance matrix of covariates $Z_T \coloneqq \sum_{\tau=0}^{T} z_\tau z_\tau^\intercal$.

Theorem 2.4 (see [89, Theorem 1])**.** *Suppose the following conditions are satisfied:* **(S1)** $\lambda_{\min}(Z_T) \to \infty$*, a.s. and* **(S2)** $\log(\lambda_{\max}(Z_T)) = o(\lambda_{\min}(Z_T))$*, a.s. Then the least squares estimate in* (2.7) *is strongly consistent with the rate of convergence:*

$$\|\hat{\beta}_T - \beta\|_\infty = \mathcal{O}\bigg(\sqrt{\frac{\log\big[\lambda_{\max}(Z_T)\big]}{\lambda_{\min}(Z_T)}}\bigg) \quad a.s.$$

Theorem 2.4 is valid for all the $\mathcal{G}_{t-1}$-measurable covariate processes $\{z_t\}_{t\geq 0}$. For the switched least squares system identification if we take $\mathcal{G}_t$ to be equal to $\mathcal{F}_t$ and verify conditions (S1) and (S2) in Theorem 2.4, then we can use Theorem 2.4 to establish its strong consistency and rate of convergence. As mentioned earlier in Remark 2.1, the empirical covariances are coupled across different components due to the system dynamics.

2.3.3 Proof of Theorem 2.1

To prove this theorem, we check the sufficient conditions in Theorem 2.4. First requirement that $X_{i,T}$ is measurable w.r.t. $\mathcal{F}_{T-1}$, follows by the definition of $X_{i,T}$. Conditions (S1) and (S2) are verified in the following.

(S1) By Proposition 2.1-(P2), we see that $\lambda_{\min}(X_{i,T}) \to \infty$ a.s.; therefore, (S1) in Theorem 2.4 is satisfied.

(S2) Proposition 2.1-(P1) and (P2) imply that there exist positive constants C_1, C_2, such that :

$$\limsup_{T\to\infty} \frac{\log(\lambda_{\max}(X_{i,T}))}{\lambda_{\min}(X_{i,T})} \leq \limsup_{T\to\infty} \frac{\log(C_1)+\log(T)}{C_2|\mathcal{T}_{i,T}|} = 0 \quad \text{a.s.},$$

where the last equality follows by Lemma 2.1 (i.e. $|\mathcal{T}_{i,T}| = \mathcal{O}(T)$, a.s.). Therefore, the second condition of Theorem 2.4 is satisfied.

Therefore, by Theorem 2.4, for each $i \in \mathcal{S}$, we have:

$$\left\|\hat{A}_{i,T} - A_i\right\|_\infty \leq \mathcal{O}\left(\sqrt{\frac{\log\left[\lambda_{\max}(X_{i,T})\right]}{\lambda_{\min}(X_{i,T})}}\right), \quad \text{a.s.} \tag{2.8}$$

which proves the claim in Theorem 2.1.

2.3.4 Proof of Corollary 2.1

Corollary 2.1 is the direct consequence of Theorem 2.1 and Proposition 2.1. Proposition 2.1-(P1) implies that $\lambda_{\max}(X_{i,T}) = \mathcal{O}(\log(T))$. By substituting $\lambda_{\max}(X_{i,T})$ with $\mathcal{O}(\log(T))$ in the right hand side of Eq. (2.8), we get that for each $i \in \mathcal{S}$, the estimation error $\|\hat{A}_{i,T} - A_i\|_\infty$ is upper-bounded by $\mathcal{O}\big(\sqrt{\log(T)/|\mathcal{T}_{i,T}|}\big)$, *a.s.*

2.3.5 Proof of Theorem 2.2

We first establish the strong consistency of the parameter $\hat{\theta}_T$. By Theorem 2.1 and the fact that $k < \infty$, we get:

$$\left\|\hat{\theta}_T - \theta\right\|_\infty \leq \max_{i\in\mathcal{S}} \mathcal{O}\left(\sqrt{\frac{\log\left[\lambda_{\max}(X_{i,T})\right]}{\lambda_{\min}(X_{i,T})}}\right), \quad \text{a.s.}$$

Therefore, the result follows by applying Theorem 2.1 to the argmax of above equation. For the second part notice that by Lemma 2.1, we know $\lim_{T\to\infty} |\mathcal{T}_{i,T}|/T = \pi_\infty(i)$, a.s. Now, by

Corollary 2.1, we get:

$$\left\|\hat{A}_{i,T} - A_i\right\|_\infty \leq \mathcal{O}\left(\sqrt{\frac{\log(T)}{|\mathcal{T}_{i,T}|}}\right) = \mathcal{O}\left(\sqrt{\frac{\log T}{\pi_\infty(i)T}}\right) \quad \text{a.s.}$$

which is the claim of Theorem 2.2.

2.4 Discussion on stability in the average sense

The main results of this chapter are derived under Assumption 2.2 i.e., the MJS system (2.1) is stable in the average sense. In this section, we discuss the connection between this notion of stability and more common forms of stability in MJS systems, i.e., mean square stability and almost sure stability.

2.4.1 Stability on the average sense and mean square stability

A common assumption on the stability of MJS systems (e.g., [115] and [70]) is mean square stability defined as following:

Definition 2.3. *The MJS system* (2.1) *is called mean square stable (MSS) if there exists a deterministic vector $x_\infty \in \mathbb{R}^n$ and a deterministic positive definite matrix $Q_\infty \in \mathbb{R}^{n\times n}$ such that for any deterministic initial state x_0 and s_0 , we have: $\lim_{\tau\to\infty}\left\|\mathbb{E}[x_\tau] - x_\infty\right\| \to 0$, and $\lim_{\tau\to\infty}\left\|\mathbb{E}[x_\tau x_\tau^\intercal] - Q_\infty\right\| \to 0$.*

Proposition 2.2 (see [45, Theorem 3.9])**.** *The system is MSS, if and only if*

$$\lambda_{\max}\left((P^\intercal \otimes I_{n^2})\,\mathrm{diag}(A_1 \otimes A_1, \ldots, A_k \otimes A_k)\right) < 1.$$

We now show that stability in the average sense is a weaker notion of stability than MSS.

Proposition 2.3. *If the MJS system* (2.1) *is mean square stable, then the system is stable in the average sense.*

Proof. Since the system is MSS, there exists a positive definite matrix $Q_\infty \in \mathbb{R}^{n\times n}$ such that $\lim_{\tau\to\infty}\mathbb{E}[x_\tau x_\tau^\intercal] = Q_\infty$, which implies $\lim_{\tau\to\infty}\mathrm{Tr}(\mathbb{E}[x_\tau x_\tau^\intercal]) = \mathrm{Tr}(Q_\infty)$. Since $\mathrm{Tr}(\mathbb{E}[xx^\intercal]) = \mathbb{E}[\mathrm{Tr}(xx^\intercal)] = \mathbb{E}[x^\intercal x]$, MSS implies that the sequence of real numbers $\{\mathbb{E}(\|x_\tau\|^2)\}_{\tau\geq 0}$ converges to $\mathrm{Tr}(Q_\infty)$ and therefore:

$$\lim_{T\to\infty}\frac{1}{T}\sum_{\tau=1}^{T}\mathbb{E}(\|x_\tau\|^2) = \mathrm{Tr}(Q_\infty) < \infty. \tag{2.9}$$

Define events

$$E_n = \Big\{\omega \in \Omega : \limsup_{T\to\infty} \frac{1}{T}\sum_{\tau=1}^{T} \|x_\tau\|^2 \leq n\Big\}, \quad \forall n \in \mathbb{N}$$

and

$$E = \bigcup_{n=0}^{\infty} E_n = \Big\{\omega \in \Omega : \limsup_{T\to\infty} \frac{1}{T}\sum_{\tau=1}^{T} \|x_\tau\|^2 < \infty\Big\}.$$

Now, by the continuity of probability measure from below, we have:

$$\mathbb{P}(E) = \mathbb{P}(\bigcup_{n=0}^{\infty} E_n) = \lim_{n\to\infty} \mathbb{P}(E_n). \tag{2.10}$$

Note that

$$\begin{aligned}
\mathbb{P}(E_n) &= \mathbb{P}\big(\limsup_{T\to\infty} \frac{1}{T}\sum_{\tau=1}^{T} \|x_\tau\|^2 \leq n\big) \\
&\overset{(a)}{\geq} \limsup_{T\to\infty} \mathbb{P}\big(\frac{1}{T}\sum_{\tau=1}^{T} \|x_\tau\|^2 \leq n\big) \\
&\overset{(b)}{\geq} 1 - \limsup_{T\to\infty} \frac{\big(\sum_{\tau=1}^{T} \mathbb{E}\|x_\tau\|^2\big)}{Tn} \\
&\overset{(c)}{\geq} 1 - \frac{\mathrm{Tr}(Q_\infty)}{n},
\end{aligned}$$

where (a) follows from reverse Fatou's lemma, (b) follows from the Markov inequality and (c) follows from Eq. (2.9). Substituting the above in equation (10), we get

$$\mathbb{P}(E) \geq \lim_{n\to\infty} \Big(1 - \frac{\mathrm{Tr}(Q_\infty)}{n}\Big) = 1.$$

Therefore $\mathbb{P}(E) = 1$, and the system is stable in the average sense. □

Remark 2.5. *Proposition 2.3 shows that MSS implies Assumption 2.2. Therefore, the results of Theorem 2.1 and 2.2 also hold when Assumption 2.2 is replaced by the assumption that the system is MSS.*

2.4.2 Stability on the average sense and almost sure stability

Consider the noise free version of the MJS system (2.1) with the following dynamics:

$$x_{t+1} = A_{s_t} x_t, \quad t \geq 0. \tag{2.11}$$

Definition 2.4. *The system* (2.11) *is called almost surely stable if, for any deterministic initial state* x_0 *and* s_0*, we have:*

$$\lim_{t\to\infty} \|x_t\| = 0, \quad a.s.$$

A common sufficient condition to check the almost sure stability of MJS system (2.11) *is given below.*

Proposition 2.4 (see [45, Theorem 3.47])**.** *If the stationary distribution* $\pi_\infty = (\pi_\infty(1), \ldots, \pi_\infty(k))$ *satisfies* **(C1)** $\pi_\infty(i) \neq 0$ *for all* i *and* **(C2)** $\prod_{i=1}^{k} \sigma_{\max}(A_i)^{\pi_\infty(i)} < 1$*, then the system* (2.11) *is almost surely stable.*

We now show that (C1) and (C2) are also sufficient conditions for stability in the average sense.

Proposition 2.5. *If the MJS system* (2.1) *satisfies (C1) and (C2), then the system is stable in the average sense.*

Proof. To simplify the notation, we assume that $x_0 = 0$ which does not entail any loss of generality. Let $\Phi(t-1, \tau+1) = A_{s_{t-1}} \cdots A_{s_{\tau+1}}$ denote the state transition matrix where we follow the convention that $\Phi(t, \tau) = I$, for $t < \tau$. Then we can write the dynamics in Eq. (2.1) of the continuous component of the state in convolutional form as:

$$x_t = \sum_{\tau=0}^{t-1} \Phi(t-1, \tau+1) w_{\tau+1}. \tag{2.12}$$

where $\|\Phi(t-1, \tau+1)\| = \|A_{s_{t-1}} \ldots A_{s_{\tau+1}}\|$, and

$$\|A_{s_{t-1}} \ldots A_{s_{\tau+1}}\| \leq \sigma_{s_{t-1}} \cdots \sigma_{s_{\tau+1}} =: \Gamma_{t-1,\tau+1} \tag{2.13}$$

where $\sigma_{s_t} = \sigma_{\max}(A_{s_t})$. In the following lemma, it is established that the conditions (C1) and (C2) in Prop. 2.5 imply that the sum of norms of the state-transition matrices are uniformly bounded.

Lemma 2.4. *Under the conditions (C1) and (C2) in Prop. 2.5, there exists a constant* $\bar{\Gamma} < \infty$ *such that for all* $T > 1$*,* $\sum_{\tau=0}^{T-1} \|\Phi(T-1, \tau+1)\| \leq \bar{\Gamma}$*, a.s.*

The proof is presented in Appendix 2.A. The following Lemma shows the implication of Assumption 2.1 on the growth rate of energy of the noise process.

Lemma 2.5 ([52, Eq. (3.1)]). *Under Assumption 2.1, we have*

$$\sum_{\tau=1}^{T} \|w_\tau\|^2 = \mathcal{O}(T), \quad a.s.$$

Using the convolution formula in Eq. (2.12), we can bound the norm of the state $\|x_t\|^2$ as following:

$$\begin{aligned}
\|x_t\|^2 =& \Big(\big\| \sum_{\tau=0}^{t-1} \Phi(t-1, \tau+1) w_{\tau+1} \big\| \Big)^2 \\
\overset{(a)}{\leq}& \Big(\sum_{\tau=0}^{t-1} \|\Phi(t-1, \tau+1) w_{\tau+1}\| \Big)^2 \\
\overset{(b)}{\leq}& \Big(\sum_{\tau=0}^{t-1} \|\Phi(t-1, \tau+1)\| \|w_{\tau+1}\| \Big)^2 \\
\overset{(c)}{\leq}& \Big(\sum_{\tau=0}^{t-1} \Gamma_{t-1,\tau+1} \|w_{\tau+1}\| \Big)^2,
\end{aligned} \tag{2.14}$$

where (a) follows from triangle inequality and (b) follow from sub-multiplicative property of the matrix norm, and (c) follows from Eq. (2.13). Now for a fixed i, $i \in \mathcal{S}$, we have:

$$\begin{aligned}
\sum_{t=1}^{T} \|x_t\|^2 \leq& \sum_{t=1}^{T} \Big(\sum_{j=0}^{t-1} \Gamma_{t-1,j+1} \|w_{j+1}\| \Big)^2 \\
\overset{(d)}{\leq}& \sum_{t=1}^{T} \Big(\sum_{j=0}^{t-1} \Gamma_{t-1,j+1} \Big) \Big(\sum_{j=0}^{t-1} \Gamma_{t-1,j+1} \|w_{j+1}\|^2 \Big) \\
\overset{(e)}{\leq}& \bar{\Gamma} \sum_{t=1}^{T} \Big(\sum_{j=0}^{t-1} \Gamma_{t-1,j+1} \|w_{j+1}\|^2 \Big) \\
\overset{(f)}{\leq}& \bar{\Gamma} \sum_{j=0}^{T-1} \Big(\sum_{t=j+1}^{T} \Gamma_{t-1,j+1} \Big) \|w_{j+1}\|^2 \\
\overset{(g)}{\leq}& \bar{\Gamma}^2 \sum_{j=0}^{T-1} \|w_{j+1}\|^2 = \mathcal{O}(T) \quad \text{a.s.},
\end{aligned}$$

where (d) follows from Cauchy-Schwarz's inequality, (e) follows from Lemma 2.4, (f) follows from changing the order of summation, and (g) follows from boundedness of sub-sums of $\sum_{\tau=0}^{T-1} \Gamma_{T-1,\tau+1}$, and Lemma 2.4. □

Remark 2.6. *Proposition 2.5 shows that (C1) and (C2) imply Assumption 2.2. Therefore, the results of Theorem 2.1 and 2.2 also hold when Assumption 2.2 is replaced by the assumption that the system satisfies (C1) and (C2).*

2.4.3 Discussion on Non-Comparable Stability Assumption

The following examples illustrate that neither MSS nor conditions (C1) and (C2) in Proposition 2.5 is stronger than the other.

Example 2.1. *Let* $\theta^{\intercal} = \{A_1, 0\}$*, and* $p = (p_1, p_2)$ *be an i.i.d. probability transition, with* $\lambda_{\max}(p_1 A_1) > 1$ *and* $x_0 \neq 0$*. Then* $\mathbb{E}[x_{\tau+1}] = \mathbb{E}[A_{\sigma_\tau} x_\tau + w_{\tau+1}] = p_1 A_1 \mathbb{E}[x_\tau] = \cdots = (p_1 A_1)^\tau \mathbb{E}(x_0)$ *, which implies* $\lim_{\tau\to\infty} \mathbb{E}(x_\tau) = \infty$*. Therefore, this system is not mean square stable. However, this system satisfies conditions (C1) and (C2) in Prop. 2.5 and therefore is stable in the average sense.*

Example 2.2. *Consider non-switched system with matrix* A*, with* $\lambda_{\max}(A) < 1$ *and* $\sigma_{\max}(A) > 1$*. This system is mean square stable, but it does not satisfy the conditions (C1) and (C2) in Proposition 2.5.*

2.5 Numerical Simulation

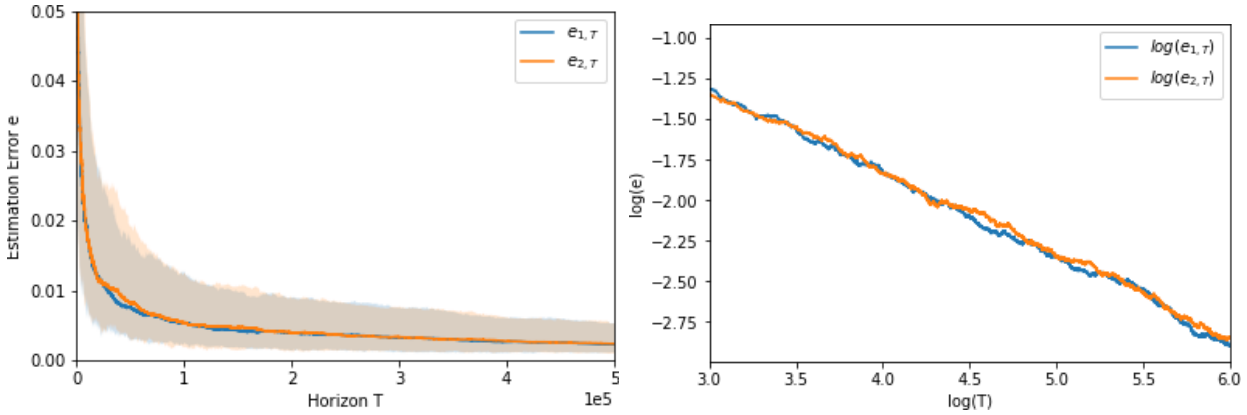
In this section, we illustrate the result of Theorem 2.1 via an example. Consider a MJS with $n = 2, k = 2$,

$$A_1 = \begin{bmatrix} 1.5 & 0 \\ 0 & 0.2 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0.01 & 0.1 \\ 0.1 & 0.1 \end{bmatrix},$$

probability transition matrix

$$P = \begin{bmatrix} 0.5 & 0.5 \\ 0.75 & 0.25 \end{bmatrix}$$

and i.i.d. $\{w_t\}_{t\geq 0}$ with $w_t \sim \mathcal{N}(0, I)$. Note that the example satisfies Assumptions 2.1 and conditions (C1) and (C2) of Proposition 2.5 (and, therefore, Assumption 2.2), but it is not mean square stable. We run the switched least squares for a horizon of $T = 10^6$ and repeat the experiment for 100 independent runs. We plot the estimation error $e_{i,T} = \|\hat{A}_{i,t} - A_1\|_\infty$ versus time in Fig. 2.1a. The plot shows that the estimation error is converging almost surely even though the system is not mean square stable. In Fig. 2.1b, logarithm of the estimation error versus logarithm of the horizon is plotted. The linearity of the graph along with approximate slope of -0.5 shows that $e_{i,T} = \tilde{O}(1/\sqrt{T})$.



(a) Error vs time.

(b) Log-error vs log-time.

Figure 2.1: Performance of switched least squares method for the example of Sec. 2.5. The solid line shows the mean across 100 runs and the shaded region shows the 25% to 75% quantile bound.

2.6 Conclusion and Future Directions

In this chapter, we investigated system identification of (autonomous) Markov jump linear systems. We proposed the switched least squares method, showed it is strongly consistent and derived the almost sure rate of convergence of $\mathcal{O}(\sqrt{\log(T)/T})$. This analysis provides a solid first step toward establishing almost sure regret bounds for adaptive control of MJS.

We derived our results assuming that system is stable in the average sense and we showed that this is a weaker assumption compared to mean square stability.

The current results are established for autonomous systems with Markov switching when the complete state of the system is observed. Interesting future research directions include relaxing these modeling assumptions and considering controlled systems under partial state observability and unobserved jump times.

Appendices to Chapter 2

2.A Proof of Lemma 2.4

Proof. Let $\sigma_i = \sigma_{\max}(A_i), i \in \{1, \ldots, k\}$. Define $\gamma_t = \sigma_{s_t}$. Then, by sub-multiplicative property of the matrix norms, we have:

$$\begin{aligned} \|\Phi(t-1, \tau+1)\| &= \|A_{s_{t-1}} \ldots A_{s_{\tau+1}}\| \\ &\leq \gamma_{t-1} \cdots \gamma_{\tau+1} =: \Gamma_{t-1,\tau+1}. \end{aligned} \tag{0.15}$$

Given numbers $m_1, \ldots, m_k$, define $f(m_1, \ldots, m_k) = \sigma_1^{m_1} \cdots \sigma_k^{m_k}$. Let $m_i(t-1, \tau+1) = \sum_{t'=\tau+1}^{t-1} \frac{\mathbb{1}\{s_{t'}=i\}}{t-\tau-1}$ denote the number of times the discrete state equals i in $[\tau+1, t-1]$. Then,

$$\begin{aligned} &\Gamma_{t-1}, \tau+1 = \gamma_{t-1} \cdots \gamma_{\tau+1} \\ &= f(m_1(t-1, \tau+1), \ldots, m_k(t-1, \tau+1))^{t-\tau-1}. \end{aligned}$$

Since $\{s_t\}_{t \geq 0}$ is aperiodic and irreducible Markov chain, by the Ergodic Theorem (Theorem 4.1 in [114]) we know that for any initial distribution π_0, $\lim_{t\to\infty} m_i(t-1, \tau+1) = \pi_\infty(i)$, a.s. Therefore, there exists a $N(\epsilon, \pi_0)$ such that for all $t - \tau - 1 \geq N(\epsilon, \pi_0)$, $|m_i(t-1, \tau+1) - \pi_\infty(i)| < \epsilon$ a.s. for all i. Define $N^*(\epsilon) = \sup_{\pi_0 \in \Delta_k} N(\epsilon, \pi_0)$, where Δ_k denotes the k-dimensional simplex. Let π^* denote the corresponding arg sup (which lies in Δ_k due to compactness). Then, $N^* = N(\epsilon, \pi^*)$ is finite due to the Ergodic Theorem. Therefore, for $t - \tau - 1 \geq N^*(\epsilon)$, $|m_i(t-1, \tau+1) - \pi_\infty(i)| < \epsilon$.

Furthermore, the rate of convergence of $m_i(t-1, \tau+1)$ to $\pi_\infty(i)$ only depends on $\tau+1$ and $t-1$ only through their difference. By the continuity of $f(\cdot)$, for any $\epsilon' > 0$, there exists a $N'(\epsilon')$ such that for all $t - \tau - 1 \geq N'(\epsilon')$, $|f(m_1(t-1, \tau+1), \cdots, m_k(t-1, \tau+1)) - f(\pi_\infty(1), \cdots, \pi_\infty(k))| < \epsilon'$ *a.s.* Hence, almost surely we have:

$$\begin{aligned} &f(m_1(t-1, \tau+1), \ldots, m_k(t-1, \tau+1)) \\ &\qquad < f(\pi_\infty(1), \ldots, \pi_\infty(k)) + \epsilon' \end{aligned}$$

By (C1) and (C2) conditions, we know $f(\pi_\infty(1), \ldots, \pi_\infty(k)) < 1$. Now we can pick ϵ' such

that $f(\pi_\infty(1),\dots,\pi_\infty(k))+\epsilon' =: \beta^* < 1$. Then for all $t \geq 1$,

$$\begin{aligned}
&\sum_{\tau=1}^{t-1} f(m_1(t-1,\tau+1),\dots,m_k(t-1,\tau+1))^{t-\tau-1} \\
\leq &\sum_{\tau=1}^{t-N(\epsilon')-1} \beta^{*t-\tau-1} + \\
&\sum_{\tau=t-N'(\epsilon')}^{t-1} f(m_1(t-1,\tau+1),\dots,m_k(t-1,\tau+1))^{t-\tau-1} \\
<&\frac{\beta^{*N'(\epsilon')}}{1-\beta^*} + \sum_{\tau=t-N'(\epsilon')}^{t-1} F_*^{t-\tau-1},
\end{aligned}$$

where $F_* = \max_{\pi(1),\dots,\pi(k)\in\Delta_k} f(\pi(1),\dots,\pi(k))$, which is clearly bounded. As a result, both terms in the right hand side are bounded which implies the statement in the claim. □

Chapter 3

Relative Almost Sure Regret Bounds for Certainty Equivalence Control of Markov Jump Systems

3.1 Overview

In this chapter, we investigate the problem of Reinforcement Learning (RL) in the Markov Jump Linear Systems framework. The results of this chapter are published in [107].

3.1.1 Organization

This chapter is organized as follows. In Sec 3.2, we review some standard results about Markov Jump Linear Systems (MJLS) that are useful in our analysis. In Sec. 3.3, we characterize the notion of almost sure regret criteria. In Sec. 3.4, we present an upper bound on the regret of adaptive linear policies. In Sec. 3.5, we present our system identification method and reinforcement learning algorithm. The main results are presented in Sec. 3.6. We conclude this chapter in Sec. 3.7.

3.2 Background on Markov Jump Linear Systems

We start by a review of stability of autonomous MJLS and the basic results for optimal control of MJLS.

3.2.1 Stability of Autonomous Markov Jump Linear Systems

Consider an autonomous discrete-time MJLS with continuous state $x_t \in \mathbb{R}^n$ and the discrete state $s_t \in \mathcal{S} = \{1, 2, \ldots, d\}$. The system starts with a known initial state (x_1, s_1).

The continuous state evolves over time according to

$$x_{t+1} = A_{s_t} x_t, \quad t \geq 1, \tag{3.1}$$

where the set $\{A_s \in \mathbb{R}^{n\times n}\}_{s\in\mathcal{S}}$ consists of the system dynamics matrices. The discrete state evolves in a time-homogeneous Markov manner according to a transition matrix H. We will refer to the above system as MJLS system $(\{A_s\}_{s\in\mathcal{S}}, H)$.

We assume that the Markov chain $\{s_t\}_{t\geq 1}$ is irreducible and aperiodic, and therefore, has a stationary distribution $\{\rho_s\}_{s\in\mathcal{S}}$.

Definition 3.1. *The MJLS system* (3.1) *is called Mean Square Stable (MSS) if for any initial state* (x_1, s_1), $\lim_{t\to\infty} \|\mathbb{E}[x_t]\| = 0$, *and* $\lim_{t\to\infty} \|\mathbb{E}[x_t x_t^\intercal]\| = 0$.

It can be shown that the two definitions of MSS in Chapter 2 and 3 are equivalent. The following characterizations of MSS follow from [45, Theorem 3.9].

Proposition 3.1. *The following conditions are equivalent:*

1. *The MJLS system in* (3.1) *is MSS.*

2. *Transition probability matrix* H *and matrices* $\{A_s\}_{s\in\mathcal{S}}$ *satisfy:*

$$\lambda_{\max}\Big((H^\intercal \otimes I_{n^2})\,\mathrm{diag}(A_1 \otimes A_1, \ldots, A_d \otimes A_d)\Big) < 1.$$

3. *The MJLS system* (3.1) *is exponentially stochastically stable , i.e., there exists* $\beta \geq 1$ *and* $0 < \zeta < 1$ *such that for any initial state* (x_1, s_1), *we have*

$$\mathbb{E}[\|x_t\|^2] \leq \beta\zeta^t \|x_0\|^2, \quad t \geq 1.$$

4. *The MJLS system* (3.1) *is stochastically stable (SS), i.e., for all initial state* (x_1, s_1), *we have*

$$\sum_{t=0}^{\infty} \mathbb{E}[\|x_t\|^2] < \infty.$$

3.2.2 Optimal Control of Markov Jump Linear Systems

Consider a discrete-time MJLS with continuous state $x_t \in \mathbb{R}^n$, discrete state $s_t \in \mathcal{S}$, control input $u_t \in \mathbb{R}^m$, and disturbance $w_t \in \mathbb{R}^n$. The system starts with a known initial state (x_1, s_1). The continuous state evolves over time according to:

$$x_{t+1} = A_{s_t} x_t + B_{s_t} u_t + w_t, \quad t \geq 1, \tag{3.2}$$

where $\{A_s \in \mathbb{R}^{n\times n}\}_{s\in\mathcal{S}}$ and $\{B_s \in \mathbb{R}^{n\times m}\}_{s\in\mathcal{S}}$ are the system dynamics matrices, and $\{w_t\}_{t\geq 1}$ is an i.i.d. process with $\mathbb{E}[w_t] = 0$ and $\mathbb{E}[w_t w_t^\intercal] = \sigma_w^2 I$. The discrete state evolves in a time-homogeneous Markov manner, independent of $\{w_t\}_{t\geq 1}$, according to a transition matrix H. We assume that the Markov chain $\{s_t\}_{t\geq 1}$ is irreducible and aperiodic, and therefore, has a stationary distribution $\{\rho_s\}_{s\in\mathcal{S}}$.

The system incurs a per-step cost

$$c(x_t, s_t, u_t) \coloneqq x_t^\intercal Q_{s_t} x_t + u_t^\intercal R_{s_t} u_t, \tag{3.3}$$

where $\{Q_s \in \mathbb{R}^{n\times n}\}_{s\in\mathcal{S}}$ and $\{R_s \in \mathbb{R}^{m\times m}\}_{s\in\mathcal{S}}$ are positive definite matrices. The objective is to design a controller which observes the state of the system and chooses control inputs to minimize the long term average cost given by

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} c(x_t, s_t, u_t)\right]. \tag{3.4}$$

3.2.2.1 Stochastic Stabilizability and Stochastic Detectability

We now define two important properties of MJLS systems:

Definition 3.2. *The MJLS system* (3.2) *is* stochastically stabilizable*, if there exists gain matrices* $\{F_s \in \mathbb{R}^{m\times n}\}_{s\in\mathcal{S}}$ *such that the autonomous MJLS system* $(\{A_s - B_s F_s\}_{s\in\mathcal{S}}, H)$ *is MSS.*

Definition 3.3. *The MJLS system* (3.2) *is* stochastically detectable *, if there exists gain matrices* $\{K_s \in \mathbb{R}^{n\times n}\}_{s\in\mathcal{S}}$ *such that the autonomous MJLS system* $(\{A_s - K_s Q_s^{1/2}\}_{s\in\mathcal{S}}, H)$ *is MSS.*

Note that one can check stochastic stability and stochastic detectability via Linear Matrix inequalities (LMIs). For instance, a check for stochastic stabilizability is given by [45, Proposition 3.42].

Proposition 3.2. *The MJLS system* (3.2) *is stochastically stabilizable if and only if there exist matrices* $\{W_s^{(2)} \in \mathbb{R}^{n\times m}\}_{s\in\mathcal{S}}$ *and positive semi-definite matrices* $\{W_s^{(1)} \in \mathbb{R}^{n\times n}\}_{s\in\mathcal{S}}$ *and*

$\{W_s^{(3)} \in \mathbb{R}^{m\times m}\}_{s\in\mathcal{S}}$ *such that:*

$$\sum_{s\in\mathcal{S}} H_{ss'}(A_sW_s^{(1)}A_s^\intercal + B_s(W_s^{(2)})^\intercal A_s^\intercal + A_sW_s^{(2)}B_s^\intercal + B_sW_s^{(3)}B_s^\intercal) < W_{s'}^{(1)} \qquad \forall s' \in \mathcal{S},$$

$$\begin{bmatrix} W_s^{(1)} & W_s^{(2)} \\ (W_s^{(2)})^\intercal & W_s^{(3)} \end{bmatrix} \geq 0, \qquad \forall s \in \mathcal{S},$$

$$W_s^{(1)} > 0, \qquad \forall s \in \mathcal{S}.$$

A similar test for stochastic detectability follows by replacing B_s by $(Q_s^{1/2})^\intercal$ in the above proposition.

3.2.2.2 Optimal Control of MJLS

We assume that the system satisfies the following assumption.

Assumption 3.1. *The MJLS system in* (3.2) *is stochastically stabilizable and stochastically detectable.*

The following result follows from [116, Theorem 45 and Theorem 51].

Theorem 3.1. *Under Assumption 3.1, the minimum value of the average cost* (3.4) *is*

$$\sigma_w^2 \sum_{s\in\mathcal{S}} \sum_{s_+\in\mathcal{S}} \rho_s H_{ss_+} \operatorname{Tr}(P_{s_+}) \tag{3.5}$$

and is achieved by the feedback policy

$$u_t = -L_{s_t}x_t, \quad t \geq 1, \tag{3.6}$$

where the gains $\{L_s\}_{s\in\mathcal{S}}$ *are given by*

$$L_s = (R_s + B_s^\intercal \bar{P}_s B_s)^{-1} B_s^\intercal \bar{P}_s A_s, \quad s \in \mathcal{S} \tag{3.7}$$

and $\{P_s\}_{s\in\mathcal{S}}$ *is the solution of the following set of algebraic Riccati equations:*

$$\bar{P}_s = \sum_{s_+\in\mathcal{S}} H_{ss_+}P_{s_+}, \quad s \in \mathcal{S}, \tag{3.8}$$

$$P_s = Q_s + A_s^\intercal \bar{P}_s A_s - A_s^\intercal \bar{P}_s B_s^\intercal (R_s + B_s^\intercal \bar{P}_s B_s)^{-1} B_s^\intercal \bar{P}_s A_s, \quad s \in \mathcal{S}. \tag{3.9}$$

As established in [116, Theorem 45], the optimal control law is stabilizing in the following sense.

Proposition 3.3. *The autonomous MJLS system* $(\{A_s - B_s L_s\}_{s\in\mathcal{S}}, H)$ *is MSS.*

Remark 3.1. *The result of Proposition 3.3 in [116, Lemma 45] states that the system* $(\{A_s - B_s L_s\}_{s\in\mathcal{S}}, H)$ *is stochastically stable. As established in Proposition 3.1, stochastic stability is equivalent to MSS, so we have stated Prop. 3.3 in terms of MSS.*

3.3 The Learning Problem

3.3.1 Some Remarks on Notation

3.3.1.1 Notation for Probability Spaces

We need a somewhat elaborate notation to describe our notion of regret. The MJLS system described above is a stochastic system with two stochastic inputs: the noise process $\{w_t\}_{t\geq 1}$ and the switching process $\{s_t\}_{t\geq 1}$. In addition, the learning algorithm may randomize while choosing control actions as well. We assume that the noise process and randomization done by the algorithm are defined on a probability space $(\Omega_1, \mathcal{F}_1, \mu_1)$ and the switching process is defined on a separate probability space $(\Omega_2, \mathcal{F}_2, \mu_2)$. Since the processes $\{w_t\}_{t\geq 1}$ and $\{s_t\}_{t\geq 1}$ and the randomization done by the algorithm are independent, we consider the probability space

$$(\Omega, \mathcal{F}, \mu) = (\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2, \mu_1 \otimes \mu_2),$$

where $\mathcal{F}_1 \otimes \mathcal{F}_2$ is the product sigma algebra given by $\sigma(D_1 \times D_2 : D_1 \in \mathcal{F}_1, D_2 \in \mathcal{F}_2)$, and $\mu_1 \otimes \mu_2$ is the product measure on $\mathcal{F}_1 \otimes \mathcal{F}_2$, i.e., for any $D_1 \in \mathcal{F}_1$, $D_2 \in \mathcal{F}_2$, we have $\mu(D_1 \times D_2) = \mu_1(D_1)\mu_2(D_2)$. We will use the tuple $(\Omega, \mathcal{F}, \mu)$ as the probability space to define all the system variables. We abbreviate almost surely with respect to measure $\mu(\cdot)$ as μ-$a.s.$ and almost surely with respect to measure $\mu_1(\cdot)$ as μ_1-$a.s.$

3.3.1.2 Notation for Policy Dependent Sample Paths

To avoid confusion, we also use a slightly elaborate notation to indicate sample paths of state and action corresponding to a specific policy. Let $\theta = \{A_s, B_s\}_{s\in\mathcal{S}}$ denote the parameters of the system dynamics. Suppose the control input u_t is chosen as a function of the history of states and actions $(x_{1:t}, s_{1:t}, u_{1:t-1})$ according to a possibly randomized history-dependent measurable policy π. Then for any $\omega = (\omega_1, \omega_2) \in \Omega$, we use the notation $\{x_t^\pi(\omega), s_t(\omega_2), u_t^\pi(\omega)\}_{t\geq 1}$ to denote the states and the control actions along the sample path ω for the system when the controller is following policy π. Note that the discrete component of state, $s_t(\omega_2)$ only depends on ω_2 and does not depend on the policy π.

When it is clear from the context, we will not explicitly indicate the dependence on θ, π, and ω.

3.3.2 Regret Definition

We are interested in the setting where the system parameters θ are unknown and the cost parameters $\{(Q_s, R_s)\}_{s\in\mathcal{S}}$ and transition matrix H are known. A learning agent observes the state (x_t, s_t) of the system and chooses the control input u_t according to a possibly history-dependent randomized measurable policy π. For any fixed realization $\omega_1 \in \Omega_1$ of the system noise and possible randomization done by the algorithm, let

$$J_T^\pi(\omega_1) = \int_{\Omega_2} \sum_{t=1}^{T} c(x_t^\pi(\omega_1, \omega_2), s_t(\omega_2), u_t^\pi(\omega_1, \omega_2))\mu_2(d\omega_2)$$

denote the performance of policy π along the sample path ω_1 for the horizon T averaged over the realizations of mode switching.

The (frequentist) **regret** of policy π is given by

$$\mathcal{R}_T^\pi(\omega_1) = J_T^\pi(\omega_1) - J_T^{\pi_\theta^*}(\omega_1),$$

where π_θ^* is the optimal policy corresponding to parameters θ.

Note that the notion of regret can be defined at different degrees of granularity. In particular, regret may be defined as a random variable which depends on the realization of the noise sequences and the randomizations done by the algorithm. Alternatively, it may be defined in terms of expectation over noise and algorithm randomization. In this chapter, we take an intermediate approach: we define regret as a random variable which depends on the realization of the process noise and the randomizations done by the algorithm, but take the expectation over the discrete switching sequence.

3.4 An Upper Bound on Regret for Adaptive Linear Policies with Persistence of Excitation

Let $\mathcal{F}_t = \sigma(x_{1:t-1}, s_{1:t-1}, u_{1:t-1})$ denote the sigma algebra generated by the observations of the history of states and actions of the learning agent at the beginning of time t. Motivated by the structure of the optimal policy presented in Theorem 3.1, we restrict attention to adaptive linear policies defined below.

Definition 3.4 (Adaptive linear policy)**.** *An adaptive linear policy π with persistence of*

excitation is characterized by a sequence of gains $\{\hat{L}_s(t) \in \mathbb{R}^{m\times n}\}_{s\in\mathcal{S},t\geq 1}$*, where* $\{\hat{L}_s(t)\}_{s\in\mathcal{S}}$ *is* $\mathcal{F}_t$*-measurable, and an independent noise process* $\{\nu_t\}_{t\geq 1}$*,* $\nu_t \in \mathbb{R}^n$*, where* $\nu_t \sim \mathcal{N}(0, \sigma_t^2 I)$*. The control input chosen by policy* π *at time* t *is given by* $u_t = -\hat{L}_{s_t}(t)x_t + \nu_t$*.*

Theorem 3.2. *Consider an adaptive linear policy* π *with persistence of excitation with gains* $\{\hat{L}_s(t)\}_{s\in\mathcal{S},t\geq 1}$ *and noise-level* $\{\sigma_t^2\}_{t\geq 1}$*. The regret of policy* π *may be decomposed as follows*

$$\mathcal{R}_T^\pi(\omega_1) = \mathcal{O}\big(\mathcal{R}_{1,T}^\pi(\omega_1)\big) + \mathcal{O}\big(\mathcal{R}_{2,T}^\pi(\omega_1)\big) + \mathcal{R}_{3,T}^\pi(\omega_1) \tag{3.10}$$

where

$$\mathcal{R}_{1,T}^\pi(\omega_1) = \int_{\Omega_2} \sum_{t=1}^{T} r_{1,t}^\pi(x_t^\pi(\omega_1,\omega_2), s_t(\omega_2))\mu_2(d\omega_2)$$

with $r_{1,t}^\pi(x_t, s_t)$ *given by*

$$x_t^\intercal(\hat{L}_{s_t}(t) - L_{s_t})^\intercal[R_{s_t} + B_{s_t}\bar{P}_{s_t}B_{s_t}](\hat{L}_{s_t}(t) - L_{s_t})x_t^\intercal,$$

and

$$\mathcal{R}_{2,T}^\pi(\omega_1) = \int_{\Omega_2} \sum_{t=1}^{T} r_{2,t}^\pi(\nu_t(\omega_1), s_t(\omega_2))\mu_2(d\omega_2)$$

with $r_{2,t}^\pi(\nu_t, s_t)$ *given by* $\nu_t^\intercal[R_{s_t} + B_{s_t}\bar{P}_{s_t}B_{s_t}]\nu_t$*, and*

$$\mathcal{R}_{3,T}^\pi(\omega_1) = \int_{\Omega_2} r_{3,t}^\pi(x_{T+1}^\pi(\omega), x_{T+1}^{\pi_\theta^*}(\omega), s_{T+1}(\omega_2))\mu_2(d\omega_2)$$

with $\omega = (\omega_1, \omega_2)$ *and* $r_{3,t}^\pi(x_{T+1}, x_{T+1}^{\pi_\theta^*}, s_{T+1})$ *given by* $(x_{T+1}^{\pi_\theta^*})^\intercal P_{s_{T+1}} x_{T+1}^{\pi_\theta^*} - x_{T+1}^\intercal P_{s_{T+1}} x_{T+1}$*, where recall that* $x^{\pi_\theta^*}$ *denotes the state corresponding to the optimal policy* π_θ^**.*

The proof is presented in Appendix 3.A.

3.5 A Certainty Equivalence Based Learning Algorithm

3.5.1 Overview of the Learning Algorithm

We consider a specific type of certainty equivalence-based learning algorithm and analyze its regret by using Theorem 3.2. The algorithm consists of two phases: a *system identification phase* which lasts for a fixed time $T^{(0)}$; and an *adaptation phase*, which last for the remainder of the time that the system is running. The adaptation phase runs in episodes, and the length of k-th episode is $\lfloor \alpha^k T^{(0)} \rfloor$, where $\alpha > 1$ is a constant. We use $t^{(k)}$ to denote the start time of episode k and use $T^{(k)}$ to denote the length of episode k.

Before describing the two phases in detail, we need to define the notion of stabilizing gains.

Definition 3.5. *A set of gain matrices* $\{\bar{L}_s \in \mathbb{R}^{m\times n}\}_{s\in\mathcal{S}}$ *is said to be* stabilizing *for the MJLS system* (3.2) *if the autonomous system* $(\{A_s - B_s\bar{L}_s\}_{s\in\mathcal{S}}, H)$ *is MSS.*

We make the following assumption:

Assumption 3.2. *The learning agent has access to a set of stabilizing controllers* $\{\bar{L}_s\}_{s\in\mathcal{S}}$.

Assumption 3.2 is a common assumption in the literature of reinforcement learning for LQR systems [27, 29, 31, 34, 71]. During the system identification phase, the control input is chosen as $u_t = -\bar{L}_{s_t}x_t + \nu_t$, where ν_t is i.i.d., zero mean Gaussian random noise with covariance $I/\sqrt{T^{(0)}}$. We then use the system identification algorithm used in the next section to generate an initial estimate $\hat{\theta}^{(0)}$.

During episode k of the adaption phase, at time $t^{(k)}$, we pick control gains $\{\hat{L}_s^{(k)}\}_{s\in\mathcal{S}}$ to be the optimal control gains corresponding to the estimate $\hat{\theta}^{(k-1)}$. During the episode, we choose the control input as $u_t = -\hat{L}_{s_t}^{(k)}x_t + \nu_t$, where ν_t is i.i.d., zero mean Gaussian random noise with covariance $I/\sqrt{T^{(k)}}$. At the end of the k-th episode, we use the system identification algorithm described in the next section to generate a new estimate $\hat{\theta}^{(k)}$ based on all the data seen in episode k.

A detailed description of the learning algorithm is presented in Algorithm 1.

3.5.2 The System Identification Algorithm

In this section, we describe the system identification algorithm used in both phases. This algorithm is a variation of the switched least squares system identification algorithm presented in Chapter 2 for autonomous system.

For uniformity of notation, we allow $k = 0$ to mean the system identification phase and set $t^{(0)} = 1$ and $\hat{L}_s^{(0)} = \bar{L}_s$ for $s \in \mathcal{S}$. Now consider a generic k-th episode, $k \in \{0, 1, \dots\}$, which is of length $T^{(k)}$. During this episode, the control input is chosen as

$$u_t = -\hat{L}_{s_t}^{(k)}x_t + \nu_t,$$

where ν_t is random noise chosen as $\nu_t \sim \mathcal{N}(0, \sigma_{(k)}^2 I)$, where $\sigma_{(k)}^2 = 1/\sqrt{T^{(k)}}$. Thus, Eq. (3.2) may be written as

$$x_{t+1} = A_{s_t}x_t - B_{s_t}\hat{L}_{s_t}^{(k)}x_t + B_{s_t}\nu_t + w_t \tag{3.11}$$

or, equivalently,

$$x_{t+1} = \eta_{s_t}^{(k)}z_t + w_t. \tag{3.12}$$

Algorithm 1: Certainty equivalence based learning algorithm

input : A set of stabilizing controllers $\{\bar{L}_s\}_{s\in\mathcal{S}}$
Time $T^{(0)}$; Scaling factor $\alpha > 1$.

System ID :

Initialize $\hat{L}_s^{(0)} = \bar{L}_s$, for all $s \in \mathcal{S}$.
Initialize $t^{(0)} = 1$.
for *time* $t \in \{t^{(0)}, \dots, t^{(0)} + T^{(0)} - 1\}$ **do**
    Sample $\nu_t \sim \mathcal{N}(0, \sigma_{(0)}^2 I)$, where $\sigma_{(0)}^2 = 1/\sqrt{T^{(0)}}$.
    Apply control input $u_t = \hat{L}_{s_t}^{(0)} x_t + \nu_t$.
Generate estimate $\hat{\theta}^{(0)}$ using (3.13) and (3.14).

Adaptation:

for *episode* $k = 1, 2, \dots$ **do**
    Initialize $t^{(k)} = t + 1$; $T^{(k)} = \lfloor \alpha^k T^{(0)} \rfloor$.
    Choose $\{\hat{L}_s^{(k)}\}_{s\in\mathcal{S}}$ using (3.7) for system $\hat{\theta}^{(k-1)}$.
    Set $\sigma_{(k)}^2 = 1/\sqrt{T^{(k)}}$.
    for *time* $t \in \{t^{(k)}, \dots, t^{(k)} + T^{(k)} - 1\}$ **do**
        Sample $\nu_t \sim \mathcal{N}(0, \sigma_{(k)}^2 I)$.
        Apply control input $u_t = \hat{L}_{s_t}^{(k)} x_t + \nu_t$.
    Generate estimate $\hat{\theta}^{(k)}$ using (3.13) and (3.14)

where $\{\eta_s^{(k)} \in \mathbb{R}^{n\times(n+m)}\}_{s\in\mathcal{S}}$ is given by

$$\eta_s^{(k)} \coloneqq [A_s - B_s \hat{L}_s^{(k)}, B_s], \quad s \in \mathcal{S},$$

and $z_t^\intercal \coloneqq [x_t^\intercal, \nu_t^\intercal] \in \mathbb{R}^{n+m}$.

At the end of the episode, we generate estimates $\hat{\eta}^{(k)} \coloneqq \{\hat{\eta}_s^{(k)} \in \mathbb{R}^{n\times(n+m)}\}_{s\in\mathcal{S}}$ by solving the following switched least squares problem:

$$\hat{\eta}^{(k)} = \underset{\eta^{(k)}=\{\eta_s^{(k)}:s\in\mathcal{S}\}}{\arg\min} \sum_{t=t^{(k)}}^{t^{(k)}+T^{(k)}-1} \|x_{t+1} - \eta_{s_t}^{(k)} z_t\|^2. \tag{3.13}$$

We then compute estimates $\{\hat{B}_s^{(k)}\}_{s\in\mathcal{S}}$ and $\{\hat{A}_s^{(k)}\}_{s\in\mathcal{S}}$ as:

$$\hat{B}_s^{(k)} = \hat{\eta}_s^{(k)} \begin{bmatrix} 0_{n\times n} \\ I_{m\times n} \end{bmatrix}, \quad \hat{A}_s^{(k)} = \hat{\eta}_s^{(k)} \begin{bmatrix} I_{n\times n} \\ \hat{L}_s^{(k)} \end{bmatrix}, \quad s \in \mathcal{S}. \tag{3.14}$$

We denote the estimated parameters as $\hat{\theta}_s^{(k)} \coloneqq [\hat{A}_s^{(k)}, \hat{B}_s^{(k)}]$, $s \in \mathcal{S}$ and use $\hat{\theta}^{(k)} \coloneqq \{\hat{\theta}_s^{(k)}\}_{s\in\mathcal{S}}$

to denote the estimated parameters of the model.

3.6 The Main Results

3.6.1 Asymptotic Regret of Certainty Equivalence Algorithm

In our analysis, we need to assume that the proposed learning algorithm at all times generates estimates such that the gains corresponding to those estimates stabilize the original system.

Definition 3.6. *Given the set of stabilizing controllers* $\{\bar{L}_s\}_{s\in\mathcal{S}}$*, time* T_0 *and scaling factor* α*, let* $\mathcal{A}_0$ *be the set of all sample paths* $\omega_1 \in \Omega_1$ *such that for almost all* $\omega_2 \in \Omega_2$ *and* $k \geq 1$ *the gains* $\{\hat{L}_s^{(k)}(\omega_1, \omega_2)\}_{s\in\mathcal{S}}$ *are stabilizing for MJLS system* (3.2)*.*

Assumption 3.3. *We assume* $\mu_1(\mathcal{A}_0) > 0$*.*

In our results below, we restrict attention to the sample paths $\omega_1 \in \mathcal{A}_0$. Note that the process $\{s_t\}_{t\geq 0}$ remains Markov on the set $\mathcal{A}_0 \times \Omega_2$ with the same transition probabilities. We assume that $\mu_1(\mathcal{A}_0) > 0$, which is weaker than the stability assumption implicitly imposed in [31] for (non-switching) LQR model, where it was assumed that $\mu(\mathcal{A}_0) = 1$.

By an argument similar to that used in Chapter 2 for autonomous systems, we can show that if the controller used in an episode is stable and the episode is asymptotically large, the estimates generated by switched least squares system identification algorithm described in Sec. 3.5.2 converge almost surely to the correct parameters. We can also characterize the rate of convergence, as shown below:

Theorem 3.3. *On the set* $\mathcal{A}_0$*, the estimate* $\hat{\theta}^{(k)}$ *is strongly consistent, i.e.*

$$\lim_{k\to\infty} \|\hat{\theta}^{(k)} - \theta\| = 0, \quad \mu_1 - a.s.$$

Furthermore, the error of the system identification method is upper bounded by:

$$\limsup_{k\to\infty} \frac{\|\hat{\theta}^{(k)} - \theta\|}{\sqrt{\log(T^{(k)})/\sigma^{(k)}T^{(k)}}} < \infty, \quad \mu_1\text{-}a.s. \tag{3.15}$$

The proof is presented in Appendix 3.B.

Following theorem establishes the regret bound for Algorithm 1. This regret matches with the regret of LQR problems established in [26, 29, 31, 34, 73] and the regret of MJLS-LQR established in [71].

Theorem 3.4. *On the set $\mathcal{A}_0$, the regret of Algorithm 1 is given by:*

$$\mathcal{R}_T^{\hat{\pi}} \leq \mathcal{O}(\sqrt{T}\log(T)) \quad \mu_1\text{-}a.s.$$

The proof is presented in Appendix 3.C.

3.6.2 Sufficient Conditions for Stability

In characterizing the almost sure regret of adaptive control problems, ensuring the stability of the system is a challenging problem. Our results in Theorems 3.3 and 3.4 are derived on the set $\mathcal{A}_0$. In this section, we try to weaken this requirement by characterizing a set which is larger than $\mathcal{A}_0$. For the MJLS system in (3.2) with parameters θ, let $L^\theta = \{L_s^\theta\}_{s\in\mathcal{S}}$ denote the set of optimal control gains. Define:

$$\mathcal{B}_\epsilon(L^\theta) \coloneqq \Big\{\{\hat{L}_s\}_{s\in\mathcal{S}} : \|\hat{L}_s - L_s^\theta\| \leq \epsilon, \forall s \in \mathcal{S}\Big\},$$

as a ball in the space of gain matrices with radius ϵ centered at L^θ.

Lemma 3.1. *[71, Lemma C.1] For the MJLS in* (3.2)*, there exists a radius ϵ_θ such that all the gains $\{\hat{L}_s\}_{s\in\mathcal{S}} \in \mathcal{B}_{\epsilon_\theta}(L^\theta)$ are stabilizing for θ.*

We define $\mathcal{B}_\delta(\theta) \coloneqq \big\{\{\hat{\theta}_s\}_{s\in\mathcal{S}} : \|\hat{\theta}_s - \theta_s\| \leq \delta, \forall s \in \mathcal{S}\big\}$. Now let δ_θ be the radius such that if $\hat{\theta} \in \mathcal{B}_{\delta_\theta}(\theta)$ then $L^{\hat{\theta}} \in \mathcal{B}_{\epsilon_\theta}(L_\theta)$.

We now characterize the connection between the assumptions on the stability and length of the identification phase $T^{(0)}$. Consider a system identification setup in which we use adaptive linear policy $\{\mathring{L}_s\}_{s\in\mathcal{S}}$ with persistent of excitation $\nu_t \sim \mathcal{N}(0, \mathring{\sigma}^2 I)$, where $\{\mathring{L}_s\}_{s\in\mathcal{S}}$ is a stabilizing controller. We get $x_{t+1} = \mathring{\eta}_{s_t} z_t + w_t$, where $\mathring{\eta}_s \coloneqq [A_s - B_s\mathring{L}_s, B_s]$. We estimate $\hat{\mathring{\eta}}_T \coloneqq \{\hat{\mathring{\eta}}_{s,T} \in \mathbb{R}^{n\times(n+m)}\}_{s\in\mathcal{S}}$ by solving:

$$\hat{\mathring{\eta}}_T = \underset{\mathring{\eta}=\{\mathring{\eta}_s : s\in\mathcal{S}\}}{\arg\min} \sum_{t=1}^{T} \|x_{t+1} - \mathring{\eta}_{s_t} z_t\|^2. \tag{3.16}$$

We generate the estimate $\hat{\theta}_T$ from $\hat{\mathring{\eta}}_T$ similarly to (3.14). To explicitly emphasize the functional dependence of $\hat{\theta}_T$ on $\mathring{L} = \{\mathring{L}_s\}_{s\in\mathcal{S}}$ and $\omega \in \Omega$, we use the notation $\hat{\theta}_T(\mathring{L}, \omega)$. By Theorem 2.2, we have that if $\{\mathring{L}_s\}_{s\in\mathcal{S}}$ is a stabilizing controller, then

$$\lim_{T\to\infty} \|\hat{\theta}_T(\mathring{L}, \omega) - \theta\| = 0, \quad \mu - a.s. \tag{3.17}$$

and the error of the system identification is upper bounded by:

$$\limsup_{T\to\infty} \frac{\|\hat{\theta}_T(\mathring{L},\omega)-\theta\|}{\sqrt{\log(T)/\mathring{\sigma}^2 T}} < \infty, \quad \mu\text{-a.s.} \tag{3.18}$$

Now for any generic stabilizing gain $\tilde{L} = \{\tilde{L}_s\}_{s\in\mathcal{S}}$, define

$$\begin{aligned} T_{\delta_\theta}(\tilde{L},\omega) &\coloneqq \inf\{T\in\mathbb{N} : \forall t\geq T, \|\hat{\theta}_T(\tilde{L},\omega)-\theta\|\leq\delta_\theta\},\\ \bar{T}_{\delta_\theta}(\omega) &\coloneqq \sup_{\tilde{L}\in\mathcal{B}_{\epsilon_\theta}(L^\theta)} T_{\delta_\theta}(\tilde{L},\omega). \end{aligned}$$

A consequence of the result in (3.18) is that for any stabilizing $\tilde{L}$, $\mathbb{P}(T_{\delta_\theta}(\tilde{L},\omega)<\infty)=1$ and consequently $\mathbb{P}(\bar{T}_{\delta_\theta}(\omega)<\infty)=1$. For any $T>0$, define

$$\mathcal{A}_{\delta_\theta}(T) \coloneqq \{\omega\in\Omega : \bar{T}_{\delta_\theta}(\omega)\leq T-1\}.$$

Proposition 3.4. *The set $\mathcal{A}_{\delta_\theta}(T)$ satisfies following properties:*

1. *If $T<T'$, we have $\mathcal{A}_{\delta_\theta}(T)\subseteq\mathcal{A}_{\delta_\theta}(T')$.*
2. *For any $\omega\in\Omega$, there exists a T_0 such that: $\omega\in\mathcal{A}_{\delta_\theta}(T_0)\subseteq\Omega$, μ-a.s.*
3. *There exists a $T_0<\infty$ such that $\Omega=\mathcal{A}_{\delta_\theta}(T_0)$, μ-a.s.*

Proof. The first statement follows from the definition of $\mathcal{A}_{\delta_\theta}(T)$. The second statement is a consequence of Theorem 2.2. To prove that, we use contradiction. Suppose there exists a subset $\Omega'\subset\Omega$ with a non-zero measure $\mu(\Omega')>0$ such that T_0 does not exist. This implies that the switched least squares algorithm is not convergent on the set Ω'. Since we assumed that $\mu(\Omega')>0$, this implies that switched least squares is not a strongly consistent estimator and that contradicts Theorem 2.2. As a result, any sample path $\omega\in\Omega$ for which T_0 does not exist should have a zero measure. To show the third statement, similarly assume there exists a subset $\Omega'\subset\Omega$ with a non-zero measure $\mu(\Omega')>0$ such that T_0 is not finite. This implies that the convergence rate in (3.18) is violated infinitely often on the set Ω'; however, this is in contradiction with the convergence rate established in Theorem 2.2. □

Theorem 3.5. ***(Sufficient condition for stability)*** *Suppose the initial stabilizing controller $\{\bar{L}_s\}_{s\in\mathcal{S}}\in\mathcal{B}_{\epsilon_\theta}(L^\theta)$, then the results of Theorem 3.3 and 3.4 are valid on the set $\mathcal{A}_{\delta_\theta}(T^{(0)})$.*

The proof is presented in Appendix 3.D.

3.7 Conclusion and Future Directions

In this chapter, we investigated the problem of simultaneous learning and control of a Markov jump linear system using complete state observation. We derived an almost sure regret decomposition for the general class of adaptive linear policies with persistence of excitation. We proposed a version of certainty equivalence controller which uses the switched least squares method for the closed-loop system identification. Our analysis shows that the error of the system identification method is $\mathcal{O}(\sqrt{\log(T)/T})$, and the regret of certainty equivalence controller is upper-bounded by $\mathcal{O}(\sqrt{T}\log(T))$ almost surely. Our guarantees are stated for specific subset of Ω. We show we can make this subset arbitrary large by increasing $T^{(0)}$. Finding an algorithm with performance guarantees independent of the set $\mathcal{A}_{\delta_\theta}(T^{(0)})$, extending these results to the case of partial observation and analyzing algorithms such as Thompson sampling with the tool developed in Theorem 3.2 is left for future works.

Appendices to Chapter 3

3.A Proof of Theorem 3.2

We start with the following completion of squares lemma, which is adapted from [117, Lemma 6.1].

Lemma 3.2. *For $x \in \mathbb{R}^n$ and $u \in \mathbb{R}^m$ and matrices A, B, S, R with appropriate dimensions, we have*

$$\begin{aligned}&u^\intercal R u + (Ax + Bu)^\intercal P(Ax + Bu) + x^\intercal Q x = \\ &(u + L(P, R, A, B)x)^\intercal [R + B^\intercal P B](u + L(P, R, A, B)x) + x^\intercal K(P, A, B, R, Q)x,\end{aligned}$$

where

$$\begin{aligned}L(P, R, A, B) &\coloneqq -(R + B^\intercal P B)^{-1} B^\intercal P A. \\ K(P, A, B, R, Q) &\coloneqq Q + A^\intercal P A - A^\intercal P B (R + B^\intercal P B)^{-1} B^\intercal P A.\end{aligned}$$

Remark 3.2. *Notice that in* (3.8)*, we have:*

$$L_s = L(\bar{P}_s, R_s, A_s, B_s), \quad P_s = K(\bar{P}_s, A_s, B_s, R_s, Q_s). \tag{0.19}$$

We assume that π and ω_1 are fixed and do not explicitly include their dependence on the terms. Instead, we will use x_t as a short-hand for $x_t^\pi(\omega)$ and x_t^* as a short-hand for $x_t^{\pi_\theta^*}(\omega)$. We also use $\tilde{s}_t$ instead of $s_t(\omega_2)$, where we use the superscript tilde to highlight the fact that we are not referring to a specific realization of the discrete state at time t rather marginalizing over all possible realizations. By recursively applying completion of squares (Lemma 3.2), we can show the following:

Lemma 3.3. *For any policy π we have*

$$\begin{aligned}&\int_{\Omega_2} \left[\sum_{t=1}^{T} c(x_t, s_t, u_t) + x_{T+1}^\intercal P_{s_{T+1}} x_{T+1} \right] \mu_2(d\omega_2) \\ &\quad = \int_{\Omega_2} \left[x_1^\intercal \bar{P}_{\tilde{s}_1} x_1 + \sum_{t=1}^{T} (u_t + L_{\tilde{s}_t} x_t)^\intercal [R_{\tilde{s}_t} + B_{\tilde{s}_t}^\intercal \bar{P}_{\tilde{s}_t} B_{\tilde{s}_t}](u_t + L_{\tilde{s}_t} x_t) \right. \\ &\qquad \left. + \sum_{t=1}^{T} \left[2 w_t^\intercal \bar{P}_{\tilde{s}_t} (A_{\tilde{s}_t} x_t + B_{\tilde{s}_t} u_t) + w_t^\intercal \bar{P}_{\tilde{s}_t} w_t \right] \right] \mu_2(d\omega_2).\end{aligned} \tag{0.20}$$

The proof is presented in Appendix 3.E. Using the decomposition in (0.20) in the expression for regret, and substituting $u_t = -\hat{L}_{\tilde{s}_t}(t)x_t + \nu_t$ for policy π and substituting $u_t = -L_{\tilde{s}_t}x_t$ for policy π^*, we get the following:

Lemma 3.4. *For any adaptive linear policy π with persistence of excitation, we have*

$$\begin{aligned} R_T^\pi(\omega_1) = \int_{\Omega_2} \Bigg[& \sum_{t=1}^{T} x_t^\intercal (\hat{L}_{\tilde{s}_t}(t) - L_{\tilde{s}_t})^\intercal [R_{\tilde{s}_t} + B_{\tilde{s}_t}^\intercal \bar{P}_{\tilde{s}_t} B_{\tilde{s}_t}] (\hat{L}_{\tilde{s}_t}(t) - L_{\tilde{s}_t}) x_t \\ & + \sum_{t=1}^{T} \left[\nu_t^\intercal [R_{\tilde{s}_t} + B_{\tilde{s}_t}^\intercal \bar{P}_{\tilde{s}_t} B_{\tilde{s}_t}] \nu_t + 2\nu_t^\intercal [R_{\tilde{s}_t} + B_{\tilde{s}_t}^\intercal \bar{P}_{\tilde{s}_t} B_{\tilde{s}_t}] (\hat{L}_{\tilde{s}_t}(t) - L_{\tilde{s}_t}) x_t \right] \\ & + \sum_{t=1}^{T} 2 w_t^\intercal \bar{P}_{\tilde{s}_t} \left[(A_{\tilde{s}_t} - B_{\tilde{s}_t} L_{\tilde{s}_t})(x_t - x_t^*) - B_{\tilde{s}_t} (\hat{L}_{\tilde{s}_t}(t) - L_{\tilde{s}_t}) x_t + B_{\tilde{s}_t} \nu_t \right] \\ & + [(x_{T+1}^*)^\intercal \bar{P}_{\tilde{s}_{T+1}} x_{T+1}^* - x_{T+1}^\intercal \bar{P}_{\tilde{s}_{T+1}} x_{T+1}] \Bigg] \mu_2(d\omega_2). \end{aligned} \tag{0.21}$$

We first recall the following result [52, Corollary 10].

Lemma 3.5. *Given a filtration $\{\mathcal{F}_t\}_{t\geq 1}$, suppose w_t is a martingale difference process adapted to $\{\mathcal{F}_t\}_{t\geq 1}$ and y_{t+1} is $\mathcal{F}_t$-measurable. Then,*

$$\sum_{t=1}^{T} y_t^\intercal w_t = \mathcal{O}\Big(\sqrt{Y_T \log(Y_T)} \Big), \quad a.s.$$

where $Y_T = \sum_{t=1}^{T} y_t^\intercal y_t$.

An implication of Lemma 3.5 is that

$$\begin{aligned} & \int_{\Omega_2} \left[\sum_{t=1}^{T} w_t^\intercal \bar{P}_{\tilde{s}_t} B_{\tilde{s}_t} (\hat{L}_{\tilde{s}_t}(t) - L_{\tilde{s}_t}) x_t \right] \mu_2(d\omega_2) \\ & \qquad = \mathcal{O}\Big(\sqrt{\mathcal{R}_{1,T}^\pi(\omega_1) \log \mathcal{R}_{1,T}^\pi(\omega_1)} \Big), \end{aligned} \tag{0.22}$$

where $\mathcal{R}_{1,T}^\pi(\omega_1)$ is defined in Theorem 3.2. By the same argument, we also have

$$\begin{aligned} & \int_{\Omega_2} \left[\sum_{t=1}^{T} \nu_t^\intercal \bar{P}_{\tilde{s}_t} B_{\tilde{s}_t} (\hat{L}_{\tilde{s}_t}(t) - L_{\tilde{s}_t}) x_t \right] \mu_2(d\omega_2) \\ & \qquad = \mathcal{O}\Big(\sqrt{\mathcal{R}_{1,T}^\pi(\omega_1) \log \mathcal{R}_{1,T}^\pi(\omega_1)} \Big), \end{aligned} \tag{0.23}$$

and

$$\int_{\Omega_2} \left[\sum_{t=1}^{T} 2w_t^\intercal \bar{P}_{\tilde{s}_t} B_{\tilde{s}_t} \nu_t \right] \mu_2(d\omega_2) = \mathcal{O}\Big(\sqrt{\mathcal{R}_{2,T}^{\pi}(\omega_1) \log \mathcal{R}_{2,T}^{\pi}(\omega_1)} \Big). \tag{0.24}$$

Now, by Prop. 3.3, the autonomous MJLS system $(\{A_s - B_s L_s\}_{s\in\mathcal{S}}, H)$ is MSS. Based on the fact that MSS implies exponential stochastic stability (Prop. 3.1), we can show that

$$\int_{\Omega_2} \left[\sum_{t=1}^{T} w_t^\intercal \bar{P}_{\tilde{s}_t} (A_{\tilde{s}_t} - B_{\tilde{s}_t} L_{\tilde{s}_t})(x_t - x_t^*) \right] \mu_2(d\omega_2) = \mathcal{O}\Big(\int_{\Omega_2} \left[\sum_{t=1}^{T} w_t^\intercal \bar{P}_{\tilde{s}_t} B_{\tilde{s}_t} (\hat{L}_{\tilde{s}_t}(t) - L_{\tilde{s}_t}) x_t \right] \mu_2(d\omega_2) \Big) \tag{0.25}$$

which is therefore also upper bounded by the right hand side of (0.22). The result of Theorem 3.2 then follows from substituting (0.22)–(0.25) in Lemma 3.4.

3.B Proof of Theorem 3.3

The proof of this theorem is based on a notion of stability in the average sense defined in Chapter 2.

Definition 3.7. *Let $\{x_t\}_{t\geq 1}$ denote the state process corresponding to the MJLS system $A_{s_t}x_t + w_t$. We say this system is stable in the average sense, if: $\sum_{t=1}^{T} \|x_t\|^2 = \mathcal{O}(T) \quad \mu$-a.s.,*

We recall Proposition 2.3 from Chapter 2.

Proposition 3.5. *If the MJLS system $(\{A_s\}_{s\in\mathcal{S}}, H)$ is MSS, then the MJLS system: $A_{s_t}x_t + w_t$ is stable in the average sense.*

Proof. By the assumptions in Theorem 3.3, we know $(\{A_s - B_s L_s\}_{s\in\mathcal{S}}, H)$ is MSS; therefore, by Proposition 3.5, and the fact that σ_t^2 is finite, we get that MJLS $x_{t+1} = (A_{s_t} - B_{s_t}\hat{L}_{s_t}^{(k)})x_t + B_{s_t}\nu_t + w_t$ is stable in the average sense. Recall that $\eta_{s_t}^{(k)} \coloneqq \left[A_{s_t} - B_{s_t}\hat{L}_{s_t}^{(k)}, B_{s_t}\right]$, and $z_t \coloneqq \begin{bmatrix} x_t \\ \nu_t \end{bmatrix}$, and we have:

$$x_{t+1} = \eta_{s_t}^{(k)} z_t + w_t. \tag{0.26}$$

Let $\mathcal{T}_{i,T}^{(k)} = \{t^{(k)} \leq t < t^{(k)} + T : s_t = i\}$ denote the time indices until the time T, when the discrete state of the system equals i at the k-th episode. Note that for $t \in \mathcal{T}_{i,T}^{(k)}$, $\eta_{s_t} = \eta_i$.

Therefore, we have:

$$\hat{\eta}_{i,T}^{(k)} \coloneqq \arg\min_{\eta} \sum_{t \in \mathcal{T}_{i,T}^{(k)}} \|x_{t+1} - \eta_i z_t\|^2, \quad \forall i \in \{1, \dots, d\}.$$

Let $Z_{i,T}$ denote $\sum_{t \in \mathcal{T}_{i,T}^{(k)}} z_t z_t^\intercal$, which we call the unnormalized empirical covariance of the augmented state process when $s_t = i$. Now we look at $\lambda_{\max}(\sum_{t \in \mathcal{T}_{i,T}^{(k)}} z_t z_t^\intercal)$ and $\lambda_{\min}(\sum_{t \in \mathcal{T}_{i,T}^{(k)}} z_t z_t^\intercal)$. We have:

$$\begin{aligned}\lambda_{\max}\Big(\sum_{t \in \mathcal{T}_{i,T}^{(k)}} z_t z_t^\intercal\Big) &\le \operatorname{tr}\Big(\sum_{t \in \mathcal{T}_{i,T}^{(k)}} z_t z_t^\intercal\Big) \\ &= \sum_{t \in \mathcal{T}_{i,T}^{(k)}} \|z_t\|^2 \le \sum_{t=1}^{T} \|z_t\|^2\end{aligned}$$

By Proposition 3.5, we know $\sum_{t=1}^{T} \|x_t\|^2 = \mathcal{O}(T)\, \mu\text{-}a.s.$ and by [52, Eq. 3.1] we know $\sum_{t=1}^{T} \|\nu_t\|^2 = \mathcal{O}(T)\, \mu\text{-}a.s.$, which implies:

$$\lambda_{\max}\Big(\sum_{t \in \mathcal{T}_{i,T}^{(k)}} z_t z_t^\intercal\Big) = \mathcal{O}(T) \quad \mu\text{-}a.s.$$

On the other hand, we have:

$$z_t z_t^\intercal = \begin{bmatrix} x_t x_t^\intercal & x_t \nu_t^\intercal \\ \nu_t x_t^\intercal & \nu_t \nu_t^\intercal \end{bmatrix}$$

Similar to Lemma 2.3, we can show $\sum_{t \in \mathcal{T}_{i,T}^{(k)}} \|x_t \nu_t^\intercal + \nu_t x_t^\intercal\| = o(T)\, \mu\text{-}a.s.$; therefore,

$$\lambda_{\min}\Big(\sum_{t \in \mathcal{T}_{i,T}^{(k)}} z_t z_t^\intercal\Big) = o\Big(\min\Big\{\lambda_{\min}\Big(\sum_{t \in \mathcal{T}_{i,T}^{(k)}} x_t x_t^\intercal\Big), \lambda_{\min}\Big(\sum_{t \in \mathcal{T}_{i,T}^{(k)}} \nu_t \nu_t^\intercal\Big)\Big\}\Big) \quad \mu\text{-}a.s.$$

By Proposition 2.1, we know $\liminf_{T \to \infty} \lambda_{\min}(\sum_{t \in \mathcal{T}_{i,T}} x_t x_t^\intercal)/\mathcal{T}_{i,T}^{(k)} \ge 0 \quad \mu\text{-}a.s.$, and since $\sigma_{(k)}^2 > 0$, we get $\liminf_{k \to \infty} \lambda_{\min}(\sum_{t \in \mathcal{T}_{i,T}^{(k)}} \nu_t \nu_t^\intercal)/\mathcal{T}_{i,T}^{(k)} \ge \sigma_{(k)}^2 \quad \mu\text{-}a.s.$ Therefore,

$$\lambda_{\min}\Big(\sum_{t \in \mathcal{T}_{i,T}^{(k)}} z_t z_t^\intercal\Big) > \sigma_{(k)}^2 \mathcal{T}_{i,T} \quad \mu\text{-}a.s.$$

Given a filtration $\{\mathcal{G}_t\}_{t \ge 0}$, consider the following regression model:

$$y_t = \beta^\intercal z_t + w_t, \quad t \ge 0, \tag{0.27}$$

where $\beta \in \mathbb{R}^n$ is an unknown parameter, $z_t \in \mathbb{R}^n$ is $\mathcal{G}_{t-1}$-measurable covariate process, y_t is the observation process, and $w_t \in \mathbb{R}$ is a noise process. Then the least squares estimate $\hat{\beta}_T$ of β is given by:

$$\hat{\beta}_T = \arg\min_{\beta^\intercal} \sum_{\tau=0}^{T} \|y_\tau - \beta^\intercal z_\tau\|^2. \tag{0.28}$$

Theorem 2.4 characterizes the rate of convergence of $\hat{\beta}_T$ to β in terms of unnormalized covariance matrix of covariates $Z_T \coloneqq \sum_{\tau=0}^{T} z_\tau z_\tau^\intercal$. Therefore, by Theorem 2.4, and the fact that $\sigma_{(k)}^2 \mathcal{T}_{i,T} = \mathcal{O}(\sigma_{(k)}^2 T)$ we get that:

$$\lim_{k\to\infty} \|\hat{\theta}_{(k)} - \theta\| = \mathcal{O}\big(\sqrt{\log(T^{(k)})/\sigma^{(k)}T^{(k)}}\big) \quad \mu_1\text{-}a.s.$$

□

3.C Proof of Theorem 3.4

Lemma 3.6. *The regret in the k-th episode satisfies:*

$$\limsup_{k\to\infty} \frac{\int_{\Omega_2} \big[\sum_{\tau=t^{(k)}}^{t^{(k)}+T^{(k)}-1} c(x_\tau, s_\tau, u_\tau)\big]\mu_2(d\omega_2)}{(T^{(k-1)})^{1/2}\log(T^{(k-1)})} < \infty \quad \mu\text{-}a.s.$$

Proof. In the k-th episode, $\hat{L}_s^{(k)}$ is computed based on the estimate $\hat{\theta}^{(k-1)}$. Assumption 3.3 implies that on the set $\mathcal{A}_0$, the set of the gains $\{\hat{L}_s^{(k)}\}_{s\in\mathcal{S}}$ is stabilizing for all $k \geq 0$. Setting $\sigma_{(k-1)}^2 = 1/\sqrt{T^{(k-1)}}$ in Theorem 3.3 implies:

$$\limsup_{k\to\infty} \frac{\|\hat{\theta}^{(k-1)} - \theta\|}{\sqrt{\log(T^{(k-1)})}/(T^{(k-1)})^{1/4}} < \infty, \quad \mu\text{-}a.s. \tag{0.29}$$

By the continuity of the gains $\hat{L}_s^{(k)}$ in the parameter $\hat{\theta}^{(k-1)}$ we get,

$$\limsup_{k\to\infty} \frac{\|\hat{L}_s^{(k)} - L_s\|}{\sqrt{\log(T^{(k-1)})}/(T^{(k-1)})^{1/4}} < \infty, \quad \mu\text{-}a.s. \tag{0.30}$$

In the proof of Theorem 3.3, we established that

$$\sum_{\tau=t^{(k)}}^{t^{(k)}+T^{(k)}-1} \|x_t\|^2 = \mathcal{O}(T^{(k)}) \quad \mu\text{-}a.s.$$

The gain $\hat{L}_s^{(k)}$ is fixed during the episode. Therefore by substituting in $\|\hat{L}_s^{(k)} - L_s\|$, $\sum_{\tau=t^{(k)}}^{t^{(k)}+T^{(k)}-1} \|x_t\|^2$

in the $\mathcal{R}^{\pi}_{1,T}(\omega_1)$, and $\sum_{\tau=t^{(k)}}^{t^{(k)}+T^{(k)}-1} \|\nu_t\|^2 = \mathcal{O}(\sqrt{T^{(k)}})$ in $\mathcal{R}^{\pi}_{2,T}$, we get the desired result. □

Let $T^{(m)} = \sum_{i=0}^{m} T^{(i)} = T^{(0)}(\alpha^{i+1}-1)/(\alpha-1)$, which implies

$$m = \mathcal{O}(\log_{\alpha}(\tilde{T}^{(k)}/T^{(0)}). \tag{0.31}$$

Since T_0 is finite, the regret incurred by the controller is also finite. As a result, we compute the regret starting from episode 1.

Proof. We have

$$\begin{aligned}
\mathcal{R}^{\pi}_{T}(\omega_1) &= \sum_{i=1}^{m} \int_{\Omega_2} \Big[\sum_{\tau=t^{(k)}}^{t^{(k)}+T^{(k)}-1} c(x_\tau, s_\tau, u_\tau) \Big] \mu_2(d\omega_2) \\
&\overset{(a)}{=} \sum_{i=1}^{m} \mathcal{O}(\sqrt{T^{(i-1)}} \log(T^{(i-1)})) \\
&\overset{(b)}{=} \sum_{i=1}^{m} \mathcal{O}(\sqrt{\alpha^{i-1}T_0} \log(\alpha^{i-1}T_0)) \\
&= \sum_{i=1}^{m} \mathcal{O}(\sqrt{\alpha^{i-1}} \log(\alpha^{i-1})) \\
&= \sum_{i=0}^{m-1} \mathcal{O}(\sqrt{\alpha^{i}} \log(\alpha^{i})) \\
&= \sum_{i=0}^{m-1} \mathcal{O}((i)(\sqrt{\alpha})^{i}) \overset{(c)}{=} \mathcal{O}((m-1)\sqrt{\alpha}^{m-1}),
\end{aligned} \tag{0.32}$$

where (a) follows from Lemma 3.6, (b) follows from definition of $T^{(i)}$, and (c) follows from:

$$\begin{aligned}
&\sum_{i=0}^{m-1} \sqrt{\alpha}^{i} = \frac{\sqrt{\alpha}^{m}-1}{\sqrt{\alpha}-1} \\
&\sum_{i=0}^{m-1} i\sqrt{\alpha}^{i} = \sqrt{\alpha}\frac{d}{d\sqrt{\alpha}}\Big(\sum_{i=0}^{m-1} \sqrt{\alpha}^{i}\Big) = \sqrt{\alpha}\frac{d}{d\sqrt{\alpha}}\Big(\frac{\sqrt{\alpha}^{m}-1}{\sqrt{\alpha}-1}\Big) \\
&= \frac{(m-1)\sqrt{\alpha}^{m+1} - m\sqrt{\alpha}^{m} + \sqrt{\alpha}}{(\sqrt{\alpha}-1)^2} = \mathcal{O}((m-1)\sqrt{\alpha}^{m-1}).
\end{aligned}$$

The result of this theorem follows by substituting the expression of m from (0.31), in (0.32). □

3.D Proof of Theorem 3.5

Proof. We prove this result by induction. We show that on the set $\mathcal{A}_{\delta_\theta}(T^{(0)})$ if $\hat{\theta}^{(k)} \in \mathcal{B}_{\delta_\theta}(\theta)$, then $\hat{\theta}^{(k+1)} \in \mathcal{B}_{\delta_\theta}(\theta)$. As the basis of induction, since $\{\bar{L}_s\}_{s\in\mathcal{S}} \in \mathcal{B}_{\epsilon_\theta}(L_\theta)$, then Theorem 3.3 and the definition of $\mathcal{A}_{\delta_\theta}(T^{(0)})$ imply that $\hat{\theta}^{(0)} \in \mathcal{B}_{\delta_\theta}(\theta)$.

Now assume that $\hat{\theta}^{(k)} \in \mathcal{B}_{\delta_\theta}(\theta)$. Lemma 3.1 implies that $\{\hat{L}_s^{(k)}\}_{s\in\mathcal{S}}$ is stabilizing. Moreover, since $T^{(k)} \geq T^{(0)}$, Theorem 3.3 and definition of $\mathcal{A}_{\epsilon_\theta}(T^{(0)})$ imply that $\|\hat{\theta}^{(k+1)} - \theta\| \leq \epsilon_\theta$. Hence $\hat{\theta}^{(k+1)} \in \mathcal{B}_{\delta_\theta}(\theta)$. This completes the proof of the induction step. □

3.E Proof of Lemma 3.3

Proof. We start by adding and subtracting $\int_{\Omega_2}\big[x_T^\intercal P_{s_T} x_T\big]\mu_2(d\omega_2)$ to $\int_{\Omega_2}\big[\sum_{t=1}^T c(x_t,s_t,u_t)\big](d\omega_2)$, we have

$$\int_{\Omega_2}\Big[\sum_{t=1}^{T} c(x_t,s_t,u_t)\Big](d\omega_2) = \int_{\Omega_2}\Big[\sum_{t=1}^{T-1} x_t^\intercal Q_{s_t}x_t + u_t^\intercal R_{s_t}u_t + x_T^\intercal P_{s_T}x_T - x_T^\intercal P_{s_T}x_T\Big]\mu_2(d\omega_2). \tag{0.33}$$

On the other hand, using the telescopic series, we can rewrite $\int_{\Omega_2}\big[x_T^\intercal P_{s_T}x_T\big]\mu_2(d\omega_2)$ as following

$$\int_{\Omega_2}\big[x_T^\intercal P_{s_T}x_T^\intercal\big](d\omega_2) = \int_{\Omega_2}\big[x_1^\intercal P_{s_1}x_1^\intercal + \sum_{t=1}^{T-1}\big(x_{t+1}^\intercal P_{s_{t+1}}x_{t+1}^\intercal - x_t^\intercal P_{s_t}x_t^\intercal\big)\big]\mu_2(d\omega_2) \tag{0.34}$$

By substituting (0.34) in (0.33), we get

$$\int_{\Omega_2}\Big[\sum_{t=1}^{T} c(x_t,s_t,u_t)\Big]\mu_2(d\omega_2) \tag{0.35}$$

$$= \int_{\Omega_2}\Big[x_1^\intercal P_{s_1}x_1 - x_T^\intercal P_{s_T}x_T + \sum_{t=1}^{T-1}\big(x_t^\intercal Q_{s_t}x_t + u_t^\intercal R_{s_t}u_t + x_{t+1}^\intercal P_{s_{t+1}}x_{t+1} - x_t^\intercal P_{s_t}x_t\big)\Big]\mu_2(d\omega_2)$$

$$= \int_{\Omega_2}\Big[x_1^\intercal P_{s_1}x_1 - x_T^\intercal P_{s_T}x_T + \sum_{t=1}^{T-2}\big(x_t^\intercal Q_{s_t}x_t + u_t^\intercal R_{s_t}u_t + x_{t+1}^\intercal P_{s_{t+1}}x_{t+1} - x_t^\intercal P_{s_t}x_t\big)\Big]\mu_2(d\omega_2)$$

$$+ \int_{\Omega_2}\sum_{s_T\in\mathcal{S}} H_{s_{T-1},s_T}\big(x_{T-1}^\intercal Q_{s_{T-1}}x_{T-1} + u_{T-1}^\intercal R_{s_{T-1}}u_{T-1} + x_T^\intercal P_{s_T}x_T - x_{T-1}^\intercal P_{s_{T-1}}x_{T-1}\big)\mu_2(d\omega_2) \tag{0.36}$$

By substituting $x_{t+1} = A_{s_t}x_t + B_{s_t}u_t + w_t$, in (0.36), we can simplify the last term of (0.36) as following

$$
\begin{aligned}
&\int_{\Omega_2} \sum_{s_T\in\mathcal{S}} H_{s_{T-1},s_T}\Big[x_{T-1}^\intercal Q_{s_{T-1}}x_{T-1} + u_{T-1}^\intercal R_{s_{T-1}}u_{T-1} - x_{T-1}^\intercal P_{s_{T-1}}x_{T-1} \\
&+\big(A_{s_{T-1}}x_{T-1} + B_{s_{T-1}}u_{T-1} + w_{T-1}\big)^\intercal P_{s_T}\big(A_{s_{T-1}}x_{T-1} + B_{s_{T-1}}u_{T-1} + w_{T-1}\big)\Big]\mu_2(d\omega_2) \\
&= \int_{\Omega_2}\Big[x_{T-1}^\intercal Q_{s_{T-1}}x_{T-1} + u_{T-1}^\intercal R_{s_{T-1}}u_{T-1} - x_{T-1}^\intercal P_{s_{T-1}}x_{T-1} + w_{T-1}^\intercal \sum_{s_T\in\mathcal{S}} H_{s_{T-1},s_T}P_{s_T}w_{T-1} \\
&\quad + \big(A_{s_{T-1}}x_{T-1} + B_{s_{T-1}}u_{T-1}\big)^\intercal \sum_{s_T\in\mathcal{S}} H_{s_{T-1},s_T}P_{s_T}\big(A_{s_{T-1}}x_{T-1} + B_{s_{T-1}}u_{T-1}\big) \\
&\quad + 2w_{T-1}^\intercal \sum_{s_T\in\mathcal{S}} H_{s_{T-1},s_T}P_{s_T}\big(A_{s_{T-1}}x_{T-1} + B_{s_{T-1}}u_{T-1}\big)\Big]\mu_2(d\omega_2) \\
&= \int_{\Omega_2}\Big[x_{T-1}^\intercal Q_{s_{T-1}}x_{T-1} + u_{T-1}^\intercal R_{s_{T-1}}u_{T-1} - x_{T-1}^\intercal P_{s_{T-1}}x_{T-1} \\
&\quad + \big(A_{s_{T-1}}x_{T-1} + B_{s_{T-1}}u_{T-1}\big)^\intercal \bar{P}_{s_{T-1}}\big(A_{s_{T-1}}x_{T-1} + B_{s_{T-1}}u_{T-1}\big) \\
&\quad + 2w_{T-1}^\intercal \bar{P}_{s_{T-1}}\big(A_{s_{T-1}}x_{T-1} + B_{s_{T-1}}u_{T-1}\big) + w_{T-1}^\intercal \bar{P}_{s_T}w_{T-1}\Big](d\omega_2). \qquad (0.37)
\end{aligned}
$$

By Lemma 3.2, we can simplify (0.37) as following

$$
\begin{aligned}
&\int_{\Omega_2}\big[(u_{T-1} + L(\bar{P}_{s_{T-1}}, R_{s_{T-1}}, A_{s_{T-1}}, B_{s_{T-1}})x_{T-1})^\intercal \\
&[R_{s_{T-1}} + B_{s_{T-1}}^\intercal \bar{P}_{s_{T-1}}B_{s_{T-1}}](u_{T-1} + L(\bar{P}_{s_{T-1}}, R_{s_{T-1}}, A_{s_{T-1}}, B_{s_{T-1}})x_{T-1}) + \\
&x_{T-1}^\intercal K(\bar{P}_{s_{T-1}}, A_{s_{T-1}}, B_{s_{T-1}}, R_{s_{T-1}}, Q_{s_{T-1}})x_{T-1} - x_{T-1}^\intercal P_{s_{T-1}}x_{T-1} + w_{T-1}^\intercal \bar{P}_{s_{T-1}}w_{T-1} \\
&+ 2w_{T-1}^\intercal \bar{P}_{s_{T-1}}\big(A_{s_{T-1}}x_{T-1} + B_{s_{T-1}}u_{T-1}\big)\Big](d\omega_2) \\
&\overset{(a)}{=} \int_{\Omega_2}\big[(u_{T-1} + L(\bar{P}_{s_{T-1}}, R_{s_{T-1}}, A_{s_{T-1}}, B_{s_{T-1}})x_{T-1})^\intercal(d\omega_2) \\
&[R_{s_{T-1}} + B_{s_{T-1}}^\intercal \bar{P}_{s_{T-1}}B_{s_{T-1}}](u_{T-1} + L(\bar{P}_{s_{T-1}}, R_{s_{T-1}}, A_{s_{T-1}}, B_{s_{T-1}})x_{T-1}) \\
&+ 2w_{T-1}^\intercal \bar{P}_{s_{T-1}}\big(A_{s_{T-1}}x_{T-1} + B_{s_{T-1}}u_{T-1}\big) + w_{T-1}^\intercal \bar{P}_{s_{T-1}}w_{T-1}\Big](d\omega_2),
\end{aligned}
$$

where (a) follows from the fact that P_{s_t} is the fixed point solution in (0.19). By repeating the same arguments for other time indices in (0.35), we get

$$\begin{aligned}
&\int_{\Omega_2} \Big[\sum_{t=1}^{T} c(x_t, s_t, u_t)\Big](d\omega_2) = \\
&= \int_{\Omega_2} \Big[x_1^\intercal P_{s_1} x_1 - x_T^\intercal P_{s_T} x_T\Big] + \\
&\int_{\Omega_2} \sum_{t=1}^{T-1} \Big[(u_t + L_{s_t} x_t)^\intercal [R_{s_t} + B_{s_t}^\intercal \bar{P}_{s_t} B_{s_t}](u_t + L_{s_t} x_t) \\
&+ 2 w_t^\intercal \bar{P}_{s_t} \big(A_{s_t} x_t + B_{s_t} u_t\big) + w_t^\intercal \bar{P}_{s_t} w_t \Big](d\omega_2).
\end{aligned}$$

□

Chapter 4

Concentration of Cumulative Reward in Markov Decision Processes

4.1 Overview

In this chapter, we investigate the concentration properties of cumulative reward in finite-state and finite action Markov decision processes. The results of this chapter are available in [118].

4.1.1 Organization

The rest of this chapter is organized as follows. The problem formulation, along with the underlying assumptions, are presented in Section 4.2. The main results for the average reward setting are presented in Section 4.3. The main results for the discounted reward setting are presented in Section 4.4. The main results for the finite-horizon setting are presented in Section 4.5. Our concluding remarks are presented in Section 4.6. Moreover, Appendix 4.A presents a background discussion on Markov chain theory. Appendix 4.B presents a background discussion on concentration of martingale sequences. Proofs of main results are presented in the remaining sections: Appendix 4.C for the average reward MDPs, Appendix 4.D for the discounted reward MDPs, and Appendix 4.E for finite-horizon MDPs.

4.2 Problem Formulation

4.2.1 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 $\pi = (\pi_0, \pi_1, \ldots)$ 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

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

4.2.2 The Average Reward Planning Setup

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

Definition 4.1. *Given a model $\mathcal{M} = (P, r)$, define $\Pi_{\mathrm{SD}} \subseteq \Pi$ to be the set of all stationary deterministic Markov policies, i.e., for any $\pi = (\pi_0, \pi_1, \ldots) \in \Pi_{\mathrm{SD}}$, 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 $\pi \in \Pi_{\mathrm{SD}}$ 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 4.2 (AROE Solvability)**.** *A Model $\mathcal{M} = (P, r)$ is said to be* AROE *(Average Reward Optimality Equation) solvable if there exists a unique optimal long-term average reward $\lambda^* \in \mathbb{R}$ and an optimal differential value function $V^* \colon \mathcal{S} \to \mathbb{R}$ that is unique up to an additive constant that satisfy:*

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

Definition 4.3. *Given a model $\mathcal{M} = (P, r)$, a policy $\pi \in \Pi_{\mathrm{SD}}$ is said to satisfy* ARPE *(Average Reward Policy Evaluation equation) if there exists a unique long-term average reward $\lambda^\pi \in \mathbb{R}$ and a differential value function $V^\pi \colon \mathcal{S} \to \mathbb{R}$ that is unique up to an additive constant that satisfy:*

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

Definition 4.4. *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 4.1 ([119, 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 4.2 ([119, 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 4.1. *The model $\mathcal{M} = (P, r)$ is AROE solvable. Hence, there exists an optimal policy $\pi^* \in \Pi_{\mathrm{AR}}$.*

Proposition 4.1 implies that under Assumption 4.1, $J^*(s)$ is constant. In the rest of this section we assume that Assumption 4.1 always holds and denote $J^*(s)$ by J^*.

4.2.3 Classification of MDPs

We present the main results of this chapter for the policy class Π_{AR} under Assumption 4.1. 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 4.1. We recall definitions of different classes of MDPs. Depending on the properties of states following the policies in Π_{SD}, we can classify MDPs into various classes.

Definition 4.5 ([120]). *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 Appendix 4.A for the details related to the definitions of Markov chains. The following proposition shows the connections between the MDP classes defined above.

Proposition 4.3 ([121, Figure 8.3.1.]). *The following statements hold:*

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

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

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

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

Proposition 4.4 ([121, Table 8.3.1.])**.** *The following properties hold:*

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

4.2.4 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.,
$$\bar{\mathcal{R}}_T^\mu(\omega) \coloneqq TJ^* - R_T^\mu(\omega). \tag{4.1}$$
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.,
$$\mathcal{R}_T^\mu(\omega) \coloneqq R_T^{\pi^*}(\omega) - R_T^\mu(\omega). \tag{4.2}$$

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 chapter, 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 (4.1)–(4.2) that

$$\mathcal{D}_T^\mu(\omega) = \mathcal{R}_T^{\pi^*}(\omega) - TJ^*, \tag{4.3}$$

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 chapter, we characterize asymptotic and non-asymptotic guarantees for the random sequence $\{\mathcal{D}_T(\omega)\}_{T\geq 1}$.

Remark 4.1. *Let $\Pi^* \subset \Pi_{AR}$ denote the set of all optimal policies that satisfy AROE. Assumption 4.1 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^*.*

4.3 Main Results for the Average Reward Setup

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

4.3.1 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

$$H^\pi \coloneqq \mathrm{sp}(V^\pi) = \max_{s\in\mathcal{S}} V^\pi(s) - \min_{s\in\mathcal{S}} V^\pi(s). \tag{4.4}$$

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

$$K^\pi \coloneqq \max_{s,s_+\in\mathcal{S}}\Big|V^\pi(s_+) - \mathbb{E}\big[V^\pi(S_+)\big|s,\pi(s)\big]\Big|. \tag{4.5}$$

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

Remark 4.2. *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 4.6 ([37])**.** *Let the expected number of steps to transition from state s to state s' under a policy $\pi \in \Pi_{\text{SD}}$ be denoted by $T^\pi(s, s')$. The diameter of $\mathcal{M}$ is defined as*

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

Lemma 4.1. *Following relationships hold between the quantities H^π, K^π, and σ^π:*

1. *For any policy $\pi \in \Pi_{\text{AR}}$, we have*

$$\sigma^\pi(s) \leq K^\pi \leq H^\pi < \infty, \quad \forall s \in \mathcal{S}. \tag{4.6}$$

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

$$\sigma^\pi(s) \leq K^\pi \leq H^\pi \leq DR_{\max}, \quad \forall s \in \mathcal{S}. \tag{4.7}$$

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

$$\sigma^*(s) \leq K^* \leq H^* \leq DR_{\max}, \quad \forall s \in \mathcal{S}. \tag{4.8}$$

The proof is presented in Appendix 4.C.1.3.

This section presents three families of results. In Section 4.3.2, we present a set of sample path properties for $R_T^\pi(\omega)$ for any policy $\pi \in \Pi_{\text{AR}}$, depicting both asymptotic and non-asymptotic concentration of $R_T^\pi(\omega)$ around its ergodic mean. In Section 4.3.3, we apply these concentration results to characterize the sample path behavior of the difference between any two policies belonging to Π_{AR}, while in Section 4.3.4, 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 $\mathcal{D}_T(\omega)$.

4.3.2 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 $\pi \in \Pi_{\text{AR}}$. The following theorem characterizes the asymptotic concentration rates of $R_T^\pi(\omega)$, establishing LLN, CLT and LIL.

Definition 4.7. *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 4.1. *For any policy $\pi \in \Pi_{\mathrm{AR}}$ and any initial state $s_0 \in \mathcal{S}$, we have following asymptotic characteristics:*

1. *(Law of Large Numbers) The empirical average of the cumulative reward converges almost surely to J^π, i.e.,*
$$\lim_{T\to\infty} \frac{R_T^\pi(\omega)}{T} = J^\pi, \quad a.s. \tag{4.9}$$
2. *(Central Limit Theorem) Assume that $\mathbb{P}(\Omega_0^\pi) = 1$. Let the stopping time ν_t be defined as $\nu_t \coloneqq \min\Big\{T \geq 1 \,:\, \Sigma_T^\pi \geq t\Big\}$. Then*
$$\lim_{T\to\infty} \frac{R_{\nu_T}^\pi(\omega) - \nu_T J^\pi}{\sqrt{T}} \xrightarrow{(d)} \mathcal{N}(0,1). \tag{4.10}$$
3. *(Law of Iterated Logarithm) For almost all $\omega \in \Omega_0^\pi$, we have*
$$\liminf_{T\to\infty} \frac{R_T^\pi(\omega) - TJ^\pi}{\sqrt{2\Sigma_T^\pi \log\log \Sigma_T^\pi}} = -1, \quad \limsup_{T\to\infty} \frac{R_T^\pi(\omega) - TJ^\pi}{\sqrt{2\Sigma_T^\pi \log\log \Sigma_T^\pi}} = 1. \tag{4.11}$$

The proof is presented in Appendix 4.C.2.

Corollary 4.1. *For any optimal policy $\pi^* \in \Pi^*$, the cumulative reward $R_T^{\pi^*}(\omega)$ satisfies the asymptotic concentration rates in (4.9)–(4.11), where in the LHS, J^π is replaced with J^*.*

Proof. Since π^* is in Π_{AR}, by Theorem 4.1, the optimal policy should satisfy the asymptotic concentration rates in (4.9)–(4.11). □

The proof of Theorem 4.1 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 4.2. *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*

$$\left|R_T^\pi - TJ^\pi - \left(V^\pi(S_0) - V^\pi(S_T)\right)\right| \le K^\pi \sqrt{2T \log \frac{2}{\delta}}. \tag{4.12}$$

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*

$$\left|R_T^\pi - TJ^\pi - \left(V^\pi(S_0) - V^\pi(S_T)\right)\right| \le \max\left\{K^\pi \sqrt{3T\left(2\log\log\frac{3T}{2} + \log\frac{2}{\delta}\right)}, (K^\pi)^2\right\}. \tag{4.13}$$

The proof is presented in Appendix 4.C.3.

Theorem 4.2 establishes a sample path dependent concentration result. The following theorem establishes a sample path independent finite-time concentration of $R_T^\pi(\omega)$ as a function of the statistical properties of V^π.

Theorem 4.3. *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*

$$\left|R_T^\pi - TJ^\pi\right| \le K^\pi \sqrt{2T \log \frac{2}{\delta}} + H^\pi. \tag{4.14}$$

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*

$$\left|R_T^\pi - TJ^\pi\right| \le \max\left\{K^\pi \sqrt{3T\left(2\log\log\frac{3T}{2} + \log\frac{2}{\delta}\right)}, (K^\pi)^2\right\} + H^\pi. \tag{4.15}$$

The proof is presented in Appendix 4.C.4.

Corollary 4.2. *For any optimal policy $\pi^* \in \Pi^*$, the cumulative reward $R_T^{\pi^*}(\omega)$ satisfies the non-asymptotic concentration rates in (4.14)–(4.15), 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 4.3, the optimal policy should satisfy the non-asymptotic concentration rates in (4.14)–(4.15). □

Corollary 4.3. *If $\mathcal{M}$ is unichain or recurrent, then any policy $\pi \in \Pi_{\mathrm{SD}}$ satisfies asymptotic concentration rates in (4.9)–(4.11) and non-asymptotic concentration rates in (4.14)–(4.15).*

Proof. By Prop. 4.4, for the unichain or recurrent model $\mathcal{M}$, we have $\Pi_{\text{AR}} = \Pi_{\text{SD}}$. As a result, any policy π which belongs to Π_{SD} also belongs to Π_{AR}. Therefore, by Theorem 4.1, the asymptotic concentration rates in (4.9)–(4.11) hold for the policy π and by Theorem 4.3, the non-asymptotic rates in (4.14)–(4.15) hold for the policy π. □

Corollary 4.4. *If $\mathcal{M}$ is recurrent, unichain, communicating, or weakly communicating, then every optimal policy $\pi^* \in \Pi^*$ satisfies asymptotic concentration rates in* (4.9)–(4.11) *and non-asymptotic concentration rates in* (4.14)–(4.15)*. (Prop. 4.4 shows that there exists at least one such policy.)*

Proof. By Prop. 4.4, 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 4.1, the asymptotic concentration rates in (4.9)–(4.11) hold for every optimal policy $\pi^* \in \Pi_{\text{AR}}$. Furthermore, by Corollary 4.2, the non-asymptotic concentration rates in (4.14)–(4.15) hold for every optimal policy $\pi^* \in \Pi_{\text{AR}}$. □

In Theorem 4.3, 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 4.5. *Suppose $\mathcal{M}$ is communicating. For any policy $\pi \in \Pi_{\text{AR}}$, following upper-bounds hold:*

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

$$\left|R_T^\pi - TJ^\pi\right| \le DR_{\max}\sqrt{2T\log\frac{2}{\delta}} + DR_{\max}. \tag{4.16}$$

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*

$$\left|R_T^\pi - TJ^\pi\right| \le \max\left\{DR_{\max}\sqrt{3T\left(2\log\log\frac{3T}{2} + \log\frac{2}{\delta}\right)}, (DR_{\max})^2\right\} + DR_{\max}. \tag{4.17}$$

The proof is presented in Appendix 4.C.5.

Corollary 4.6. *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* (4.16)–(4.17)*, where in the LHS, J^π is replaced with J^*.*

The proof is presented in Appendix 4.C.6. In the Corollary 4.5, 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.

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

As an implication of the results presented in the Section 4.3.2, 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 4.7. *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*

$$\Big|\big|R_T^{\pi_1} - R_T^{\pi_2}\big| - \big|TJ^{\pi_1} - TJ^{\pi_2}\big|\Big| \leq K^{\pi_1}\sqrt{2T\log\frac{4}{\delta}} + H^{\pi_1} + K^{\pi_2}\sqrt{2T\log\frac{4}{\delta}} + H^{\pi_2}. \tag{4.18}$$

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

$$\begin{aligned}\Big|\big|R_T^{\pi_1} - R_T^{\pi_2}\big| - \big|TJ^{\pi_1} - TJ^{\pi_2}\big|\Big| \leq{}& \max\Big\{K^{\pi_1}\sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{4}{\delta}\Big)}, (K^{\pi_1})^2\Big\} + H^{\pi_1} \\ &+ \max\Big\{K^{\pi_2}\sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{4}{\delta}\Big)}, (K^{\pi_2})^2\Big\} + H^{\pi_2}.\end{aligned} \tag{4.19}$$

The proof is presented in Appendix 4.C.7.

Corollary 4.8. *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*

$$\big|R_T^{\pi_1^*} - R_T^{\pi_2^*}\big| \leq 2\Big(K^*\sqrt{2T\log\frac{4}{\delta}} + H^*\Big). \tag{4.20}$$

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

$$\left| R_T^{\pi_1^*} - R_T^{\pi_2^*} \right| \leq 2\Big(\max\Big\{ K^* \sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{4}{\delta}\Big)}, (K^*)^2 \Big\} + H^* \Big). \tag{4.21}$$

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, $T\left|J^{\pi_1^*} - J^{\pi_2^*}\right| = 0$. As a result, by Corollary 4.7, the difference $\left|R_T^{\pi_1^*} - R_T^{\pi_2^*}\right|$ satisfies the non-asymptotic concentration rates in Corollary 4.7 with the RHS of (4.18)–(4.19) being simplified to RHS of (4.20)–(4.21). □

Remark 4.3. *Similar to the Corollary 4.5, by imposing the assumption that* $\mathcal{M}$ *is communicating or weakly communicating, we can derive the counterpart of* (4.18)–(4.19) *and* (4.20)–(4.21) *in terms of* $DR_{\max}$ *respectively. For brevity, we omit this result.*

4.3.4 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 4.1, we characterize the asymptotic concentration rates of $\{\mathcal{D}_T(\omega)\}_{T\geq 1}$, establishing LLN, CLT and LIL.

Definition 4.8. *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 4.4. *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*

$$\liminf_{T\to\infty} \frac{\mathcal{D}_T(\omega)}{\sqrt{2\Sigma_T^* \log\log \Sigma_T^*}} = -1, \quad \limsup_{T\to\infty} \frac{\mathcal{D}_T(\omega)}{\sqrt{2\Sigma_T^* \log\log \Sigma_T^*}} = 1. \tag{4.22}$$

Proof is presented in Appendix 4.C.8.

In addition to the asymptotic results presented in Theorem 4.4, we present non-asymptotic guarantees for the sequence $\{\mathcal{D}_T(\omega)\}_{T\geq 1}$. Similar to Theorem 4.3, 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 4.5. *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 Appendix 4.C.9. As mentioned earlier, the difference $\mathcal{D}_T(\omega)$ does not depend on the learning policy μ. Therefore, the results of Theorem 4.5 do not depend on the choice of the learning policy either.

In Theorem 4.5, the upper-bounds are established in terms of K^* and H^*. Similar to Corollary 4.5, 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 4.9. *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 Appendix 4.C.10.

Remark 4.4. *Notice that conditions of Corollary 4.9 are weaker than the conditions of Corollary 4.5. As a result, Corollary 4.9 can be applied to broader classes of $\mathcal{M}$. This difference originates from the difference between items (2) and (3) in Lemma 4.1.*

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 4.6. *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 Appendix 4.C.11.

4.4 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.

4.4.1 System Model

Consider a discounted reward MDP with state space $\mathcal{S}$ and action space $\mathcal{A}$. Similar to Section 4.2, 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 Section 4.2. 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 $\pi \in \Pi_{\text{SD}}$ 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 4.9. *A discounted model $\mathcal{M}$ is said to satisfy DROE (Discounted Reward Op-*

timality Equation) if there exists an optimal discounted value function $V_\gamma^* : \mathcal{S} \to \mathbb{R}$ *that satisfies:*

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

Definition 4.10. *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:*

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

Proposition 4.5 ([119, 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^**.*

4.4.2 Sample Path Characteristics of Any Policy

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

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

$$H^{\pi,\gamma} \coloneqq \mathrm{sp}(V_\gamma^\pi) = \max_{s \in \mathcal{S}} V_\gamma^\pi(s) - \min_{s \in \mathcal{S}} V_\gamma^\pi(s). \tag{4.23}$$

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

$$K^{\pi,\gamma} \coloneqq \max_{s, s_+ \in \mathcal{S}} \Big| V_\gamma^\pi(s_+) - \mathbb{E}\big[V_\gamma^\pi(S_+) \bigm| s, \pi(s) \big] \Big|. \tag{4.24}$$

For any optimal policy $\pi^* \in \Pi_{\mathrm{SD}}$, we denote these corresponding quantities by $H^{*,\gamma}$, and $K^{*,\gamma}$. Similar to the results in Theorem 4.2 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 4.7. *For any policy $\pi \in \Pi_{\mathrm{SD}}$ and any $s \in \mathcal{S}$, we have:*

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

$$\Big|R_T^{\pi,\gamma} - \big(V_\gamma^\pi(S_0) - \gamma^T V_\gamma^\pi(S_T)\big)\Big| \le K^{\pi,\gamma}\sqrt{2f^\gamma(T)\log\frac{2}{\delta}}. \tag{4.25}$$

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

$$\begin{aligned}&\Big|R_T^{\pi,\gamma} - \big(V_\gamma^\pi(S_0) - \gamma^T V_\gamma^\pi(S_T)\big)\Big| \\ &\qquad \le \max\left\{K^{\pi,\gamma}\sqrt{3f^\gamma(T)\Big(2\log\log\big(\frac{3}{2}f^\gamma(T)\big) + \log\frac{2}{\delta}\Big)}, (K^{\pi,\gamma})^2\right\}.\end{aligned} \tag{4.26}$$

The proof is presented in Appendix 4.D.1.

Corollary 4.10. *For any policy $\pi \in \Pi_{\mathrm{SD}}$ and any $s \in \mathcal{S}$, we have:*

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

$$\Big|R_T^{\pi,\gamma} - V_\gamma^\pi(S_0)\Big| \le K^{\pi,\gamma}\sqrt{2f^\gamma(T)\log\frac{2}{\delta}} + \frac{\gamma^T}{1-\gamma}R_{\max}. \tag{4.27}$$

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

$$\begin{aligned}&\Big|R_T^{\pi,\gamma} - V_\gamma^\pi(S_0)\Big| \\ &\qquad \le \max\left\{K^{\pi,\gamma}\sqrt{3f^\gamma(T)\big(2\log\log(\frac{3}{2}f^\gamma(T)) + \log\frac{2}{\delta}\big)}, (K^{\pi,\gamma})^2\right\} + \frac{\gamma^T}{1-\gamma}R_{\max}.\end{aligned} \tag{4.28}$$

The proof is presented in Appendix 4.D.2.

Corollary 4.11. *For any optimal policy $\pi^* \in \Pi_{\text{SD}}$, the discounted cumulative reward $R_T^{\pi^*,\gamma}(\omega)$ satisfies the non-asymptotic concentration rates in (4.25)–(4.28), 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 4.7 and Corollary 4.10, the optimal policy satisfies the non-asymptotic concentration rates in (4.25)–(4.28). □

4.4.3 Sample Path Behavior of Performance Difference of Two Stationary Policies

As an implication of the results presented in the Section 4.4.2, 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 4.12. *Consider two policies $\pi_1, \pi_2 \in \Pi_{\text{SD}}$. Let $\{S_t^{\pi_1}\}_{t\geq 0}$ and $\{S_t^{\pi_2}\}_{t\geq 0}$ denote the random sequences of the states encountered by policy π_1 and π_2 respectively. Following upper-bounds hold for the difference between the discounted cumulative reward received by the two policies.*

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

$$\begin{aligned}&\Big|\big|R_T^{\pi_1,\gamma} - R_T^{\pi_2,\gamma}\big| - \big|\big[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\big] - \big[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\big]\big|\Big| \\ &\leq K^{\pi_1,\gamma}\sqrt{2f^\gamma(T)\log\frac{4}{\delta}} + K^{\pi_2,\gamma}\sqrt{2f^\gamma(T)\log\frac{4}{\delta}}. \end{aligned} \tag{4.29}$$

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

$$T_0^{\pi_i}(\frac{\delta}{2}) \coloneqq \min\Big\{T' \geq 1 : f^\gamma(T') > \frac{173}{K^{\pi_i,\gamma}}\log\frac{8}{\delta}\Big\}, \quad i \in \{1,2\}. \tag{4.30}$$

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

$$\begin{aligned}&\Big|\big|R_T^{\pi_1,\gamma} - R_T^{\pi_2,\gamma}\big| - \big|\big[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\big] - \big[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\big]\big|\Big| \\ &\leq \max\Big\{K^{\pi_1,\gamma}\sqrt{3f^\gamma(T)\Big(2\log\log\frac{3}{2}f^\gamma(T) + \log\frac{4}{\delta}\Big)}, (K^{\pi_1,\gamma})^2\Big\} \\ &+ \max\Big\{K^{\pi_2,\gamma}\sqrt{3f^\gamma(T)\Big(2\log\log\frac{3}{2}f^\gamma(T) + \log\frac{4}{\delta}\Big)}, (K^{\pi_2,\gamma})^2\Big\}. \end{aligned} \tag{4.31}$$

The proof if presented in Appendix 4.D.3.

Corollary 4.13. *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 $\left|R_T^{\pi_1^*,\gamma} - R_T^{\pi_2^*,\gamma}\right|$, we have:*

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

$$\Big|\left|R_T^{\pi_1^*,\gamma} - R_T^{\pi_2^*,\gamma}\right| - \gamma^T\left|V_\gamma^*(S_T^{\pi_2^*}) - V_\gamma^*(S_T^{\pi_1^*})\right|\Big| \leq 2\Big(K^{*,\gamma}\sqrt{2f^\gamma(T)\log\frac{4}{\delta}}\Big). \tag{4.32}$$

2. *Consider $T_0^{\pi^*}(\frac{\delta}{2})$ defined in (4.30). For any $\delta \in (0,1)$, for all $T \geq T_0^{\pi^*}(\frac{\delta}{2})$, with probability at least $1-\delta$, we have*

$$\begin{aligned}&\Big|\left|R_T^{\pi_1^*,\gamma} - R_T^{\pi_2^*,\gamma}\right| - \gamma^T\left|V_\gamma^*(S_T^{\pi_2^*}) - V_\gamma^*(S_T^{\pi_1^*})\right|\Big| \\ &\leq 2\Big(\max\Big\{K^{*,\gamma}\sqrt{3f^\gamma(T)\Big(2\log\log\big(\frac{3}{2}f^\gamma(T)\big) + \log\frac{4}{\delta}\Big)}, (K^{*,\gamma})^2\Big\}\Big).\end{aligned} \tag{4.33}$$

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 4.12, the difference $\left|R_T^{\pi_1^*,\gamma} - R_T^{\pi_2^*,\gamma}\right|$ satisfies the non-asymptotic concentration rates in Corollary 4.12 with the RHS of (4.29) and (4.31) being simplified to RHS of (4.32)–(4.33). □

4.4.4 Vanishing Discount Analysis

In order to observe the connection between the upper-bounds established in Theorem 4.2 and Theorem 4.7, 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 4.14. *For any policy $\pi \in \Pi_{\mathrm{AR}}$, we have the following asymptotic relations between the bounds in Theorem 4.2 and Theorem 4.7.*

1. *As γ goes to 1 from below, the quantity in the LHS of (4.25)–(4.26) converges to the LHS of (4.12), 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 (4.25) converges to the RHS in (4.12), i.e.,*

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

3. *As γ goes to 1 from below, the RHS in (4.26) converges to the RHS in (4.13), i.e.,*

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

Proof is presented in Appendix 4.D.4.

Remark 4.5. *The non-asymptotic characterizations are established in Theorem 4.7. 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 4.14 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.*

4.5 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.

4.5.1 System Model

Consider an MDP with state space $\mathcal{S}$ and action space $\mathcal{A}$. Similar to Section 4.2, 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 Section 4.2.

Definition 4.11. *Given a model $\mathcal{M} = (P, r, h)$, define Π_{FD} to be the set of finite-horizon deterministic policies, i.e., for any $\pi = (\pi_0, \pi_1, \ldots, \pi_h) \in \Pi_{\text{FD}}$, we have $\pi_t : \mathcal{S} \to \mathcal{A}$ (i.e., $A_t = \pi_t(S_t)$), but π_t may depend upon t.*

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

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

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

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

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 4.12. *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 4.13. *Given a policy $\pi \in \Pi_{\text{FD}}$, the sequence of finite-horizon value functions $\{V_t^{\pi,h}\}_{t=0}^{h+1} : \mathcal{S} \to \mathbb{R}$ corresponding to the policy π is defined as follows*

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

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

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

Proposition 4.6 ([122])**.** *Let $\pi^* = (\pi_0^*, \pi_1^*, \ldots, \pi_h^*) \in \Pi_{\text{FD}}$ 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 $s \in \mathcal{S}$, $J^{\pi^*,h}(s) = J^{*,h}(s)$.*

4.5.2 Sample Path Characteristics of Any Policy

For any policy $\pi \in \Pi_{\text{FD}}$, 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

$$H_t^{\pi,h} \coloneqq \text{sp}(V_t^{\pi,h}), \quad \forall t \in \{0, 1, \ldots, h\}. \tag{4.34}$$

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

$$K_t^{\pi,h} \coloneqq \max_{s,s_+} \Big| V_t^{\pi,h}(s_+) - \mathbb{E}\big[V_t^{\pi,h}(S_+) \bigm| s, \pi_t(s)\big] \Big|, \quad \forall t \in \{0, 1, \ldots, h\}. \tag{4.35}$$

Similar to the results in Theorem 4.3 and Theorem 4.7 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

$$\bar{K}_T^{\pi,h} = \max_{0 \le t \le T} K_t^{\pi,h}, \quad \bar{H}_T^{\pi,h} = \max_{0 \le t \le T} H_t^{\pi,h}, \tag{4.36}$$

and let

$$g^{\pi,h}(T) \coloneqq \frac{\sum_{t=1}^{T} (K_t^{\pi,h})^2}{\big(\bar{K}_T^{\pi,h}\big)^2}. \tag{4.37}$$

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

that

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

In this section, we show that with high-probability $R_T^{\pi,h}$ concentrates around $V_0^{\pi,h}(S_0) - V_T^{\pi,h}(S_T)$ and characterize the concentration rate. Following theorem is analogous to the concentration bounds in average reward setup given in Theorem 4.2 and concentration bounds in discounted reward setup given in Theorem 4.7.

Theorem 4.8. *For any policy $\pi \in \Pi_{\text{FD}}$, we have:*

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

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

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

$$T_0^{\pi,h}(\delta) := \min\Big\{T' \geq 1 : g^{\pi,h}(T') \geq 173\log\frac{4}{\delta}\Big\}. \tag{4.38}$$

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

$$\begin{aligned}&\Big|R_T^{\pi,h} - \big(V_0^{\pi,h}(S_0) - V_T^{\pi,h}(S_T)\big)\Big| \\ &\qquad \leq \max\Big\{\bar{K}_T^{\pi,h}\sqrt{3g^{\pi,h}(T)\Big(2\log\log(\frac{3}{2}g^{\pi,h}(T)) + \log\frac{2}{\delta}\Big)}, (\bar{K}_T^{\pi,h})^2\Big\}.\end{aligned} \tag{4.39}$$

The proof is presented in Appendix 4.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 4.3 and concentration bounds in the discounted reward setup given in Corollary 4.10.

Corollary 4.15. *For any policy $\pi \in \Pi_{\text{FD}}$, we have:*

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

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

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 (4.38). Then*

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

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

The proof is presented in Appendix 4.E.2.

4.5.3 Sample Path Behavior of Performance Difference of Two Policies

As an implication of the results presented in Section 4.5.2, 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 4.16. *Consider two policies* $\pi_1, \pi_2 \in \Pi_{\mathrm{FD}}$*. Let* $\{S_t^{\pi_1}\}_{t=0}^h$ *and* $\{S_t^{\pi_2}\}_{t=0}^h$ *denote the random sequences of the states encountered by 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*

$$\begin{aligned}&\left|\left|R_T^{\pi_1,h} - R_T^{\pi_2,h}\right| - \left|\left[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\right] - \left[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\right]\right|\right| \\ &\le \bar{K}_T^{\pi_1,h}\sqrt{2g^{\pi_1,h}(T)\log\frac{4}{\delta}} + \bar{K}_T^{\pi_2,h}\sqrt{2g^{\pi_2,h}(T)\log\frac{4}{\delta}}. \end{aligned} \tag{4.40}$$

2. *For any* $\delta \in (0,1)$*, if* $\min\left\{g^{\pi_1,h}(h), g^{\pi_2,h}(h)\right\} \ge 173\log\frac{8}{\delta}$*, define* $T_0^{\pi,h}(\delta)$ *as specified in* (4.38) *and let*

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

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

$$\begin{aligned}&\left|\left|R_T^{\pi_1,h} - R_T^{\pi_2,h}\right| - \left|\left[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\right] - \left[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\right]\right|\right| \\ &\le \max\left\{\bar{K}_T^{\pi_1,h}\sqrt{3g^{\pi_1,h}(T)\Big(2\log\log(\frac{3}{2}g^{\pi_1,h}(T)) + \log\frac{4}{\delta}\Big)}, (\bar{K}_T^{\pi_1,h})^2\right\} \\ &+ \max\left\{\bar{K}_T^{\pi_2,h}\sqrt{3g^{\pi_2,h}(T)\Big(2\log\log(\frac{3}{2}g^{\pi_2,h}(T)) + \log\frac{4}{\delta}\Big)}, (\bar{K}_T^{\pi_2,h})^2\right\}. \end{aligned} \tag{4.41}$$

The proof is presented in Appendix 4.E.3.

Corollary 4.17. *Consider two optimal 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 states encountered by optimal policies π_1^* and π_2^*. To simplify the expression, we assume the system starts at a fixed initial state, i.e., $S_0^{\pi_1^*} = S_0^{\pi_2^*}$. Then for the difference between the cumulative rewards received by the two optimal policies $\left|R_T^{\pi_1^*,h} - R_T^{\pi_2^*,h}\right|$, we have:*

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

$$\left|\left|R_T^{\pi_1^*,h} - R_T^{\pi_2^*,h}\right| - \left|V_T^{*,h}(S_T^{\pi_2^*}) - V_T^{*,h}(S_T^{\pi_1^*})\right|\right| \leq 2\Big(\bar{K}_T^{*,h}\sqrt{2g^{*,h}(T)\log\frac{4}{\delta}}\Big). \tag{4.42}$$

2. *For any $\delta \in (0,1)$, if $g^{*,h}(h) \geq 173\log\frac{4}{\delta}$, define $T_0^{\pi^*,h}(\delta)$ as specified in (4.38). Then with probability at least $1-\delta$, for all $T_0^{\pi^*,h}(\delta) \leq T \leq h+1$, we have*

$$\begin{aligned}&\left|\left|R_T^{\pi_1^*,h} - R_T^{\pi_2^*,h}\right| - \left|V_T^{*,h}(S_T^{\pi_2^*}) - V_T^{*,h}(S_T^{\pi_1^*})\right|\right| \\ \leq& 2\Big(\max\Big\{\bar{K}_T^{*,h}\sqrt{3g^{*,h}(T)\Big(2\log\log(\frac{3}{2}g^{*,h}(T)) + \log\frac{4}{\delta}\Big)}, (\bar{K}_T^{*,h})^2\Big\}\Big).\end{aligned} \tag{4.43}$$

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

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

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

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

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 4.16, the difference $\left|R_T^{\pi_1^*,h} - R_T^{\pi_2^*,h}\right|$ satisfies the non-asymptotic concentration rates in Corollary 4.16 with the RHS of (4.40)–(4.41) being simplified to RHS of (4.42)–(4.43). □

4.6 Conclusion

In this chapter, 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.

Appendices to Chapter 4

4.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* [114].
- 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.

4.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 4.14. *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 4.15. *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 Appendices 4.B.1 and 4.B.2, we present the asymptotic and non-asymptotic concentration characteristics of martingales with bounded MDS.

4.B.1 Asymptotic Concentration

4.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 4.9 (see [113, 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:

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

4.B.1.2 Central Limit Theorem

Following theorem characterizes a version of Central Limit Theorem for Martingale Sequences with corresponding bounded MDS.

Theorem 4.10 (see [123, 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

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

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

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

4.B.1.3 Law of Iterated Logarithm

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

Theorem 4.11 (see [124, 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 5.2. 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 Appendix 4.B.2.

4.B.2 Non-Asymptotic Concentration

4.B.2.1 Azuma-Hoeffding Inequality

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

Theorem 4.12 (see [125, 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 4.12, we get following equivalent form of Azuma-Hoeffding inequality.

Corollary 4.18. *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*

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

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 4.12 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.

4.B.2.2 Non-Asymptotic Law of Iterated Logarithm

The following result is a finite-time analogue of the 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 4.13 (see [126, 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*

$$\Big|\sum_{t=1}^{T} Y_t\Big| \leq \sqrt{3\Big(\sum_{t=1}^{T} c_t^2\Big)\Big(2\log\log\frac{3\sum_{t=1}^{T} c_t^2}{2\big|\sum_{t=1}^{T} Y_t\big|} + \log\frac{2}{\delta}\Big)}. \tag{0.44}$$

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

Corollary 4.19. *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*

$$\Big|\sum_{t=1}^{T} Y_t\Big| \leq C\max\Bigg\{\sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{2}{\delta}\Big)}, C\Bigg\}. \tag{0.45}$$

Proof. This corollary follows from Theorem 4.13, by substituting the sequence $\{c_t\}_{t\geq 1}$ with the constant C on the RHS of (0.44). 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 4.13, with probability at least $1-\delta$, we

get:

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

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

$$\Big|\sum_{t=1}^{T} Y_t\Big| \le \max\left\{C\sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{2}{\delta}\Big)}, C^2\right\}. \tag{0.46}$$

□

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

4.C.1 Preliminary Results

4.C.1.1 Martingale Decomposition

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

Definition 4.16. *Let filtration $\mathcal{F} = \{\mathcal{F}_t\}_{t\ge 0}$ 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 $\{M_t^\pi\}_{t\ge 1}$ as follows*

$$M_t^\pi \coloneqq V^\pi(S_t) - \mathbb{E}\big[V^\pi(S_t) \,\big|\, S_{t-1}, \pi(S_{t-1})\big], \quad \forall t \ge 1, \tag{0.47}$$

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

Lemma 4.2. *Sequence $\{M_t^\pi\}_{t\ge 1}$ is an MDS.*

Proof. By the definition of $\{\mathcal{F}_t\}_{t\ge 0}$, 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 $\{M_t^\pi\}_{t\ge 0}$ is an MDS with respect to the filtration $\{\mathcal{F}_t\}_{t\ge 0}$. □

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

Lemma 4.3. *Given any policy $\pi \in \Pi_{\mathrm{AR}}$, we can rewrite the cumulative reward R_T^π as follows*

$$R_T^\pi = TJ^\pi + \sum_{t=1}^{T} M_t^\pi + V^\pi(S_0) - V^\pi(S_T). \tag{0.48}$$

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 (0.47).

□

4.C.1.2 A Consequence of The Union Bound

Lemma 4.4. *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

$$\mathbb{P}\Big(|X_T| > h_1(T, \delta/2)\Big) < \frac{\delta}{2}. \tag{0.49}$$

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

$$\mathbb{P}\Big(|Y_T| > h_2(T, \delta/2)\Big) < \frac{\delta}{2}. \tag{0.50}$$

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 (0.49)–(0.50), we get

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

□

4.C.1.3 Proof of Lemma 4.1

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

$$\sigma_\pi(s) \overset{(a)}{\leq} K^\pi \overset{(b)}{\leq} H^\pi \overset{(c)}{\leq} \infty, \quad \forall s \in \mathcal{S}. \tag{0.51}$$

Proof of Part 1-(a): By the definition of K^π in Eq. (4.5), 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

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

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

$$V^\pi(s) - \mathbb{E}[V^\pi(S_+)|s,\pi(s)] \le V^\pi(s) - \min_{s\in\mathcal{S}} V^\pi(s) \le \max_{s\in\mathcal{S}} V^\pi(s) - \min_{s\in\mathcal{S}} V^\pi(s), \quad \forall s \in \mathcal{S}. \tag{0.52}$$

Similarly, we have

$$\mathbb{E}[V^\pi(S_+)|s,\pi(s)] - V^\pi(s) \le \max_{s\in\mathcal{S}} V^\pi(s) - V^\pi(s) \le \max_{s\in\mathcal{S}} V^\pi(s) - \min_{s\in\mathcal{S}} V^\pi(s), \quad \forall s \in \mathcal{S}. \tag{0.53}$$

Therefore (0.52)–(0.53) imply that

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

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 [121], we have:

$$V^\pi(s) = \mathbb{E}\Big[\sum_{t=0}^{\infty}[r(S_t,A_t) - J^\pi] \;\Big|\; S_0 = s\Big]. \tag{0.54}$$

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

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

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 $H^\pi \leq DR_{\max}$.

Proof of Part 3 The result of this part follows from [8, Theorem 4], where it is shown that for weakly communicating $\mathcal{M}$, we have $H^* \leq DR_{\max}$.

4.C.2 Proof of Theorem 4.1

4.C.2.1 Proof of Part 1

By Lemma 4.3, for any policy $\pi \in \Pi_{\text{AR}}$, 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 (4.5) and Lemma 4.1 we have

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

Therefore

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

As a result by choosing $p = 2$ and $a_t = t$ in Theorem 4.9, we have

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

Furthermore, Lemma 4.1 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

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

4.C.2.2 Proof of Part 2

To prove this part, we verify the conditions of Theorem 5.2 for the MDS $\{M_t^\pi\}_{t\geq 0}$. By Lemma 4.1, 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 4.1, we have

$$\frac{\sum_{t=1}^{\nu_T} M_t^\pi}{\sqrt{T}} \xrightarrow{(d)} \mathcal{N}(0,1). \tag{0.55}$$

Since by Lemma 4.1, $V^\pi(S_t)$ has bounded support for all $t \geq 1$, we get

$$\frac{V^\pi(S_0) - V^\pi(S_T)}{\sqrt{T}} \to 0, \quad a.s. \tag{0.56}$$

By combining (0.55) and (0.56) and by using Theorem 4.14, we get

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

4.C.2.3 Proof of Part 3

We verify the conditions of Theorem 4.11 for the MDS $\{M_t^\pi\}_{t\geq 0}$. By Lemma 4.1, we have

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

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

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

As a result, by using the definition of increasing process $\{\Sigma_t^\pi\}_{t\geq 0}$ and Theorem 4.11, we get

$$\liminf_{T\to\infty} \frac{\sum_{t=1}^{T} M_t^\pi}{\sqrt{2\Sigma_t^\pi \log\log \Sigma_t^\pi}} = -1, \quad \limsup_{T\to\infty} \frac{\sum_{t=1}^{T} M_t^\pi}{\sqrt{2\Sigma_t^\pi \log\log \Sigma_t^\pi}} = 1. \tag{0.57}$$

Since by Lemma 4.1, $V^\pi(S_t)$ has bounded support for all $t \geq 1$, we get

$$\lim_{T\to\infty} \frac{V^\pi(S_0) - V^\pi(S_T)}{\sqrt{2\Sigma_t^\pi \log\log \Sigma_t^\pi}} = 0, \quad \text{a.s.} \tag{0.58}$$

By combining (0.57) and (0.58), we get

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

4.C.3 Proof of Theorem 4.2

4.C.3.1 Proof of Part 1

By Lemma 4.3, 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

$$\big|R_T^\pi(\omega) - TJ^\pi - \big(V^\pi(S_0) - V^\pi(S_T)\big)\big| = \big|\sum_{t=1}^{T} M_t^\pi\big|. \tag{0.59}$$

In order to upper-bound the term $\big|\sum_{t=1}^{T} M_t^\pi\big|$, we verify the conditions of Corollary 4.18. By (4.5) and Lemma 4.1 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 4.18 implies that for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\Big|\sum_{t=1}^{T} M_t^\pi\Big| \leq \sqrt{2T(K^\pi)^2 \log(\frac{2}{\delta})}. \tag{0.60}$$

By combining (0.59) and (0.60), with probability at least $1-\delta$, we have

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

4.C.3.2 Proof of Part 2

Similar to the proof of Part 1, by lemma 4.3, we have

$$\left|R_T^\pi(\omega) - TJ^\pi - \big(V^\pi(S_0) - V^\pi(S_T)\big)\right| = \Big|\sum_{t=1}^{T} M_t^\pi\Big| \tag{0.61}$$

Moreover, MDS $\{M_t^\pi\}_{t\geq 0}$ is a uniformly bounded MDS with respect to the constant K^π. Therefore, Corollary 4.19 implies that for any $\delta \in (0,1)$, for all $T \geq T_0(\delta) \coloneqq \left\lceil \frac{173}{K^\pi} \log \frac{4}{\delta} \right\rceil$, with probability at least $1-\delta$, we have

$$\Big|\sum_{t=1}^{T} M_t^\pi\Big| \leq \max\Big\{K^\pi \sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{2}{\delta}\Big)}, (K^\pi)^2\Big\}. \tag{0.62}$$

By combining (0.61) and (0.62), with probability at least $1-\delta$, we have

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

4.C.4 Proof of Theorem 4.3

4.C.4.1 Proof of Part 1

By lemma 4.3, 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

$$\begin{aligned}
\left|R_T^\pi(\omega) - TJ^\pi\right| &= \Big|\sum_{t=1}^{T} M_t^\pi + V^\pi(S_0) - V^\pi(S_T)\Big| \\
&\overset{(a)}{\leq} \Big|\sum_{t=1}^{T} M_t^\pi\Big| + \Big|V^\pi(S_0) - V^\pi(S_T)\Big| \\
&\overset{(b)}{\leq} \Big|\sum_{t=1}^{T} M_t^\pi\Big| + H^\pi,
\end{aligned} \tag{0.63}$$

where (a) follows from the triangle inequality and (b) follows from the definition of H^π. In order to upper-bound the term $\big|\sum_{t=1}^{T} M_t^\pi\big|$, we verify the conditions of Corollary 4.18. By

(4.5) and Lemma 4.1 we have

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

As a result, MDS $\{M_t^\pi\}_{t\geq 1}$ is a uniformly bounded MDS with respect to the constant K^π. Therefore, Corollary 4.18 implies that for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\left|\sum_{t=1}^{T} M_t^\pi\right| \leq \sqrt{2T(K^\pi)^2 \log(\frac{2}{\delta})}. \tag{0.64}$$

By combining (0.63) and (0.64), with probability at least $1-\delta$, we have

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

4.C.4.2 Proof of Part 2

Similar to the proof of Part 1, by lemma 4.3, we have

$$\left|R_T^\pi(\omega) - TJ^\pi\right| \leq \left|\sum_{t=1}^{T} M_t^\pi\right| + H^\pi. \tag{0.65}$$

Moreover, MDS $\{M_t^\pi\}_{t\geq 0}$ is a uniformly bounded MDS with respect to the constant K^π. Therefore, Corollary 4.19 implies that for any $\delta \in (0,1)$, for all $T \geq T_0(\delta) \coloneqq \left\lceil \frac{173}{K^\pi} \log \frac{4}{\delta} \right\rceil$, with probability at least $1-\delta$, we have

$$\left|\sum_{t=1}^{T} M_t^\pi\right| \leq \max\left\{K^\pi \sqrt{3T\Big(2\log\log\frac{3T}{2} + \log\frac{2}{\delta}\Big)}, (K^\pi)^2\right\}. \tag{0.66}$$

By combining (0.65) and (0.66), with probability at least $1-\delta$, we have

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

4.C.5 Proof of Corollary 4.5

4.C.5.1 Proof of Part 1

Since $\mathcal{M}$ is communicating, by Lemma 4.1, for any policy $\pi \in \Pi_{\mathrm{AR}}$, we have

$$\left|M_t^\pi\right| \leq K^\pi \leq DR_{\max}, \quad \forall t \geq 1. \tag{0.67}$$

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 4.3, Part 1, and substituting H^π with $DR_{\max}$ in the RHS of (0.63) and replacing K^π with $DR_{\max}$ in the RHS of (0.64), we get that with probability at least $1-\delta$, we have:

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

4.C.5.2 Proof of Part 2

Since $\mathcal{M}$ is communicating, by Lemma 4.1, for any policy $\pi \in \Pi_{\text{AR}}$, we have

$$\left|M_t^\pi\right| \leq K^\pi \leq DR_{\max}, \quad \forall t \geq 1. \tag{0.68}$$

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 4.3, Part 2, and substituting H^π with $DR_{\max}$ in the RHS of (0.65) and substituting K^π with $DR_{\max}$ in the RHS of (0.66), we prove the claim, i.e, 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

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

4.C.6 Proof of Corollary 4.6

In the case of communicating $\mathcal{M}$, since $\pi^* \in \Pi_{\text{AR}}$, by Corollary 4.5, we get that $R_T^{\pi^*}(\omega)$ satisfies the non-asymptotic concentration rates in (4.16)–(4.17).

In the case of weakly communicating $\mathcal{M}$, by Lemma 4.1, for any optimal policy $\pi^* \in \Pi_{\text{AR}}$, we have

$$\left|M_t^{\pi^*}\right| = \left|V^*(S_t) - \mathbb{E}\left[V^*(S_t) \mid S_{t-1}, \pi(S_{t-1})\right]\right| \leq K^* \leq DR_{\max}, \quad \forall t \geq 1. \tag{0.69}$$

As a result, the MDS $\{M_t^{\pi^*}\}_{t\geq 1}$ is uniformly bounded MDS with respect to the constant $DR_{\max}$. Therefore, by repeating the arguments of the proof of Corollary 4.5, Part 1 and Part 2 for the optimal policy $\pi^* \in \Pi_{\text{AR}}$, we prove that $\left|R_T^{\pi^*}(\omega) - TJ^*\right|$ satisfies the non-asymptotic concentration rates in (4.16)–(4.17).

4.C.7 Proof of Corollary 4.7

4.C.7.1 Proof of Part 1

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

$$\begin{aligned}\left|R_T^{\pi_1} - R_T^{\pi_2}\right| =& \left|R_T^{\pi_1} - TJ^{\pi_1} + TJ^{\pi_1} - TJ^{\pi_2} + TJ^{\pi_2} - R_T^{\pi_2}\right| \\ \overset{(a)}{\leq}& \left|R_T^{\pi_1} - TJ^{\pi_1}\right| + \left|TJ^{\pi_1} - TJ^{\pi_2}\right| + \left|TJ^{\pi_2} - R_T^{\pi_2}\right|, \end{aligned} \tag{0.70}$$

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

$$\begin{aligned}\left|TJ^{\pi_1} - TJ^{\pi_2}\right| =& \left|TJ^{\pi_1} - R_T^{\pi_1} + R_T^{\pi_1} - R_T^{\pi_2} + R_T^{\pi_2} - TJ^{\pi_2}\right| \\ \overset{(b)}{\leq}& \left|TJ^{\pi_1} - R_T^{\pi_1}\right| + \left|R_T^{\pi_1} - R_T^{\pi_2}\right| + \left|R_T^{\pi_2} - TJ^{\pi_2}\right|, \end{aligned} \tag{0.71}$$

where (b) follows from the triangle inequality. (0.70)–(0.71) imply that

$$\left|\left|R_T^{\pi_1} - R_T^{\pi_2}\right| - \left|TJ^{\pi_1} - TJ^{\pi_2}\right|\right| \leq \left|R_T^{\pi_1} - TJ^{\pi_1}\right| + \left|R_T^{\pi_2} - TJ^{\pi_2}\right|. \tag{0.72}$$

By Theorem 4.3, 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 4.4 and (0.72), we get that for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

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

4.C.7.2 Proof of Part 2

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

$$\Big|\big|R_T^{\pi_1} - R_T^{\pi_2}\big| - \big|TJ^{\pi_1} - TJ^{\pi_2}\big|\Big| \leq \big|R_T^{\pi_1} - TJ^{\pi_1}\big| + \big|R_T^{\pi_2} - TJ^{\pi_2}\big|.$$

By Theorem 4.3, 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

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

Similarly, 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

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

As a result, by applying Lemma 4.4 and (0.72), we get that for all $T \geq T_0(\delta) \coloneqq \max\Big\{\Big\lceil \frac{173}{K^{\pi_1}} \log \frac{8}{\delta} \Big\rceil, \Big\lceil \frac{173}{K^{\pi_2}} \log \frac{8}{\delta} \Big\rceil\Big\}$, with probability at least $1 - \delta$, we have

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

4.C.8 Proof of Theorem 4.4

By Corollary 4.1, for any optimal policy $\pi^* \in \Pi_{\text{AR}}$, the quantity R^{π^*} satisfies the asymptotic concentration rates in (4.9)–(4.11). On the other hand, by (4.3), 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 (4.9)–(4.11), we get that for any learning policy μ, these asymptotic concentration rates also hold for the difference $\mathcal{D}_T(\omega)$ of cumula-

tive regret and interim cumulative regret.

4.C.9 Proof of Theorem 4.5

By Corollary 4.2, for any optimal policy $\pi^* \in \Pi_{\text{AR}}$, the quantity $|R_T^{\pi^*} - TJ^*|$ satisfies the asymptotic concentration rates in (4.14)–(4.15). On the other hand, by (4.3), 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 (4.14)–(4.15), 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.

4.C.10 Proof of Corollary 4.9

By Corollary 4.6, for the weakly communicating $\mathcal{M}$, for any optimal policy $\pi^* \in \Pi_{\text{AR}}$, the quantity $|R_T^{\pi^*} - TJ^*|$ satisfies the non-asymptotic concentration rates in (4.16)–(4.17). On the other hand, by (4.3), 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 (4.16)–(4.17), 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. 4.3, we have that if $\mathcal{M}$ is recurrent, unichain, or communicating it is also weakly communicating. As a result, these non-asymptotic concentration bounds hold for all the cases.

4.C.11 Proof of Theorem 4.6

4.C.11.1 Proof of Part 1

This part of the theorem is a consequence of Theorem 4.4. Recall that by definition, we have

$$\mathcal{D}_T(\omega) = \mathcal{R}_T^{\mu}(\omega) - \bar{\mathcal{R}}_T^{\mu}(\omega). \tag{0.73}$$

On the other hand, we can rewrite the law of iterated logarithm in Theorem 4.4 using the $\tilde{\mathcal{O}}(\cdot)$ notation as follows

$$\mathcal{D}_T(\omega) \leq \tilde{\mathcal{O}}(\sqrt{T}), \quad a.s. \tag{0.74}$$

As a result, for any learning policy μ that satisfies $R_T^\mu(\omega) \leq \tilde{\mathcal{O}}(\sqrt{T})$, almost surely, (0.73)–(0.74) 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, (0.73)–(0.74) imply that $R_T^\mu(\omega) \leq \tilde{\mathcal{O}}(\sqrt{T})$. Therefore, statements 1 and 2 are equivalent.

4.C.11.2 Proof of Part 2

Proof of this part is a consequence of Theorem 4.5. 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

$$R_T^\mu(\omega) \leq h_1(\delta_1, T), \tag{0.75}$$

where for a fixed δ_1, we have $h_1(\delta_1, T) = \tilde{\mathcal{O}}(\sqrt{T})$. Moreover, by Theorem 4.5, 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

$$D_T(\omega) \leq h_2(\delta_2, T), \tag{0.76}$$

where for a fixed δ_2, we have $h_2(\delta_2, T) = \tilde{\mathcal{O}}(\sqrt{T})$. As a result, by (0.73), (0.75)–(0.76), and Lemma 4.4, we get that for any $\delta \in (0,1)$, for all $T \geq \max\big\{T_1(\delta/2), T_2(\delta/2)\big\}$, with probability at least $1 - \delta$, we have

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

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

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

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

4.D Proof of Main Results for Discounted Reward Setup

4.D.1 Proof of Theorem 4.7

4.D.1.1 Preliminary Results

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

Definition 4.17. *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 $\pi \in \Pi_{\text{SD}}$, let V_γ^π denote the corresponding discounted value function. We define the sequence $\{N_t^{\pi,\gamma}\}_{t\geq 1}$ as follows*

$$N_t^{\pi,\gamma} \coloneqq \Big[V_\gamma^\pi(S_t) - \mathbb{E}\big[V_\gamma^\pi(S_t) \;\big|\; S_{t-1}, \pi(S_{t-1})\big]\Big], \quad \forall t \geq 1, \tag{0.77}$$

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

Lemma 4.5. *Sequence $\{\gamma^t N_t^{\pi,\gamma}\}_{t\geq 1}$ is an MDS.*

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

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

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

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

Lemma 4.6. *Given any policy $\pi \in \Pi_{\text{SD}}$, we can rewrite the discounted cumulative return $R_T^{\pi,\gamma}$ as follows*

$$R_T^{\pi,\gamma}(\omega) = \sum_{t=1}^{T} \gamma^t N_t^{\pi,\gamma} + V_\gamma^\pi(S_0) - \gamma^T V_\gamma^\pi(S_T). \tag{0.78}$$

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

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

Repeating similar steps as in the proof of Lemma 4.3, 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 $\{N_t^{\pi,\gamma}\}_{t\geq 0}$. $\square$

4.D.1.2 Proof of Theorem 4.7

Proof of this theorem follows from the martingale decomposition stated in Lemma 4.6 and the concentration bounds stated in Corollary 4.18 and Theorem 4.13.

Proof of Part 1 By Lemma 4.6, 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

$$\Big| R_T^{\pi,\gamma}(\omega) - \big(V_\gamma^\pi(S_0) - \gamma^T V_\gamma^\pi(S_T)\big) \Big| = \Big| \sum_{t=1}^{T} \gamma^t N_t^{\pi,\gamma} \Big|. \tag{0.79}$$

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 4.18. By (4.24) and Lemma 4.1, we have

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

As a result, MDS $\{\gamma^t N_t^{\pi,\gamma}\}_{t\geq 1}$ is a sequentially bounded MDS with respect to the sequence $\{\gamma^t K^{\pi,\gamma}\}_{t\geq 1}$. Therefore, Corollary 4.18 implies that for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\begin{aligned}\Big|\sum_{t=1}^{T}\gamma^t N_t^{\pi,\gamma}\Big| &\leq \sqrt{2\sum_{t=1}^{T}(K^{\pi,\gamma})^2\gamma^{2t}\log\frac{2}{\delta}} \\ &= K^{\pi,\gamma}\sqrt{2\sum_{t=1}^{T}\gamma^{2t}\log\frac{2}{\delta}} \\ &= K^{\pi,\gamma}\sqrt{2f^{\gamma}(T)\log\frac{2}{\delta}}. \end{aligned} \tag{0.80}$$

As a result, by combining (0.79) and (0.80), we get that with probability at least $1-\delta$, we have

$$\Big|R_T^{\pi,\gamma}(\omega) - \big(V_\gamma^\pi(S_0) - \gamma^T V_\gamma^\pi(S_T)\big)\Big| \leq K^{\pi,\gamma}\sqrt{2f^{\gamma}(T)\log\frac{2}{\delta}}. \tag{0.81}$$

Proof of Part 2: Similar to the proof of Part 1, by Lemma 4.6, we have

$$\Big|R_T^{\pi,\gamma}(\omega) - \big(V_\gamma^\pi(S_0) - \gamma^T V_\gamma^\pi(S_T)\big)\Big| = \Big|\sum_{t=1}^{T}\gamma^t N_t^{\pi,\gamma}\Big|. \tag{0.82}$$

Moreover, MDS $\{\gamma^t N_t^{\pi,\gamma}\}_{t\geq 1}$ is a sequentially bounded MDS with respect to the sequence $\{\gamma^t K^{\pi,\gamma}\}_{t\geq 1}$. Therefore, Theorem 4.13 implies that for any $\delta \in (0,1)$, for all $T \geq T_0(\delta) \coloneqq \min\big\{T \geq 1 : f^{\gamma}(T) > \frac{173}{K^{\pi,\gamma}}\log\frac{4}{\delta}\big\}$, with probability at least $1-\delta$, we have

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

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

$|\sum_{t=1}^{T}\gamma^t N_t^{\pi,\gamma}| \geq (K^{\pi,\gamma})^2$, we get:

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

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

$$\Big|\sum_{t=1}^{T}\gamma^t N_t^{\pi,\gamma}\Big| \leq \max\Big\{K^{\pi,\gamma}\sqrt{3f^{\gamma}(T)\big(2\log\log\frac{3}{2}f^{\gamma}(T)+\log\frac{2}{\delta}\big)},(K^{\pi,\gamma})^2\Big\}. \tag{0.83}$$

By combining (0.82)–(0.83), with probability at least $1-\delta$, we have

$$\begin{aligned}&\Big|R_T^{\pi,\gamma}(\omega)-\big(V_{\gamma}^{\pi}(S_0)-\gamma^T V_{\gamma}^{\pi}(S_T)\big)\Big|\\ &\qquad\leq \max\Big\{K^{\pi,\gamma}\sqrt{3f^{\gamma}(T)\big(2\log\log(\frac{3}{2}f^{\gamma}(T))+\log\frac{2}{\delta}\big)},(K^{\pi,\gamma})^2\Big\}.\end{aligned} \tag{0.84}$$

4.D.2 Proof of Corollary 4.10

Proof of Part 1: By Lemma 4.6, 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

$$\Big|R_T^{\pi,\gamma}(\omega)-V_{\gamma}^{\pi}(S_0)\Big| \overset{(a)}{\leq} \Big|\sum_{t=1}^{T}\gamma^t N_t^{\pi,\gamma}\Big|+\Big|\gamma^T V_{\gamma}^{\pi}(S_T)\Big|, \tag{0.85}$$

where (a) follows from the triangle inequality. In the proof of Theorem 4.7, Part 1, we showed that with probability at least $1-\delta$, we have

$$\Big|\sum_{t=1}^{T}\gamma^t N_t^{\pi,\gamma}\Big| \leq K^{\pi,\gamma}\sqrt{2f^{\gamma}(T)\log\frac{2}{\delta}}. \tag{0.86}$$

Moreover, we have

$$\gamma^T V_\gamma^\pi(S_T) = \gamma^T \mathbb{E}^\pi\Big[\lim_{T\to\infty}\sum_{t=0}^{T-1}\gamma^t r(S_t,A_t) \;\Big|\; S_0 = S_T\Big]$$
$$= \gamma^T \mathbb{E}^\pi\Big[\lim_{T\to\infty}\sum_{t=0}^{T-1}\gamma^t R_{\max} \;\Big|\; S_0 = S_T\Big] \le \frac{\gamma^T}{1-\gamma}R_{\max}. \tag{0.87}$$

By combining (0.85)–(0.87), 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 4.6, we have

$$\Big|R_T^{\pi,\gamma}(\omega) - V_\gamma^\pi(S_0)\Big| \le \Big|\sum_{t=1}^{T}\gamma^t N_t^{\pi,\gamma}\Big| + \Big|\gamma^T V_\gamma^\pi(S_T)\Big|. \tag{0.88}$$

Moreover, we have

$$\Big|\gamma^T V_\gamma^\pi(S_T)\Big| \le \gamma^T\frac{R_{\max}}{1-\gamma}. \tag{0.89}$$

In addition, from proof of Theorem 4.7, Part 2, for any $\delta\in(0,1)$, for all $T \ge T_0(\delta) := \min\big\{T' \ge 1 : f^\gamma(T') > \frac{173}{K^{\pi,\gamma}}\log\frac{4}{\delta}\big\}$, with probability at least $1-\delta$, we have

$$\Big|\sum_{t=1}^{T}\gamma^t N_t^{\pi,\gamma}\Big| \le \max\Big\{K^{\pi,\gamma}\sqrt{3f^\gamma(T)\big(2\log\log\frac{3}{2}f^\gamma(T) + \log\frac{2}{\delta}\big)}, (K^{\pi,\gamma})^2\Big\}. \tag{0.90}$$

By combining (0.88)–(0.90), with probability at least $1-\delta$, we have

$$\Big|R_T^{\pi,\gamma}(\omega) - V_\gamma^\pi(S_0)\Big|$$
$$\le \max\Big\{K^{\pi,\gamma}\sqrt{3f^\gamma(T)\big(2\log\log(\frac{3}{2}f^\gamma(T)) + \log\frac{2}{\delta}\big)}, (K^{\pi,\gamma})^2\Big\} + \frac{\gamma^T}{1-\gamma}R_{\max}. \tag{0.91}$$

4.D.3 Proof of Corollary 4.12

4.D.3.1 Proof of Part 1

Consider two policies $\pi_1, \pi_2 \in \Pi_{\text{SD}}$. Let $\{S_t^{\pi_1}\}_{t\geq 0}$ and $\{S_t^{\pi_2}\}_{t\geq 0}$ denote the random sequences of states encountered by following policies π_1 and π_2. We have

$$\begin{aligned}
\left|R_T^{\pi_1,\gamma} - R_T^{\pi_2,\gamma}\right| &\overset{(a)}{=} \Big|R_T^{\pi_1,\gamma} - \big[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\big] + \big[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\big] \\
&\quad - \big[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\big] + \big[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\big] - R_T^{\pi_2,\gamma}\Big| \\
&\overset{(b)}{\leq} \Big|R_T^{\pi_1,\gamma} - \big[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\big]\Big| + \Big|\big[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\big] - R_T^{\pi_2,\gamma}\Big| \\
&\quad + \Big|\big[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\big] - \big[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\big]\Big|, \qquad (0.92)
\end{aligned}$$

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

$$\begin{aligned}
&\Big|\big[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\big] - \big[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\big]\Big| \overset{(a)}{=} \\
&\Big|\big[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\big] - R_T^{\pi_1,\gamma} + R_T^{\pi_1,\gamma} - R_T^{\pi_2,\gamma} + R_T^{\pi_2,\gamma} - \big[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\big]\Big| \\
&\overset{(b)}{\leq} \Big|R_T^{\pi_1,\gamma} - \big[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\big]\Big| + \Big|R_T^{\pi_2,\gamma} - \big[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\big]\Big| \\
&+ \Big|R_T^{\pi_1,\gamma} - R_T^{\pi_2,\gamma}\Big|, \qquad (0.93)
\end{aligned}$$

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. (0.92)–(0.93) imply that

$$\begin{aligned}
&\Big|\big|R_T^{\pi_1,\gamma} - R_T^{\pi_2,\gamma}\big| - \big|\big[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\big] - \big[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\big]\big|\Big| \\
\leq &\Big|R_T^{\pi_1,\gamma} - \big[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\big]\Big| + \Big|R_T^{\pi_2,\gamma} - \big[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\big]\Big|. \qquad (0.94)
\end{aligned}$$

By Theorem 4.7, we know that for any $\delta_1 \in (0,1)$, with probability at least $1-\delta_1$, we have

$$\Big|R_T^{\pi_1,\gamma} - \big[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\big]\Big| \leq K^{\pi_1,\gamma}\sqrt{2f^\gamma(T)\log\frac{2}{\delta_1}}. \qquad (0.95)$$

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

$$\Big|R_T^{\pi_2,\gamma} - \big[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\big]\Big| \leq K^{\pi_2,\gamma}\sqrt{2f^\gamma(T)\log\frac{2}{\delta_2}}. \qquad (0.96)$$

As a result, by applying Lemma 4.4 and (0.94)–(0.96), we get that for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

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

4.D.3.2 Proof of Part 2

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

$$\begin{aligned}
&\Big|\big|R_T^{\pi_1,\gamma} - R_T^{\pi_2,\gamma}\big| - \big|\big[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\big] - \big[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\big]\big|\Big| \\
\le&\Big|R_T^{\pi_1,\gamma} - \big[V_\gamma^{\pi_1}(S_0^{\pi_1}) - \gamma^T V_\gamma^{\pi_1}(S_T^{\pi_1})\big]\Big| + \Big|R_T^{\pi_2,\gamma} - \big[V_\gamma^{\pi_2}(S_0^{\pi_2}) - \gamma^T V_\gamma^{\pi_2}(S_T^{\pi_2})\big]\Big|. \qquad (0.97)
\end{aligned}$$

By Theorem 4.7, for any $\delta_1 \in (0,1)$, for all $T \ge T_0^{\pi_1}(\delta_1) \coloneqq \min\big\{T' \ge 1 : f^\gamma(T') > \frac{173}{K^{\pi_1,\gamma}}\log\frac{4}{\delta_1}\big\}$, with probability at least $1-\delta_1$, we have:

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

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

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

As a result, by applying Lemma 4.4, we get that for all $T \geq T_0^{\pi}(\delta) \coloneqq \max\left\{T_0^{\pi_1}(\frac{\delta}{2}), T_0^{\pi_2}(\frac{\delta}{2})\right\}$, with probability at least $1-\delta$, we have

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

4.D.4 Proof of Corollary 4.14

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 the following preliminary lemma.

4.D.4.1 Preliminary Lemma

Lemma 4.7. *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*

$$\lim_{\gamma\uparrow 1} f(T,\gamma) = T. \tag{0.98}$$

4. *We have*

$$\lim_{\gamma\uparrow 1} R_T^{\pi,\gamma} = R_T^\pi. \tag{0.99}$$

Proof. **of Part 1**: From the Laurent series expansion ([119, Proposition 5.1.2], for any policy $\pi \in \Pi_{\mathrm{SD}}$, 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 ([119, Proposition 5.1.2], for any policy $\pi\in\Pi_{\mathrm{SD}}$, 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

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

□

4.D.4.2 Proof of Corollary 4.14

Proof of Part 1: By Lemma 4.7, Part 4, for all $T \geq 1$, we have

$$\lim_{\gamma \uparrow 1} \left[R_T^{\pi,\gamma}\right] = R_T^{\pi}. \tag{0.100}$$

Moreover, we have

$$\begin{aligned}\lim_{\gamma \uparrow 1} \left[V_\gamma^\pi(S_0) - \gamma^T V_\gamma^\pi(S_T)\right] &= \lim_{\gamma \uparrow 1} \left[V_\gamma^\pi(S_0) - V_\gamma^\pi(S_T) + V_\gamma^\pi(S_T) - \gamma^T V_\gamma^\pi(S_T)\right] \\ &\overset{(a)}{=} V^\pi(S_0) - V^\pi(S_T) + TJ^\pi + V^\pi(S_T) - V^\pi(S_T) \\ &= TJ^\pi + V^\pi(S_0) - V^\pi(S_T), \end{aligned} \tag{0.101}$$

where (a) follows from Lemma 4.7, Parts 1 and 2. The result of this part follows by substituting (0.100)–(0.101) on the LHS of (4.25).

Proof of Part 2: By Lemma 4.7, Part 2, for all $s_1, s_2 \in \mathcal{S}$, we have

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

This implies that

$$\lim_{\gamma \uparrow 1} \left[K^{\pi,\gamma}\right] = K^\pi. \tag{0.102}$$

Moreover, by Lemma 4.7, Part 3, we have

$$\lim_{\gamma \uparrow 1} f^\gamma(T) = T. \tag{0.103}$$

The result of this part follows by substituting (0.102)–(0.103) on the RHS of (4.25).

Proof of Part 3: The result of this part follows by substituting (0.102)–(0.103) on the RHS of (4.26).

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

4.E.1 Proof of Theorem 4.8

4.E.1.1 Preliminary Results

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

Definition 4.18. *Let filtration $\mathcal{F} = \{\mathcal{F}_t\}_{t=0}^h$ be defined as $\mathcal{F}_t \coloneqq \sigma(S_{0:t}, A_{0:t})$. 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*

$$W_t^{\pi,h} \coloneqq \Big[V_t^{\pi,h}(S_t) - \mathbb{E}\big[V_t^{\pi,h}(S_t) \mid S_{t-1}, \pi_{t-1}(S_{t-1})\big]\Big], \quad \forall t \in \{1, \dots, h+1\}, \tag{0.104}$$

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

Lemma 4.8. *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

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

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

$$R_T^{\pi,h}(\omega) = \sum_{t=1}^{T} W_t^{\pi,h} + V_0^{\pi,h}(S_0) - V_T^{\pi,h}(S_T). \tag{0.105}$$

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

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

□

For any $1 \leq T \leq h+1$, by repeating similar steps as in the proof of Lemma 4.3, 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 (0.104).

4.E.1.2 Proof of Theorem 4.8

Proof of this theorem follows from the martingale decomposition stated in Lemma 4.9 and the concentration bounds stated in Theorem 4.12 and Theorem 4.13.

Proof of Part 1 By Lemma 4.9, 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

$$\Big| R_T^{\pi,h}(\omega) - \big(V_0^{\pi,h}(S_0) - V_T^{\pi,h}(S_T)\big) \Big| = \Big| \sum_{t=1}^{T} W_t^{\pi,h} \Big|. \tag{0.106}$$

In order to upper-bound the term $\big|\sum_{t=1}^{T} W_t^{\pi,h}\big|$, we verify the conditions of Corollary 4.18. By (4.35), we have

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

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 4.18 implies that for any $\delta \in (0,1)$, with probability at least

$1-\delta$, we have

$$\Big|\sum_{t=1}^{T} W_t^{\pi,h}\Big| \le \sqrt{2\sum_{t=1}^{T}(K_t^{\pi,h})^2 \log\frac{2}{\delta}}$$
$$\overset{(a)}{=} \bar{K}_T^{\pi,h}\sqrt{2g^{\pi,h}(T)\log\frac{2}{\delta}}, \tag{0.107}$$

where (a) follows from (4.37). By combining (0.106) and (0.107), we get that with probability at least $1-\delta$, we have

$$\Big|R_T^{\pi,h} - \big(V_0^{\pi,h}(S_0) - V_T^{\pi,h}(S_T)\big)\Big| \le \sqrt{2g^{\pi,h}(T)\log\frac{2}{\delta}}. \tag{0.108}$$

Proof of Part 2: Similar to the proof of Part 1, by Lemma 4.9, we have

$$\Big|R_T^{\pi,h} - \big(V_0^{\pi,h}(S_0) - V_T^{\pi,h}(S_T)\big)\Big| = \Big|\sum_{t=1}^{T} W_t^{\pi,h}\Big|. \tag{0.109}$$

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 4.13 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 $|\sum_{t=1}^{T} W_t^{\pi,h}| \leq (\bar{K}_T^{\pi,h})^2$ or $|\sum_{t=1}^{T} W_t^{\pi,h}| \geq (\bar{K}_T^{\pi,\gamma})^2$. If $|\sum_{t=1}^{T} W_t^{\pi,h}| \geq (\bar{K}_T^{\pi,\gamma})^2$, we get:

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

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

$$\Big|\sum_{t=1}^{T} W_t^{\pi,h}\Big| \leq \max\Big\{\bar{K}_T^{\pi,h}\sqrt{3g^{\pi,h}(T)\big(2\log\log\big(\frac{3}{2}g^{\pi,h}(T)\big)+\log\frac{2}{\delta}\big)}, (\bar{K}_T^{\pi,h})^2\Big\}. \tag{0.110}$$

By combining (0.109)–(0.110), with probability at least $1-\delta$, we have

$$\begin{aligned}&\Big|R_T^{\pi,h}(\omega) - \big(V_0^{\pi,h}(S_0) - V_T^{\pi,h}(S_T)\big)\Big| \\ &\qquad \leq \max\Big\{\bar{K}_T^{\pi,h}\sqrt{3g^{\pi,h}(T)\big(2\log\log(\frac{3}{2}g^{\pi,h}(T))+\log\frac{2}{\delta}\big)}, (\bar{K}_T^{\pi,h})^2\Big\}.\end{aligned} \tag{0.111}$$

4.E.2 Proof of Corollary 4.15

Proof of Part 1 By Lemma 4.9, 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

$$\Big|R_T^{\pi,h}(\omega) - V_0^{\pi,h}(S_0)\Big| \overset{(a)}{\leq} \Big|\sum_{t=1}^{T} W_t^{\pi,h}\Big| + \Big|V_T^{\pi,h}(S_T)\Big|, \tag{0.112}$$

where (a) follows from the triangle inequality. In the proof of Theorem 4.8, Part 1, we showed that with probability at least $1-\delta$, we have

$$\Big|\sum_{t=1}^{T} W_t^{\pi,h}\Big| \le \sqrt{2\sum_{t=1}^{T}(K_t^{\pi,h})^2 \log\frac{2}{\delta}} \overset{(b)}{\le} \bar{K}_T^{\pi,h}\sqrt{2T\log\frac{2}{\delta}}, \tag{0.113}$$

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

$$V_T^{\pi,h}(S_T) \le \bar{H}_T^{\pi,h}, \quad \forall t \le T. \tag{0.114}$$

By combining (0.112)–(0.114), 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 4.9, we have

$$\Big|R_T^{\pi,h}(\omega) - V_0^{\pi,h}(S_0)\Big| \le \Big|\sum_{t=1}^{T} W_t^{\pi,h}\Big| + \Big|V_T^{\pi,h}(S_T)\Big|, \tag{0.115}$$

Moreover, we have

$$V_T^{\pi,h}(S_T) \le \bar{H}_T^{\pi,h}. \tag{0.116}$$

In addition, from proof of Theorem 4.8, 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

$$\Big|\sum_{t=1}^{T} W_t^{\pi,h}\Big| \le \max\left\{\bar{K}_T^{\pi,h}\sqrt{3g^{\pi,h}(T)\big(2\log\log\big(\frac{3}{2}g^{\pi,h}(T)\big)+\log\frac{2}{\delta}\big)}, (\bar{K}_T^{\pi,h})^2\right\} \overset{(c)}{\le} \max\left\{\bar{K}_T^{\pi,h}\sqrt{3T\big(2\log\log\big(\frac{3T}{2}\big)+\log\frac{2}{\delta}\big)}, (\bar{K}_T^{\pi,h})^2\right\}, \tag{0.117}$$

where (c) follows from the fact that $g^{\pi,h}(T) \le T$. By combining (0.115)–(0.117), with probability at least $1-\delta$, we have

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

4.E.3 Proof of Corollary 4.16

4.E.3.1 Proof of Part 1

Consider two policies $\pi_1, \pi_2 \in \Pi_{\text{SD}}$. Let $\{S_t^{\pi_1}\}_{t\geq 0}$ and $\{S_t^{\pi_2}\}_{t\geq 0}$ denote the random sequence of states encountered by following policies π_1 and π_2. We have

$$\begin{aligned}\left|R_T^{\pi_1,h} - R_T^{\pi_2,h}\right| &\overset{(a)}{=} \Big|R_T^{\pi_1,h} - \big[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\big] + \big[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\big] \\ &\quad - \big[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\big] + \big[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\big] - R_T^{\pi_2,h}\Big| \\ &\overset{(b)}{\leq} \Big|R_T^{\pi_1,h} - \big[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\big]\Big| + \Big|\big[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\big] - R_T^{\pi_2,h}\Big| \\ &\quad + \Big|\big[V_0^{\pi_1}(S_0^{\pi_1}) - V_T^{\pi_1}(S_T^{\pi_1})\big] - \big[V_0^{\pi_2}(S_0^{\pi_2}) - V_T^{\pi_2}(S_T^{\pi_2})\big]\Big|, \end{aligned} \tag{0.118}$$

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

$$\begin{aligned}&\Big|\big[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\big] - \big[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\big]\Big| \overset{(a)}{=} \\ &\Big|\big[V_0^{\pi_1}(S_0^{\pi_1}) - V_T^{\pi_1}(S_T^{\pi_1})\big] - R_T^{\pi_1,h} + R_T^{\pi_1,h} - R_T^{\pi_2,h} + R_T^{\pi_2,T} - \big[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\big]\Big| \\ \overset{(b)}{\leq}&\Big|R_T^{\pi_1,h} - \big[V_0^{\pi_1}(S_0^{\pi_1}) - V_T^{\pi_1}(S_T^{\pi_1})\big]\Big| + \Big|R_T^{\pi_2,h} - \big[V_0^{\pi_2}(S_0^{\pi_2}) - V_T^{\pi_2}(S_T^{\pi_2})\big]\Big| \\ &+\Big|R_T^{\pi_1,h} - R_T^{\pi_2,h}\Big|, \end{aligned} \tag{0.119}$$

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. (0.118)–(0.119) imply that

$$\begin{aligned}&\Big|\big|R_T^{\pi_1,h} - R_T^{\pi_2,h}\big| - \big|\big[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\big] - \big[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\big]\big|\Big| \\ \leq&\Big|R_T^{\pi_1,h} - \big[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\big]\Big| + \Big|R_T^{\pi_2,h} - \big[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_\gamma^{\pi_2,h}(S_T^{\pi_2})\big]\Big|. \end{aligned} \tag{0.120}$$

By Theorem 4.8, we know that for any $\delta_1 \in (0,1)$, with probability at least $1-\delta_1$, we have

$$\Big|R_T^{\pi_1,h} - \big(V_0^{\pi_1,h}(S_0) - V_T^{\pi_1,h}(S_T)\big)\Big| \leq \bar{K}_T^{\pi_1,h}\sqrt{2g^{\pi_1,h}(T)\log\frac{2}{\delta_1}}. \tag{0.121}$$

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

$$\Big|R_T^{\pi_2,h} - \big(V_0^{\pi_2,h}(S_0) - V_T^{\pi_2,h}(S_T)\big)\Big| \leq \bar{K}_T^{\pi_2,h}\sqrt{2g^{\pi_2,h}(T)\log\frac{2}{\delta_2}}. \tag{0.122}$$

As a result, by applying Lemma 4.4 and (0.120)–(0.122), we get that for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

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

4.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

$$\begin{aligned}
&\Big|\big|R_T^{\pi_1,h} - R_T^{\pi_2,h}\big| - \big|\big[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\big] - \big[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\big]\big|\Big| \\
\le&\Big|R_T^{\pi_1,h} - \big[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\big]\Big| + \Big|R_T^{\pi_2,h} - \big[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_\gamma^{\pi_2,h}(S_T^{\pi_2})\big]\Big|. \qquad (0.123)
\end{aligned}$$

By Corollary 4.16, for any $\delta_1 \in (0,1)$, if $g^{\pi_1,h}(h) \ge 173\log\frac{4}{\delta_1}$, let

$$T_0^{\pi,h}(\delta_1) \coloneqq \min\Big\{T' \ge 1 : g^{\pi,h}(T') \ge 173\log\frac{4}{\delta_1}\Big\}. \qquad (0.124)$$

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

$$\begin{aligned}
&\Big|R_T^{\pi_1,h} - \big(V_0^{\pi_1,h}(S_0) - V_T^{\pi_1,h}(S_T)\big)\Big| \\
&\qquad \le \max\Bigg\{\bar{K}_T^{\pi_1,h}\sqrt{3g^{\pi_1,h}(T)\big(2\log\log(\frac{3}{2}g^{\pi_1,h}(T)) + \log\frac{2}{\delta_1}\big)}, (\bar{K}_T^{\pi_1,h})^2\Bigg\}. \qquad (0.125)
\end{aligned}$$

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

$$\begin{aligned}
&\Big|R_T^{\pi_2,h} - \big(V_0^{\pi_2,h}(S_0) - V_T^{\pi_2,h}(S_T)\big)\Big| \\
&\qquad \le \max\Bigg\{\bar{K}_T^{\pi_2,h}\sqrt{3g^{\pi_2,h}(T)\big(2\log\log(\frac{3}{2}g^{\pi_2,h}(T)) + \log\frac{2}{\delta_2}\big)}, (\bar{K}_T^{\pi_2,h})^2\Bigg\}. \qquad (0.126)
\end{aligned}$$

As a result, by applying Lemma 4.4, 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

$$\begin{aligned}
&\Big|\big|R_T^{\pi_1,h} - R_T^{\pi_2,h}\big| - \big|\big[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\big] - \big[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\big]\big|\Big| \\
\le &\Big|R_T^{\pi_1,h} - \big[V_0^{\pi_1,h}(S_0^{\pi_1}) - V_T^{\pi_1,h}(S_T^{\pi_1})\big]\Big| + \Big|R_T^{\pi_2,h} - \big[V_0^{\pi_2,h}(S_0^{\pi_2}) - V_T^{\pi_2,h}(S_T^{\pi_2})\big]\Big| \\
\le &\max\left\{\bar{K}_T^{\pi_1,h}\sqrt{3g^{\pi_1,h}(T)\big(2\log\log(\frac{3}{2}g^{\pi_1,h}(T)) + \log\frac{4}{\delta}\big)}, (\bar{K}_T^{\pi_1,h})^2\right\} \\
&+ \max\left\{\bar{K}_T^{\pi_2,h}\sqrt{3g^{\pi_2,h}(T)\big(2\log\log(\frac{3}{2}g^{\pi_2,h}(T)) + \log\frac{4}{\delta}\big)}, (\bar{K}_T^{\pi_2,h})^2\right\}. \qquad (0.127)
\end{aligned}$$

4.F Miscellaneous Theorems

4.F.1 Slutsky's Theorem

Theorem 4.14 (see [127, Theorem 7.7.1]). *If $X_t \xrightarrow{(d)} X$ and $Y_t \xrightarrow{(d)} c$, where $c \in \mathbb{R}$ (equivalently $Y_t \xrightarrow{(P)} c$) then we have*

1. $X_t + Y_t \xrightarrow{(d)} X + c$.
2. $X_t Y_t \xrightarrow{(d)} cX$.
3. $\frac{X_t}{Y_t} \xrightarrow{(d)} \frac{X}{c}$, *if* $c \neq 0$.

Remark 4.6. *Since convergence in the almost-sure sense implies convergence in probability, same results hold when $Y_t \xrightarrow{(a.s.)} c$.*

Chapter 5

Asymptotic Normality of Cumulative Cost in Linear Quadratic Regulators

5.1 Overview

In this chapter, we investigate the asymptotic normality of cumulative cost in Linear Quadratic Regulator (LQR) framework. The results of this chapter are published in [106].

5.1.1 Organization

The rest of this chapter is organized as follows. In Section 5.2, we present the system model, assumptions, and the main results. In Section 5.3, we present preliminary results on the cost decomposition, implications of our assumption on the noise process, a preliminary on the central limit theorem for martingale difference sequences, and the proof of the main result. Our concluding remarks are presented in Section 5.4.

5.2 Problem Formulation and Main Result

5.2.1 System Model

Consider a discrete-time linear time-invariant system with full state observation. Let $x_t \in \mathbb{R}^n$ and $u_t \in \mathbb{R}^d$ denote the state and control input at time t. The system starts at a known initial state x_0 and it evolves according to the following dynamics:

$$x_{t+1} = Ax_t + Bu_t + Dv_{t+1}, \quad t \geq 0, \tag{5.1}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times d}$, and $D \in \mathbb{R}^{n\times n}$ are the system dynamic matrices and $\{v_t\}_{t\geq 1}$, $v_{t+1} \in \mathbb{R}^n$, is an independent and identically distributed (i.i.d.) zero-mean noise process

with unit covariance I. At each time t, the system incurs a per-step cost of

$$c(x_t, u_t) = x_t^\intercal Q x_t + u_t^\intercal R u_t,$$

where $Q \succeq 0$ and $R \succ 0$.

We assume that the control inputs are chosen according to a time-homogeneous (and measurable) policy $\pi \colon \mathbb{R}^n \to \mathbb{R}^d$, i.e.,

$$u_t = \pi(x_t).$$

Let Π denote the set of all such policies. For a fixed policy $\pi \in \Pi$, let $\{x_t^\pi\}_{t \geq 0}$ and $\{u_t^\pi\}_{t \geq 0}$ denote the sequence of states and control inputs generated over time. Let

$$\mathcal{C}(\pi, T) \coloneqq \sum_{t=0}^{T-1} c(x_t^\pi, u_t^\pi),$$

denote the cumulative cost incurred by policy π up to time T. Note that our definition of $\mathcal{C}(\pi, T)$ does not include an expectation, so $\mathcal{C}(\pi, T)$ is a random variable. The long-term average performance of policy $\pi \in \Pi$ is given by

$$J(\pi) \coloneqq \limsup_{T \to \infty} \frac{1}{T} \mathbb{E}[\mathcal{C}(\pi, T)],$$

where the expectation is with respect to the noise process $\{v_t\}_{t \geq 1}$. Let

$$J^* = \inf_{\pi \in \Pi} J(\pi),$$

denote the optimal performance. A policy $\pi^* \in \Pi$ is called optimal if $J(\pi^*) = J^*$.

We impose the following standard assumption on the model.

Assumption 5.1. *The pair of matrices (A, B) is controllable, and the pair of matrices $(A, Q^{1/2})$ is observable.*

It is well known (e.g., see [1]) that under Assumption 5.1, the optimal policy exists, is unique, and is given by

$$\pi^*(x_t) = -L^* x_t, \tag{5.2}$$

where the optimal gain L^* is given by

$$L^* = (R + B^\intercal S B)^{-1} B^\intercal S A, \tag{5.3}$$

where S is the unique fixed point of the Discrete Algebraic Riccatti Equation (DARE) given by:

$$P = A^\intercal PA - A^\intercal PB(R + B^\intercal PB)^{-1}B^\intercal PA + Q. \tag{5.4}$$

Moreover the optimal value J^* is given by:

$$J^* = \mathrm{Tr}(SDD^\intercal). \tag{5.5}$$

5.2.2 Main Result

The classical result described above characterizes the behavior of the expected value of $\mathcal{C}(\pi^*, T)$; in particular,

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}[\mathcal{C}(\pi^*, T)] = \mathrm{Tr}(SDD^\intercal) = J^*. \tag{5.6}$$

Our main result characterizes a much stronger *distributional* behavior of $\mathcal{C}(\pi^*, T)$. In particular, we will show that under a mild assumption, loosely speaking, the stochastic process $\mathcal{C}(\pi^*, T)$ converges in distribution to a Gaussian random variable. We will present this statement more precisely in this section.

For our analysis, we impose the following additional assumption on the noise process $\{v_t\}_{t\geq 1}$.

Assumption 5.2. *In addition to being i.i.d. across time and having a unit covariance, the noise sequence $\{v_t\}_{t\geq 1}$ satisfies the following conditions for each time t:*

(A1) *The components of v_t are independent and admit a density f_v that is even.*

(A2) *v_t is uniformly bounded, that is, there exists a $K_v \in \mathbb{R}_+$ such that $\|v_t\| \leq K_v$ almost surely.*

(A3) *For matrices D and S, we have $\mathrm{Var}(v_t^\intercal D^\intercal SDv_t) \neq 0$.*

For the ease of notation, let $\{(x_t^*, u_t^*)\}_{t\geq 0}$ denote the (stochastic) trajectory $\{(x_t^{\pi^*}, u_t^{\pi^*})\}_{t\geq 0}$ of the optimal policy, $w_t = Dv_t$ denote the noise at time t, and $A^* = A - BL^*$ denote the closed loop dynamics under the optimal policy. Define:

$$M \coloneqq \mathbb{E}[w_t^\intercal Sw_t w_t^\intercal Sw_t] - \left(\mathbb{E}[w_t^\intercal Sw_t]\right)^2$$

which is a scalar constant. We now define a process $\{N_T\}_{T\geq 1}$ where:

$$N_T \coloneqq \sum_{t=0}^{T-1}\Big[M + 4(A^*x_t^*)^\intercal SDD^\intercal SA^*x_t^*\Big]$$

and let $\{\nu_T\}_{T\geq 1}$ be a stopping time corresponding to $\{N_T\}_{T\geq 1}$ given by

$$\nu_T \coloneqq \min_{\tau\geq 1}\left\{\tau : \sum_{t=1}^{\tau} N_t \geq T\right\}. \tag{5.7}$$

Our main result is the following theorem.

Theorem 5.1. *We have that*

$$\frac{\mathcal{C}(\pi^*, \nu_T) - \nu_T J^*}{\sqrt{T}} \xrightarrow{(d)} \mathcal{N}(0,1) \text{ as } T \to \infty.$$

The proof is presented in Section 5.3.

Above theorem is presented in terms of the stopping time in Eq. (5.7). In the following lemma, we establish the growth rate of this stopping time in the almost sure sense.

Lemma 5.1. *The stopping time* $\{\nu_T\}_{T\geq 1}$ *satisfies:*

$$\nu_T \asymp T, \quad a.s.$$

The proof is presented in Appendix 5.A.

Theorem 5.1 and Lemma 5.1 together give a complete picture of distributional behavior of $\mathcal{C}(\pi^*, \nu_T)$, which in the order, matches with the asymptotic normality results in other frameworks.

5.3 Proof of Theorem 5.1

In this section we present the proof of Theorem 5.1. Our proof relies on three techniques: (i) a completion of square argument to establish a decomposition of the cumulative cost, similar to one used in [117]; (ii) some implications of noise having an even density; and (iii) the CLT for bounded martingale difference sequences [128].

5.3.1 Decomposition of Cumulative Cost

The following lemma provides a decomposition of the cumulative cost of any arbitrary policy π.

Lemma 5.2. *For any* $\pi \in \Pi$*, we have*

$$\mathcal{C}(\pi, T) = x_0^\intercal S x_0 - (x_T^\pi)^\intercal S x_T^\pi$$

$$+\sum_{t=0}^{T-1}\left[(u_t^\pi + L^* x_t^\pi)^\intercal (R + B^\intercal S B)(u_t^\pi + L^* x_t^\pi)\right.$$
$$+\sum_{t=0}^{T-1}\left[2(Ax_t^\pi + Bu_t^\pi)^\intercal S w_{t+1} + w_{t+1}^\intercal S w_{t+1}\right],$$

where matrices L^* and S are given by (5.3) *and* (5.4)*.*

The proof is similar to the decomposition of $\mathbb{E}[\mathcal{C}(\pi, T)]$ presented in [117] and is presented in Appendix 5.B for completeness.

In the following Lemma, we use Lemma 5.2 to characterize the cumulative cost function induced by the optimal policy $\mathcal{C}(\pi^*, T)$.

Lemma 5.3. *For the optimal policy π^*, we have*

$$\mathcal{C}(\pi^*, T) = x_0^\intercal S x_0 - (x_T^*)^\intercal S x_T^* + \sum_{t=0}^{T-1}\left[2(A^* x_t^*)^\intercal S w_{t+1} + w_{t+1}^\intercal S w_{t+1}\right].$$

Proof. The result follows by substituting $u_t^* = -L^* x_t^*$ in Lemma 5.2, and substituting $x_t^{\pi^*}$ with x_t^*.

□

5.3.2 Implications of the Assumption on the Noise

The assumed symmetry on the components of v_t (i.e., the components of v_t admitting a density f_v that is even) has important implications in our analysis. We show this structure implies that a certain cubic transformation of the noise has zero mean. Following lemma summarizes these structures.

Lemma 5.4. *Under Assumption 5.2, we have the following for any time t:*

1. *For any odd $k \in \mathbb{N}$ and any component $i \in \{1, \ldots, n\}$, $\mathbb{E}[v_t(i)^k] = 0$.*
2. *For any $i, j \in \{1, \ldots, n\}$, $i \neq j$, $\mathbb{E}[v_t(i) v_t(j)^2] = 0$.*
3. *For any arbitrary matrix M, let $y_t = M v_t$, then $\mathbb{E}[y_t y_t^\intercal y_t] = \mathbf{0}$.*

Proof is presented in Appendix 5.C.

Furthermore, the boundedness assumption on the noise sequence $\{v_t\}_{t\geq 1}$ implies the boundedness of optimal state trajectory $\{x_t^*\}_{t\geq 0}$. This is presented in the following lemma.

Lemma 5.5. *Under Assumption 5.2, there exists a universal constant $K_x \in \mathbb{R}_+$ (which depends only on K_v and x_0) such that*

$$\|x_t^*\| \leq K_x, \quad a.s., \quad \forall t \geq 0.$$

This is a classic result and its proof exists in many resources. We included a proof in Appendix 5.D for completeness.

5.3.3 CLT for Martingale Difference Sequences

The usual CLT for martingale difference sequences is the Lindeberg-Levy CLT for triangular array of martingale difference sequences. In our analysis, we use an implication of Lindeberg-Levy CLT stated in [128]. Since this version of the CLT is not as well known, we restate it below for completeness.

Let $\{\delta_t\}_{t\geq 1}$, $\delta_t \in \mathbb{R}$, be a martingale difference sequence adapted to some filtration sequence $\{\mathcal{G}_t\}_{t\geq 0}$, i.e.:

$$\mathbb{E}[\delta_t | \mathcal{G}_{t-1}] = 0.$$

In addition, for all $t \geq 1$, let $\Delta_t \coloneqq \sum_{\tau=1}^{t} \delta_\tau$ denote the martingale process corresponding to $\{\delta_t\}_{t\geq 1}$. Let $\rho_t^2 \coloneqq \mathbb{E}[\delta_t^2 | \mathcal{G}_{t-1}]$ denote the conditional variance of δ_t. For any $T \geq 0$, define the stopping time μ_T as:

$$\mu_T = \min_{\tau \geq 1} \left\{ \tau : \sum_{t=1}^{\tau} \rho_t^2 \geq T \right\}.$$

The following theorem states a version of central limit theorem for the martingale sequence $\{\Delta_t\}_{t\geq 1}$.

Theorem 5.2 (see [128, Theorem 35.11])**.** *Suppose the martingale difference sequence $\{\delta_t\}_{t\geq 1}$ satisfies the following conditions:*

(C1) *For all $t \geq 1$, $|\delta_t|$ is uniformly bounded, i.e., there exists a $K_\delta \in \mathbb{R}_+$, such that:*

$$|\delta_t| \leq K_\delta, \quad a.s.$$

(C2) *We have:*

$$\sum_{t=1}^{\infty} \mathbb{E}[\delta_t^2 | \mathcal{G}_{t-1}] = \infty.$$

Then we have:

$$\frac{\Delta_{\mu_T}}{\sqrt{T}} \xrightarrow{(d)} \mathcal{N}(0,1) \text{ as } T \to \infty.$$

In the subsequent subsection, we show some of the terms in the cumulative cost $\mathcal{C}(\pi^*, T)$ satisfy martingale difference property. We then use Theorem 5.2 to derive the distribution of the cumulative cost.

5.3.4 Preliminary Results

Define the filtration to be the sigma field generated by the sequence of states and control actions, i.e., $\mathcal{F}_t \coloneqq \sigma(x^*_{0:t}, u^*_{0:t})$. Using Lemma 5.3 and the fact that $J^* = \mathbb{E}[w_{t+1}^\intercal S w_{t+1}]$, we rewrite $\mathcal{C}(\pi^*, T) - TJ^*$ as following:

$$\begin{aligned}\mathcal{C}(\pi^*, T) - TJ^* &= x_0^\intercal S x_0 - (x_T^*)^\intercal S x_T^* \\ &+ \sum_{t=0}^{T-1} \Big[2(A^* x_t^*)^\intercal w_{t+1} + w_{t+1}^\intercal S w_{t+1} - \mathbb{E}[w_{t+1}^\intercal S w_{t+1}]\Big].\end{aligned}$$

To simplify the algebra, we define following intermediate variables for $t \geq 0$:

$$a_{t+1} \coloneqq w_{t+1}^\intercal S w_{t+1}, \tag{5.8}$$

$$b_{t+1} \coloneqq 2(A^* x_t^*)^\intercal S w_{t+1}, \tag{5.9}$$

$$c_{t+1} \coloneqq \mathbb{E}[w_{t+1}^\intercal S w_{t+1}], \tag{5.10}$$

$$z_{t+1} \coloneqq a_{t+1} + b_{t+1} - c_{t+1}. \tag{5.11}$$

As a result of above reparametrization, we have:

$$\mathcal{C}(\pi^*, T) - TJ^* = \sum_{t=0}^{T-1} z_{t+1} + (x_0)^\intercal S(x_0) - (x_T^*)^\intercal S(x_T^*).$$

We show that the sequence $\{z_t\}_{t\geq 1}$ is a martingale difference sequence satisfying conditions (C1) and (C2) in Theorem 5.2. We first establish the properties of variables a_{t+1}, b_{t+1}, and c_{t+1} in the following proposition.

Proposition 5.1. *For all $t \geq 0$, we have:*

(P1) $\mathbb{E}[b_{t+1}|\mathcal{F}_t] = 0$.

(P2) $\mathbb{E}[a_{t+1}|\mathcal{F}_t] = c_{t+1}$.

(P3) $\mathbb{E}[a_{t+1}^2|\mathcal{F}_t] = \mathbb{E}[a_{t+1}^2]$.

(P4) $\mathbb{E}[c_{t+1}a_{t+1}|\mathcal{F}_t] = c_{t+1}^2$.

(P5) $\mathbb{E}[c_{t+1}b_{t+1}|\mathcal{F}_t] = 0$.

(P6) $\mathbb{E}[a_{t+1}b_{t+1}|\mathcal{F}_t] = 0$.

Proof. These properties are the consequences of the assumption on the noise process.

(P1) Follows by the fact that x_t^* is $\mathcal{F}_t$-measurable and based on Assumption 5.2, $w_{t+1} = Dv_{t+1}$ is zero mean and independent of $\mathcal{F}_t$.

(P2) Follows from independence of w_{t+1} from $\mathcal{F}_t$, and the definition of c_{t+1}.

(P3) Follows from independence of w_{t+1} from $\mathcal{F}_t$.

(P4) Follows from following equations:

$$\mathbb{E}[c_{t+1}a_{t+1}|\mathcal{F}_t] \overset{(a)}{=} c_{t+1}\mathbb{E}[a_{t+1}|\mathcal{F}_t] \overset{(b)}{=} c_{t+1}^2,$$

where (a) follows from the fact that c_{t+1} is not a random variable and (b) follows from Property (P2).

(P5) Follows from following equations:

$$\mathbb{E}[c_{t+1}b_{t+1}|\mathcal{F}_t] \overset{(c)}{=} c_{t+1}\mathbb{E}[b_{t+1}|\mathcal{F}_t] \overset{(d)}{=} 0,$$

where (c) follows from the fact that c_{t+1} is not a random variable and (d) follows from Property (P1).

(P6) Follows from Lemma 5.4. To show this, let:

$$y_{t+1} \coloneqq S^{1/2}Dv_{t+1} = S^{1/2}w_{t+1}$$

we have:

$$\begin{aligned}\mathbb{E}[a_{t+1}b_{t+1}|\mathcal{F}_t] &\overset{(e)}{=} \mathbb{E}[2(x_t^*)^\intercal(A^*)^\intercal S^{1/2}S^{1/2}w_{t+1}w_{t+1}^\intercal S^{1/2}S^{1/2}w_{t+1}|\mathcal{F}_t] \\ &\overset{(f)}{=} 2(x_t^*)^\intercal(A^*)^\intercal S^{1/2}\mathbb{E}[y_t y_t^\intercal y_t] \overset{(g)}{=} 0,\end{aligned}$$

where (e) follows from the fact that $S \succ 0$, (f) follows from the fact that $S^{1/2}$ is symmetric, and (g) follows from Lemma 5.4 part (3).

□

5.3.5 Proof of Theorem 5.1

To prove the theorem, we first verify the conditions of Theorem 5.2 for the sequence $\{z_t\}_{t\geq 1}$. First, recall that by definition, $z_{t+1} = a_{t+1} + b_{t+1} - c_{t+1}$. We have:

$$\mathbb{E}[z_{t+1}|\mathcal{F}_t] = \mathbb{E}[a_{t+1} - c_{t+1}|\mathcal{F}_t] + \mathbb{E}[b_{t+1}|\mathcal{F}_t] \overset{(a)}{=} 0,$$

where (a) follows from Properties (P1) and (P2) in Proposition 5.1. We now verify conditions (C1) and (C2) in Theorem 5.2.

5.3.5.1 Verifying (C1)

We know a_{t+1} and c_{t+1} are uniformly bounded by (A2) in Assumption 5.2. By Lemma 5.5 and (A2) in Assumption 5.2, we know $|b_{t+1}|$ is uniformly bounded. As a result, $|z_{t+1}|$ is uniformly bounded almost surely.

5.3.5.2 Verifying (C2)

We compute the conditional expectation of z_{t+1}^2 given the filtration $\mathcal{F}_t$ as following:

$$\begin{aligned}
&\mathbb{E}[z_{t+1}^2|\mathcal{F}_t] = \mathbb{E}[(a_{t+1} + b_{t+1} - c_{t+1})^2|\mathcal{F}_t] \\
=&\mathbb{E}[a_{t+1}^2|\mathcal{F}_t] + \mathbb{E}[b_{t+1}^2|\mathcal{F}_t] + \mathbb{E}[c_{t+1}^2|\mathcal{F}_t] \\
+&2\mathbb{E}[a_{t+1}b_{t+1}|\mathcal{F}_t] - 2\mathbb{E}[c_{t+1}a_{t+1}|\mathcal{F}_t] - 2\mathbb{E}[c_{t+1}b_{t+1}|\mathcal{F}_t] \\
\overset{(b)}{=}&\mathbb{E}[a_{t+1}^2|\mathcal{F}_t] + \mathbb{E}[b_{t+1}^2|\mathcal{F}_t] + \mathbb{E}[c_{t+1}^2|\mathcal{F}_t] - 2\mathbb{E}[a_{t+1}c_{t+1}|\mathcal{F}_t] \\
\overset{(c)}{=}&\mathbb{E}[a_{t+1}^2] - c_{t+1}^2 + \mathbb{E}[b_{t+1}^2|\mathcal{F}_t]
\end{aligned} \tag{5.12}$$

where (b) follows from properties (P5) and (P6) in Proposition 5.1 and (c) follows from properties (P3) and (P4). Now the term $\mathbb{E}[a_{t+1}^2] - c_{t+1}^2$ is independent of t and depends only on the density f_v. Therefore, by Jensen's inequality and (A3) in Assumption 5.2, we know that there exists an $\underline{\epsilon} > 0$, such that:

$$\mathbb{E}[a_{t+1}^2] - c_{t+1}^2 > \underline{\epsilon}, \tag{5.13}$$

for all $t \geq 0$. By definition, we know $\mathbb{E}[b_{t+1}^2|\mathcal{F}_t] \geq 0$ for all $t \geq 0$. As a result, we have:

$$\sum_{t=0}^{T-1} z_{t+1} \geq T\underline{\epsilon}.$$

Implying that: $\lim_{T\to\infty} \sum_{t=0}^{T-1} \mathbb{E}[z_{t+1}^2|\mathcal{F}_t] = \infty$, almost surely, verifying the condition (C2).

5.3.5.3 Concluding the proof

Since the conditions (C1) and (C2) hold for the sequence $\{z_t\}_{t\geq 1}$, by Theorem 5.2, we have:

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

By Lemma 5.5, we know $(x_T^*)^\intercal S(x_T^*)$ is almost surely bounded for all $T \geq 0$. Moreover $x_0^T S x_0$ is a constant. Therefore, we have:

$$\lim_{T\to\infty} \frac{x_0^\intercal S x_0 - (x_T^*)^\intercal S x_T^*}{\sqrt{T}} \to 0, \quad a.s.$$

As a result, by using Slutsky's Theorem (see [127, Theorem 7.7.3]), we get:

$$\frac{\mathcal{C}(\nu_T, \pi^*) - \nu_T J^*}{\sqrt{T}} \xrightarrow{(d)} \mathcal{N}(0,1).$$

Remark 5.1. *In the proof of Theorem 5.1, each of the two sequences $\{a_{t+1} - c_{t+1}\}_{t\geq 0}$ and $\{b_{t+1}\}_{t\geq 0}$ is a martingale difference sequence. However, these two sequences are dependent, and therefore, the fact that each of them converges in distribution does not trivially imply that their summation also converges in distribution. As a result, applying Theorem 5.2 on each of these sequences individually would not imply the desired result. Therefore, characterizing the behavior of the sequence $\{a_{t+1} + b_{t+1} - c_{t+1}\}_{t\geq 0}$ similar to the approach in our proof is necessary.*

5.4 Conclusion

In this chapter we have established the asymptotic normality of the cumulative cost in the LQR framework. We have shown that under mild assumptions on the noise process, asymptotic normality holds for the distribution of the cumulative cost only using first principles. Our result gives a complete description of the cost distribution induced by the optimal policy. We believe this analysis opens new doors to understanding the distributional behavior of the cumulative cost and may pave the way to derive confidence bounds for this framework. These confidence bounds can be used in risk-averse or distributional reinforcement learning within this setup. A natural extension of this work is to derive similar results for larger classes of policies or to weaken the assumption on the noise sequence to be Gaussian or sub-Gaussian.

Appendices to Chapter 5

5.A Proof of Lemma 5.1

Using Eq. (5.12), we have:

$$\mathbb{E}[z_{t+1}^2|\mathcal{F}_t] = \mathbb{E}[a_{t+1}^2] - c_{t+1}^2 + \mathbb{E}[b_{t+1}^2|\mathcal{F}_t].$$

By (A3) in Assumption 5.2 and Jensen's inequality, we know there exists a $\underline{\epsilon} > 0$ such that $\mathbb{E}[a_{t+1}^2] - c_{t+1}^2 > \underline{\epsilon}$. Since $\mathbb{E}[b_{t+1}^2|\mathcal{F}_t] > 0$, we have:

$$\liminf_{T\to\infty} \frac{N_T}{T} = \liminf_{T\to\infty} \frac{\sum_{t=0}^{T-1} \mathbb{E}[z_{t+1}^2|\mathcal{F}_t]}{T} \geq \underline{\epsilon} > 0, \quad a.s.$$

From the definition of b_{t+1}, it is clear that there exists a constant $C \in \mathbb{R}_+$ such that $\mathbb{E}[b_{t+1}^2|\mathcal{F}_t] \leq C\|x_t\|^2$ for all $t \geq 0$. As a result, by following arguments similar to the proof of Proposition 2.5, we have:

$$\limsup_{T\to\infty} \frac{\sum_{t=0}^{T-1} \mathbb{E}[b_{t+1}^2|\mathcal{F}_t]}{T} < \infty, \quad a.s.$$

Since the term $\mathbb{E}[a_{t+1}^2] - c_{t+1}^2$ is independent of t and only depends on the density f_v, there exists an $\bar{\epsilon} > 0$, such that:

$$\mathbb{E}[a_{t+1}^2] - c_{t+1}^2 < \bar{\epsilon}.$$

As a result,

$$\limsup_{T\to\infty} \frac{N_T}{T} = \limsup_{T\to\infty} \frac{\sum_{t=0}^{T-1} \mathbb{E}[b_{t+1}^2|\mathcal{F}_t]}{T} + \bar{\epsilon} < \infty,$$

almost surely, implying that $N_T \asymp T$ and therefore $\nu_T \asymp T$, almost surely.

5.B Proof of Lemma 5.2

5.B.1 Preliminary Result

The proof of this lemma is similar to the regret decomposition in Chapter 3. Following algebraic lemma is adapted from [129, Lemma 6.1].

Lemma .6. *We have following statements:*

1. *(Algebraic completion of square) For $x \in \mathbb{R}^n$ and $u \in \mathbb{R}^d$ and matrices A, B, S, R with appropriate dimensions, we have*

$$\begin{aligned} &u^\intercal R u + (Ax+Bu)^\intercal P (Ax+Bu) + x^\intercal Q x \\ =&[u+L(P,R,A,B)x]^\intercal [R+B^\intercal P B][u+L(P,R,A,B)x] \\ &+ x^\intercal K(P,A,B,R,Q)x, \end{aligned} \tag{0.14}$$

with $L(P,R,A,B) \coloneqq -[R+B^\intercal P B]^{-1} B^\intercal P A$, and $K(P,A,B,R,Q)$ defined as:

$$K(P,A,B,R,Q) = Q + A^\intercal P A - A^\intercal P B (R + B^\intercal P B)^{-1} B^\intercal P A.$$

2. *The Discrete Algebraic Riccati Equation (DARE) in Eq. (5.4), i.e. $K(P,A,B,R,Q) = P$ has a unique positive definite fixed point solution $S \succeq 0$. As a result, we have:*

$$\begin{aligned} &u^\intercal R u + (Ax+Bu)^\intercal S (Ax+Bu) + x^\intercal Q x \\ =&[u+L(S,R,A,B)x]^\intercal [R+B^\intercal S B][u+L(S,R,A,B)x] + x^\intercal S x \end{aligned}$$

5.B.2 Proof of Lemma 5.2

Proof. The proof follows by applying Lemma .6. We start by adding and subtracting the term $(x_T^\pi)^\intercal S(x_T^\pi)$ to the expression. Recall that $\{x_t^\pi\}_{t\geq 0}$ and $\{u_t^\pi\}_{t\geq 0}$ denote the sequences of states and actions induced by the policy π. We have:

$$\begin{aligned} \mathcal{C}(\pi, T) &= \sum_{t=0}^{T-1} \left[(x_t^\pi)^\intercal Q(x_t^\pi) + (u_t^\pi)^\intercal R(u_t^\pi)\right] + (x_T^\pi)^\intercal S(x_T^\pi) - (x_T^\pi)^\intercal S(x_T^\pi) \\ &= \sum_{t=0}^{T-2} \left[(x_t^\pi)^\intercal Q(x_t^\pi) + (u_t^\pi)^\intercal R(u_t^\pi)\right] - (x_T^\pi)^\intercal S x_T^\pi \\ &\quad + \left[(x_{T-1}^\pi)^\intercal Q(x_{T-1}^\pi) + (u_{T-1}^\pi)^\intercal R(u_{T-1}^\pi) + (x_T^\pi)^\intercal S(x_T^\pi)\right] \\ &= \Big[\sum_{t=0}^{T-2} (x_t^\pi)^\intercal Q(x_t^\pi) + (u_t^\pi)^\intercal R(u_t^\pi)\Big] - (x_T^\pi)^\intercal S(x_T^\pi) \\ &\quad + (x_{T-1}^\pi)^\intercal Q(x_{T-1}^\pi) + (u_{T-1}^\pi)^\intercal R(u_{T-1}^\pi) \\ &\quad + (Ax_{T-1}^\pi + Bu_{T-1}^\pi + w_T)^\intercal S(Ax_{T-1}^\pi + Bu_{T-1}^\pi + w_T) \\ &\overset{(a)}{=} \Big[\sum_{t=0}^{T-2} (x_t^\pi)^\intercal Q(x_t^\pi) + (u_t^\pi)^\intercal R(u_t^\pi)\Big] + (x_{T-1}^\pi)^\intercal S(x_{T-1}^\pi) - (x_T^\pi)^\intercal S(x_T^\pi) \\ &\quad + \Big[(u_{T-1}^\pi + L^* x_{T-1}^\pi)^\intercal (R + B^\intercal S B)(u_{T-1}^\pi + L^* x_{T-1}^\pi) \end{aligned}$$

$$+ w_T^\intercal S w_T + 2(Ax_{T-1}^\pi + Bu_{T-1}^\pi)^\intercal S w_T \Big],$$

where (a) follows from Lemma .6, with L^* being the RHS of Eq. (5.3). By repeating the same argument, we get:

$$\mathcal{C}(\pi, T) = x_0^\intercal S x_0 - x_T^\intercal S x_T$$
$$+ \sum_{t=1}^{T-1} \Big[(u_t^\pi + L^* x_t^\pi)^\intercal (R + B^\intercal S B)(u_t^\pi + L^* x_t^\pi) + 2(Ax_t^\pi + Bu_t^\pi)^\intercal S w_{t+1} + w_{t+1}^\intercal S w_{t+1} \Big]. \square$$

5.C Proof of Lemma 5.4

For an odd n, Assumption 5.2, implies that for all $1 \leq i \leq n$ and for all $t \geq 0$, we have:

$$\mathbb{E}[v_t(i)^k] = \int_{-K_v}^{K_v} v^k f_v(v) dv.$$

1) Proof of part (1): The Probability Density Function (PDF) f_v is an even function and for odd $k \in \mathbb{N}$, v^k is an odd function . As a result, $v^k f_v$ is an odd function, and integrating an odd function from $-K_v$ to K_v is 0.
2) Proof of part (2): For all $i \neq j$, we have:

$$\mathbb{E}[v_t(i) v_t(j)^2] \overset{(a)}{=} \mathbb{E}[v_t(i)] \mathbb{E}[v_t(j)^2] \overset{(b)}{=} 0,$$

where (a) follows from the independence of the components of v_t, and (b) follows from part (1) of this lemma.
3) Proof of part (3): Let $m(i, j)$ denote the (i, j)-th component of M. Then Recall that we have

$$y_t(i) = [M v_t](i) = \sum_{j=1}^{n} m(i, j) v_t(j).$$

It is clear that $\mathbb{E}[y_t(i)] = 0$ for all $t \geq 0$ by the linearity of the expectation operator. We show that for all $i \in \{1, \ldots, n\}$ and all $t \geq 0$, we have: $\mathbb{E}[y_t(i)^3] = 0$. By multinomial theorem, we have:

$$\mathbb{E}\Big[y_t(i)^3\Big] = \mathbb{E}\Big[\Big(\sum_{j=1}^{n} m(i, j) v_t(j)\Big)^3\Big]$$
$$= \mathbb{E}\Bigg[\sum_{k_1 + \cdots + k_n = 3} \binom{3}{k_1, \ldots, k_n} (m(i, 1) v_t(1))^{k_1} \ldots (m(i, n) v_t(n))^{k_n}\Bigg].$$

Where the notation $\sum_{k_1+\cdots+k_n=3}$ denotes all possible tuples $(k_1,\ldots,k_n)$ such that $k_1+\cdots+k_n=3$. Let the tuple $(k_1',\ldots,k_n')$ be a decreasing permutation of $(k_1,\ldots,k_n)$, i.e.,

$$k_1' \geq k_2' \geq \cdots \geq k_n'.$$

Since $k_1+\cdots+k_n=3$, there are only 3 choices for the tuple $(k_1',\ldots,k_n')$. These choices are $(3,0,\ldots,0)$ or $(2,1,\ldots,0)$ or $(1,1,1,0,\ldots,0)$. By Parts (1) and (2), we get:

1. For any $i\in\{1,\ldots,n\}$, $\mathbb{E}[v_t(i)^3]=0$.

2. For any $i,j\in\{1,\ldots,n\}$, $i\neq j$, $\mathbb{E}[v_t(i)^2v_t(j)]=0$.

3. For any $i,j,k\in\{1,\ldots,n\}$, $i\neq j\neq k$, $\mathbb{E}[v_t(i)v_t(j)v_t(k)]=0$.

This implies that all the permutations which are mapped to the tuples $(3,0,\cdots,0)$ or $(2,1,\cdots,0)$ or $(1,1,1,0,\cdots,0)$ have zero expected value; therefore, $\mathbb{E}[y_t(i)^3]=0$. Next we show for all $i,j\in\{1,\ldots,n\}$ such that $i\neq j$, we have: $\mathbb{E}[y_t(i)^2y_t(j)]=0$. By using the multinomial theorem, we have:

$$\begin{aligned}\mathbb{E}\Big[y_t(i)^2\Big] &= \mathbb{E}\bigg[\Big(\sum_{j=1}^{n} m(i,j)v_t(j)\Big)^2\bigg]\\ =&\mathbb{E}\Bigg[\sum_{k_1+\cdots+k_n=2}\binom{2}{k_1,\ldots,k_n}(m(i,1)v_t(1))^{k_1}\ldots(m(i,n)v_t(n))^{k_n}\Bigg].\end{aligned}$$

Again let the tuple $(k_1',\ldots,k_n')$ be a decreasing permutation of $(k_1,\ldots,k_n)$. Since $k_1+\cdots+k_n=2$, there are only 2 choices for the tuple $(k_1',\ldots,k_n')$. These choices are $(2,0,\ldots,0)$ or $(1,1,0,\ldots,0)$. Now since $y_t(j)=\sum_{k=1}^{n} m(j,k)v_t(k)$, expanding $y_t(i)^2y_t(j)$ and ordering the permutations we again end up with 3 choices for $(k_1',\ldots,k_n')$, i.e., $(3,0,\ldots,0)$, $(2,1,\ldots,0)$, and $(1,1,1,0,\ldots,0)$. By repeating the arguments similar to the previous part, we have that $\mathbb{E}[y(i)^2y(j)]=0$. At last, since

$$\mathbb{E}[yy^{\intercal}y]=\begin{bmatrix}y(1)\\ \vdots\\ y(n)\end{bmatrix}\Big(y(1)^2+\cdots y(n)^2\Big), \tag{0.15}$$

all the terms are either of the form $\mathbb{E}[y(i)^3]$ or $\mathbb{E}[y(i)^2y(j)]$, $i\neq j$, implying that:

$$\mathbb{E}[yy^{\intercal}y]=\mathbf{0}.$$

5.D Proof of Lemma 5.5

Given that $\|v_t\| \leq K_v$, we have that $\|w_t\| \leq \|D\|\|v_t\| =: K_w$. Let $\rho_{\max} = \lambda_{\max}(A^*) < 1$ (recall $A^* = A - BL^*$) since L^* is a stabilizing controller gain. Pick an $\varepsilon > 0$ such that $\rho_{\max} + \varepsilon < 1$. Then, by Gelfand's formula, we know that there exists a T_0 such that for all $t > T_0$, $\|(A^*)^t\| < \rho_{\max} + \varepsilon$. By the convolutional form of the output, we have that for $T > T_0$,

$$\begin{aligned}
\|x_T\| &= \|(A^*)^T x_0\| + \left\| \sum_{\tau=1}^{T} (A^*)^\tau w_{T-\tau} \right\| \\
&\leq \|(A^*)^T\|\|x_0\| + \sum_{\tau=1}^{T} \|(A^*)^\tau\|\|w_{T-\tau}\| \\
&\leq \|(A^*)^T\|\|x_0\| + K_w \sum_{\tau=1}^{T} \|(A^*)^\tau\| \\
&\leq (\rho_{\max} + \varepsilon)^T \|x_0\| + K_w \sum_{\tau=1}^{T} (\rho_{\max} + \varepsilon)^\tau \\
&\overset{(a)}{\leq} (\rho_{\max} + \varepsilon)^{T_0} \|x_0\| + \frac{K_w}{1 - (\rho_{\max} + \varepsilon)} =: K_x
\end{aligned}$$

where (a) uses the fact that $\rho_{\max} + \varepsilon < 1$.

Chapter 6

Conclusions and Future Research

6.1 Conclusion

In this thesis, we investigated fundamental challenges in the learning and control of Markov Jump Linear Systems (MJLS) and analyzed the concentration of cumulative rewards in Markov Decision Processes (MDP) and linear systems. Our work focused on (i) solving the problem of system identification in MJLS and establishing theoretical convergence guarantees, (ii) integrating the identification algorithms into a model-based reinforcement learning framework, along with deriving theoretical sub-linear regret bounds for the proposed approach, and (iii) examining the concentration properties of cumulative rewards under various planning policies and explored their implications for the learning process. We believe these contributions are essential for advancing our understanding of the challenges in integrating learning into control of the dynamical systems.

Beyond the immediate results of this thesis, the auxiliary and intermediate results established here may be of independent interest. The decompositions for cumulative regret and cumulative reward introduced in this work can be leveraged to derive guarantees for other algorithms in comparable settings. Additionally, the decomposition of some of the investigated problems into martingale structures, combined with the application of martingale convergence and concentration techniques, can be leveraged to derive new results in stochastic systems, control theory, and learning algorithms.

6.2 Summary of Results

6.2.1 Learning in Markov Jump Linear Systems

We investigated the problem of system identification (or learning) in autonomous Markov jump linear systems. We proposed a variant of least squares algorithm called switched

least squares specifically tailored for this framework. We proved strong consistency of this algorithm and derived its almost-sure rate of convergence. Our analysis involves using the notion of stability in the average sense in MJLS and establishing that it is a sufficient condition for the convergence of the switched least squares algorithm. Additionally, we explored the relationships between stability in the average sense and other notions of stability, such as mean square stability and almost-sure stability, providing other sufficient conditions for the convergence of switched least squares algorithm. The proof techniques in this chapter rely on classical regression convergence arguments, the relationship between the growth rate of the covariate process and system stability, and the results on growth rate of martingale difference sequences.

6.2.2 Learning and Control in Markov Jump Linear Systems

We investigated the problem of model-based reinforcement learning within the framework of Markov jump linear systems (MJLS). In this context, we first derived a general regret decomposition for the class of adaptive linear state-feedback policies. We then introduced a model-based reinforcement learning algorithm based on certainty-equivalence algorithm that uses the switched least squares method for system identification. For this algorithm, we provided almost-sure sub-linear upper bounds on regret.

As part of our analysis, we established convergence guarantees for the switched least squares algorithm in the closed-loop system identification setting. The proof techniques used in this chapter are based on a novel regret decomposition, the convergence rate of the system identification method, and convergence of martingale sequences. Our results hold on a specific subset of the sample space, and we further explored the relationship between system stability, the duration of the initial system identification phase, and the properties of this subset.

6.2.3 Concentration of Reward in Markov Decision Processes

We investigated the problem of cumulative reward concentration in finite-state and finite-action Markov decision processes (MDPs). In the average reward framework, we established both asymptotic and non-asymptotic concentration properties of cumulative rewards. In the asymptotic setup, we established the law of large numbers, the central limit theorem, and the law of iterated logarithm for cumulative rewards. In the non-asymptotic setup, we derived Azuma–Hoeffding-type inequalities and a non-asymptotic version of the law of iterated logarithm. While our results are initially established for a subset of stationary policies, through analyzing different categories of MDPs, we proposed sufficient conditions to

extend these concentration results to broader subsets of stationary policies, expanding their application. In addition, we investigated two implications of our results, (i) we provided a concentration result for the difference between the cumulative reward of any two stationary policies, and (ii) we proved that two common notions of regret defined in the reinforcement learning literature are rate-equivalent.

In addition, we extended the non-asymptotic concentration results to two other MDP setups (i.e., finite-horizon setup and the infinite-horizon discounted setup) and investigated their implications. Our proof techniques rely on a novel martingale decomposition of the cumulative reward, the properties of the solutions of policy evaluation fixed-point equation and asymptotic and non-asymptotic concentration of martingale sequences.

6.2.4 Concentration of Cost in Linear Quadratic Regulators

We investigated the asymptotic concentration properties of cumulative cost in the framework of Linear Quadratic Regulators (LQRs). Since in the LQR framework, the state and action spaces are continuous and non-compact, the concentration results established in the finite-state and finite-action MDP setup are not directly applicable to this framework. By using a different approach, we established a central limit theorem for the cumulative cost of the optimal policy. Our proof techniques rely on a decomposition of cumulative cost for the LQR problem, properties of i.i.d. noise sequences with even densities, and central limit theorem for martingale sequences.

6.3 Future Work

This thesis explored some of the key challenges in integrating learning with the control of dynamical systems. The methodologies and theoretical findings in this thesis establish a solid basis for further research in this domain. In the sequel, we outline several future research directions closely related to this work.

6.3.1 System Identification

The results presented in Chapter 2 are focused on the system identification of Markov jump linear systems in a full-state observation framework, where both continuous and discrete states are directly observed by the agent. However, in some applications, this assumption may not be practical. Investigating the system identification problem for Markov jump linear systems in a partial state observation framework is both a challenging and important direction for future research. Furthermore, investigating the performance of switched least

squares system identification in settings where switching is not necessarily governed by a Markov process is another interesting direction. Applying the system identification algorithm proposed in this thesis to practical scenarios, such as network control systems and other dynamical systems with abrupt changes in their dynamics can show the potential of this algorithm in industry. Moreover, most current system identification results, including those in this thesis, rely on stability as a sufficient condition. Exploring ways to relax these assumptions or establish their necessity represents another valuable area for future research.

6.3.2 Control of Dynamical Systems

Model-based and model-free reinforcement learning within the framework of linear quadratic regulators has been extensively studied. However, extending these approaches to a broader range of nonlinear and time-varying systems remains an open challenge. In this thesis, we focused on reinforcement learning in Markov jump linear systems as an example of such systems. A natural extension of this work is to adapt the developed algorithms and guarantees to other classes of complex systems, such as Markov jump nonlinear systems, control-affine systems, and bilinear systems. These systems have diverse applications, and advancing reinforcement learning techniques in these settings is both theoretically significant and practically important.

One of the key challenges in our analyses was ensuring the almost sure stability of systems throughout the learning process, i.e., guaranteeing that the system remains stable at all times during learning. This is particularly challenging because the underlying system parameters are unknown. Ensuring such a notion of stability is a fundamental requirement in many safety-critical applications. Although existing literature offers algorithms and guarantees for sub-linear regret and stability with high probability, it remains unclear whether these approaches can ensure an almost surely stable controller. We believe that proposing algorithms with guarantees on the almost sure stability of systems is of significant practical importance and a critical direction for further research.

6.3.3 Concentration of Cumulative Reward in MDPs

In Chapter 3, we derived the concentration properties of cumulative reward in finite-state and finite-action MDPs. A key research direction is to explore the applications of these concentration bounds in areas such as safe, robust, and risk-averse planning and reinforcement learning. High-probability concentration bounds can provide essential safety or robustness guarantees for various algorithms, paving the way for combining learning and control in safety-critical systems and high-stakes applications.

A potential improvement to the current results is the derivation of sharper non-asymptotic upper bounds. The results in this chapter are obtained using the Azuma–Hoeffding inequality and the non-asymptotic version of the law of the iterated logarithm. By introducing additional assumptions on the underlying MDP and reward function, it may be possible to apply more refined martingale concentration bounds, leading to tighter upper-bounds.

In this chapter, we primarily focused on deriving concentration bounds within the planning setup. In the literature, non-asymptotic concentration bounds have been derived for estimation and the effects of model perturbation. An interesting research direction would be to integrate these two families of results, providing a complete picture of the concentration behavior of algorithms in reinforcement learning and adaptive control. Developing such bounds would significantly enhance the practical applicability of these methods by offering robust theoretical guarantees.

The main assumptions in this chapter are the assumptions on the underlying MDP, i.e., (i) the state and action spaces are discrete spaces, and (ii) these spaces are finite. A promising research direction is to extend our results to more general MDP frameworks with non-compact continuous state and action spaces. This type of extension is partially explored in the context of linear quadratic regulators in Chapter 5. Generalizing these results to broader MDP frameworks represents a significant and valuable challenge for future research.

6.3.4 Asymptotic Normality of Cost in LQR

In Chapter 5, we derived a central limit theorem for the cumulative cost in linear quadratic regulators by imposing specific assumptions on the noise sequence, including bounded support. A valuable research direction would be to explore relaxing these assumptions and extending the results to light-tailed noise sequences. Another promising line of research is to examine the application of the established theorem in safe or risk-averse planning and learning in Linear Quadratic Regulators.

Bibliography

[1] P. E. Caines, *Linear stochastic systems*. SIAM, 2018.

[2] L. Ljung, *System identification*. Springer, 1998.

[3] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd. Cambridge, MA: MIT Press, 2018, ISBN: 978-0-262-03924-6.

[4] N. S. Nise, *Control systems engineering*. John Wiley & Sons, 2020.

[5] T. Jaksch, R. Ortner, and P. Auer, "Near-optimal regret bounds for reinforcement learning", *Journal of Machine Learning Research*, vol. 11, no. 51, pp. 1563–1600, 2010.

[6] P. Auer and R. Ortner, "Logarithmic online regret bounds for undiscounted reinforcement learning", in *Advances in Neural Information Processing Systems*, vol. 19, MIT Press, 2006.

[7] S. Filippi, O. Cappé, and A. Garivier, "Optimism in reinforcement learning and Kullback-Leibler divergence", in *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, IEEE, 2010, pp. 115–122.

[8] P. L. Bartlett and A. Tewari, "Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps", *arXiv preprint arXiv:1205.2661*, 2012.

[9] D. Russo and B. Van Roy, "Learning to optimize via posterior sampling", *Mathematics of Operations Research*, vol. 39, no. 4, pp. 1221–1243, 2014.

[10] I. Osband, D. Russo, and B. Van Roy, "(More) efficient reinforcement learning via posterior sampling", in *Advances in Neural Information Processing Systems*, vol. 26, Curran Associates, Inc., 2013.

[11] K. Lakshmanan, R. Ortner, and D. Ryabko, "Improved regret bounds for undiscounted continuous reinforcement learning", in *International Conference on Machine Learning*, PMLR, 2015, pp. 524–532.

[12] I. Osband, B. V. Roy, and Z. Wen, "Generalization and exploration via randomized value functions", in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 48, New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 2377–2386.

[13] Y. Ouyang, M. Gagrani, A. Nayyar, and R. Jain, "Learning unknown Markov decision processes: A Thompson sampling approach", in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, Curran Associates, Inc., 2017.

[14] G. Theocharous, Z. Wen, Y. Abbasi-Yadkori, and N. Vlassis, "Posterior sampling for large scale reinforcement learning", *arXiv preprint arXiv:1711.07979*, 2017.

[15] S. Agrawal and R. Jia, "Optimistic posterior sampling for reinforcement learning: Worst-case regret bounds", in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[16] M. S. Talebi and O.-A. Maillard, "Variance-aware regret bounds for undiscounted reinforcement learning in MDPs", in *Algorithmic Learning Theory*, PMLR, 2018, pp. 770–805.

[17] R. Fruit, M. Pirotta, A. Lazaric, and R. Ortner, "Efficient bias-span-constrained exploration-exploitation in reinforcement learning", in *International Conference on Machine Learning*, PMLR, 2018, pp. 1578–1586.

[18] Z. Zhang and X. Ji, "Regret minimization for reinforcement learning by evaluating the optimal bias function", in *Advances in Neural Information Processing Systems*, vol. 32, Curran Associates, Inc., 2019.

[19] J. Qian, R. Fruit, M. Pirotta, and A. Lazaric, "Exploration bonus for regret minimization in discrete and continuous average reward MDPs", in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., 2019.

[20] R. Fruit, "Exploration-exploitation dilemma in reinforcement learning under various form of prior knowledge", Ph.D. dissertation, Université de Lille 1, Sciences et Technologies; CRIStAL UMR 9189, 2019.

[21] A. Zanette and E. Brunskill, "Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds", in *International Conference on Machine Learning*, PMLR, 2019, pp. 7304–7312.

[22] R. Fruit, M. Pirotta, and A. Lazaric, "Improved analysis of UCRL2 with empirical bernstein inequality", *arXiv preprint arXiv:2007.05456*, 2020.

[23] H. Bourel, O. Maillard, and M. S. Talebi, "Tightening exploration in upper confidence reinforcement learning", in *International Conference on Machine Learning*, PMLR, 2020, pp. 1056–1066.

[24] Z. Zhang and Q. Xie, "Sharper model-free reinforcement learning for average-reward Markov decision processes", in *The Thirty Sixth Annual Conference on Learning Theory*, PMLR, 2023, pp. 5476–5477.

[25] V. Boone and Z. Zhang, "Achieving tractable minimax optimal regret in average reward MDPs", in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[26] Y. Abbasi-Yadkori and C. Szepesvári, "Regret bounds for the adaptive control of linear quadratic systems", in *Proceedings of the 24th Annual Conference on Learning Theory*, 2011, pp. 1–26.

[27] M. Abeille and A. Lazaric, "Improved regret bounds for Thompson sampling in linear quadratic control problems", in *International Conference on Machine Learning*, PMLR, 2018, pp. 1–9.

[28] H. Mania, S. Tu, and B. Recht, "Certainty equivalence is efficient for linear quadratic control", *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[29] A. Cohen, T. Koren, and Y. Mansour, "Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret", in *International Conference on Machine Learning*, PMLR, 2019, pp. 1300–1309.

[30] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "On the sample complexity of the linear quadratic regulator", *Foundations of Computational Mathematics*, vol. 20, no. 4, pp. 633–679, 2020.

[31] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "On adaptive linear–quadratic regulators", *Automatica*, vol. 117, p. 108 982, 2020.

[32] M. Fazel, R. Ge, S. Kakade, and M. Mesbahi, "Global convergence of policy gradient methods for the linear quadratic regulator", in *International Conference on Machine Learning*, PMLR, 2018, pp. 1467–1476.

[33] Y. Ouyang, M. Gagrani, and R. Jain, "Control of unknown linear systems with Thompson sampling", in *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, IEEE, 2017, pp. 1198–1205.

[34] M. Simchowitz and D. Foster, "Naive exploration is optimal for online LQR", in *International Conference on Machine Learning*, PMLR, 2020, pp. 8937–8948.

[35] S. Lale, K. Azizzadenesheli, B. Hassibi, and A. Anandkumar, "Logarithmic regret bound in partially observable linear dynamical systems", *Advances in Neural Information Processing Systems*, vol. 33, pp. 20 876–20 888, 2020.

[36] Y. Abbasi-Yadkori, P. Bartlett, K. Bhatia, N. Lazic, C. Szepesvari, and G. Weisz, "Politex: Regret bounds for policy iteration using expert prediction", in *International Conference on Machine Learning*, PMLR, 2019, pp. 3692–3702.

[37] P. L. Bartlett and A. 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)*, AUAI Press, 2009, pp. 35–42.

[38] L. Ljung, "On the consistency of prediction error identification methods", in *Mathematics in Science and Engineering*, vol. 126, Elsevier, 1976, pp. 121–164.

[39] S. Oymak and N. Ozay, "Non-asymptotic identification of LTI systems from a single trajectory", in *2019 American Control Conference (ACC)*, IEEE, 2019, pp. 5655–5661.

[40] G. S. Deaecto, M. Souza, and J. C. Geromel, "Discrete-time switched linear systems state feedback design with application to networked control", *IEEE Transactions on Automatic Control*, vol. 60, no. 3, pp. 877–881, 2014.

[41] C. De Persis and P. Tesi, "Input-to-state stabilizing control under denial-of-service", *IEEE Transactions on Automatic Control*, vol. 60, no. 11, pp. 2930–2944, 2015.

[42] A. Cetinkaya, H. Ishii, and T. Hayakawa, "Analysis of stochastic switched systems with application to networked control under jamming attacks", *IEEE Trans. Autom. Control*, vol. 64, no. 5, pp. 2013–2028, 2018.

[43] Y. Fang, K. A. Loparo, and X. Feng, "Almost sure and δ-moment stability of jump linear systems", *International Journal of Control*, vol. 59, no. 5, pp. 1281–1307, 1994.

[44] Y. Fang, "A new general sufficient condition for almost sure stability of jump linear systems", *IEEE Transactions on Automatic Control*, vol. 42, no. 3, pp. 378–382, 1997.

[45] O. L. V. Costa, M. D. Fragoso, and R. P. Marques, *Discrete-time Markov jump linear systems*. Springer Science & Business Media, 2006.

[46] H. J. Chizeck, A. S. Willsky, and D. Castanon, "Discrete-time Markovian jump linear quadratic optimal control", *International Journal of Control*, vol. 43, no. 1, pp. 213–231, 1986.

[47] G. Goodwin, P. Ramadge, and P. Caines, "Discrete-time multivariable adaptive control", *IEEE Transactions on Automatic Control,*, vol. 25, no. 3, pp. 449–456, 1980.

[48] J. Rissanen and P. Caines, "The strong consistency of maximum likelihood estimators for ARMA processes", *Annals of Statistics*, pp. 297–315, 1979.

[49] P. E. Caines and L. Ljung, "Prediction error estimators: Asymptotic normality and accuracy", in *IEEE Conference on Decision and Control*, IEEE, 1976, pp. 652–658.

[50] B. Ho and R. E. Kálmán, "Effective construction of linear state-variable models from input/output functions", *at-Automatisierungstechnik*, vol. 14, no. 1-12, pp. 545–548, 1966.

[51] A. Lindquist and G. Picci, "State space models for Gaussian stochastic processes", in *Stochastic Systems: The Mathematics of Filtering and Identification and Applications*, M. Hazewinkel and J. C. Willems, Eds., Dordrecht: Springer Netherlands, 1981, pp. 169–204, ISBN: 978-94-009-8546-9.

[52] T. L. Lai and C. Z. Wei, "Asymptotic properties of multivariate weighted sums with applications to stochastic regression in linear dynamic systems", *Multivariate Analysis VI*, pp. 375–393, 1985.

[53] G. Ferrari-Trecate, M. Muselli, D. Liberati, and M. Morari, "A clustering technique for the identification of piecewise affine systems", *Automatica*, vol. 39, no. 2, pp. 205–217, 2003.

[54] F. Lauer and G. Bloch, "A new hybrid system identification algorithm with automatic tuning", *IFAC Proceedings Volumes*, vol. 41, no. 2, pp. 10 207–10 212, 2008.

[55] R. Vidal, "Recursive identification of switched arx systems", *Automatica*, vol. 44, no. 9, pp. 2274–2287, 2008.

[56] J. Roll, A. Bemporad, and L. Ljung, "Identification of piecewise affine systems via mixed-integer programming", *Automatica*, vol. 40, no. 1, pp. 37–50, 2004.

[57] N. Ozay, C. Lagoa, and M. Sznaier, "Set membership identification of switched linear systems with known number of subsystems", *Automatica*, vol. 51, pp. 180–191, 2015.

[58] R. Vidal, S. Soatto, Y. Ma, and S. Sastry, "An algebraic geometric approach to the identification of a class of linear hybrid systems", in *42nd IEEE International Conference on Decision and Control (IEEE Cat. No. 03CH37475)*, IEEE, vol. 1, 2003, pp. 167–172.

[59] A. L. Juloski, S. Weiland, and W. M. H. Heemels, "A bayesian approach to identification of hybrid systems", *IEEE Transactions on Automatic Control*, vol. 50, no. 10, pp. 1520–1533, 2005.

[60] A. Bemporad, A. Garulli, S. Paoletti, and A. Vicino, "A bounded-error approach to piecewise affine system identification", *IEEE Transactions on Automatic Control*, vol. 50, no. 10, pp. 1567–1580, 2005.

[61] F. Lauer and G. Bloch, "Switched and piecewise nonlinear hybrid system identification", in *International workshop on hybrid systems: computation and control*, Springer, 2008, pp. 330–343.

[62] N. Ozay, C. Lagoa, and M. Sznaier, "Robust identification of switched affine systems via moments-based convex optimization", in *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, IEEE, 2009, pp. 4686–4691.

[63] F. Lauer, G. Bloch, and R. Vidal, "A continuous optimization framework for hybrid system identification", *Automatica*, vol. 47, no. 3, pp. 608–613, 2011.

[64] F. Lauer, G. Bloch, F. Lauer, and G. Bloch, "Hybrid system identification", *Hybrid System Identification: Theory and Algorithms for Learning Switching Models*, pp. 77–101, 2019.

[65] P. E. Caines and H.-F. Chen, "Optimal adaptive LQG control for systems with finite state process parameters", *IEEE Transactions on Automatic Control*, vol. 30, no. 2, pp. 185–189, 1985.

[66] P. E. Caines and J.-F. Zhang, "On the adaptive control of jump parameter systems via nonlinear filtering", *SIAM Journal on Control and Optimization*, vol. 33, no. 6, pp. 1758–1777, 1995.

[67] F. Xue and L. Guo, "Necessary and sufficient conditions for adaptive stablizability of jump linear systems", *Communications in Information and Systems*, vol. 1, no. 2, pp. 205–224, 2001.

[68] B. Sayedana, M. Afshari, P. E. Caines, and A. Mahajan, "Thompson-sampling based reinforcement learning for networked control of unknown linear systems", in *2022 IEEE 61st Conference on Decision and Control (CDC)*, 2022, pp. 723–730. DOI: 10.1109/CDC51059.2022.9992565.

[69] S. Shi, O. Mazhar, and B. De Schutter, "Finite-sample analysis of identification of switched linear systems with arbitrary or restricted switching", *IEEE Control Systems Letters*, vol. 7, pp. 121–126, 2023. DOI: 10.1109/LCSYS.2022.3187511.

[70] T. Sarkar, A. Rakhlin, and M. Dahleh, "Nonparametric system identification of stochastic switched linear systems", in *2019 IEEE 58th Conference on Decision and Control (CDC)*, IEEE, 2019, pp. 3623–3628.

[71] Y. Sattar, Z. Du, D. A. Tarzanagh, L. Balzano, N. Ozay, and S. Oymak, "Identification and adaptive control of Markov jump systems: Sample complexity and regret bounds", *arXiv preprint arXiv:2111.07018*, 2021.

[72] Y. Sattar, S. Oymak, and N. Ozay, "Finite sample identification of bilinear dynamical systems", in *2022 IEEE 61st Conference on Decision and Control (CDC)*, IEEE, 2022, pp. 6705–6711.

[73] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Optimism-based adaptive regulation of linear-quadratic systems", *IEEE Trans. Autom. Control*, vol. 66, no. 4, pp. 1802–1808, 2020.

[74] Z. Du, Y. Sattar, D. A. Tarzanagh, L. Balzano, S. Oymak, and N. Ozay, "Certainty equivalent quadratic control for Markov jump systems", *arXiv preprint arXiv:2105.12358*, 2021.

[75] J. P. Jansch-Porto, B. Hu, and G. E. Dullerud, "Policy optimization for Markovian jump linear quadratic control: Gradient method and global convergence", *IEEE Transactions on Automatic Control*, vol. 68, no. 4, pp. 2475–2482, 2022.

[76] A. Ruszczyński, "Risk-averse dynamic programming for Markov decision processes", *Mathematical Programming*, vol. 125, pp. 235–261, 2010.

[77] F. J. Beutler and K. W. Ross, "Optimal policies for controlled markov chains with a constraint", *Journal of Mathematical Analysis and Applications*, vol. 112, no. 1, pp. 236–252, 1985.

[78] E. Altman, *Constrained Markov Decision Processes*. New York: Routledge, 2021, ISBN: 978-1-351-45082-3. DOI: 10.1201/9781315140223.

[79] M. G. Bellemare, W. Dabney, and R. Munos, "A distributional perspective on reinforcement learning", in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 70, PMLR, Jun. 2017, pp. 449–458.

[80] M. G. Bellemare, W. Dabney, and M. Rowland, *Distributional Reinforcement Learning*. MIT Press, 2023, http://www.distributional-rl.org.

[81] M. J. Sobel, "The variance of discounted Markov decision processes", *Journal of Applied Probability*, vol. 19, no. 4, pp. 794–802, 1982.

[82] M. Duflo, *Random Iterative Models* (Applications of Mathematics). Berlin, Heidelberg: Springer Science & Business Media, 2013, vol. 34, ISBN: 978-3-662-03203-0.

[83] S. P. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability* (Cambridge Mathematical Library). Cambridge: Springer Science & Business Media, 2012, ISBN: 978-1-4612-4244-9.

[84] O. Hernández-Lerma and J. B. Lasserre, *Further Topics on Discrete-Time Markov Control Processes* (Applications of Mathematics). Berlin, Heidelberg: Springer Science & Business Media, 2012, vol. 42, ISBN: 978-1-4612-7067-1.

[85] R. A. Fisher, "On the mathematical foundations of theoretical statistics", *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, vol. 222, no. 594-604, pp. 309–368, 1922.

[86] H. Cramér, *Mathematical methods of statistics*. Princeton university press, 1999, vol. 26.

[87] P. J. Huber, "The behavior of maximum likelihood estimates under nonstandard conditions", in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Berkeley, CA: University of California Press, vol. 1, 1967, pp. 221–233.

[88] F. Eicker, "Asymptotic normality and consistency of the least squares estimators for families of linear regressions", *The Annals of Mathematical Statistics*, vol. 34, no. 2, pp. 447–456, 1963.

[89] T. L. Lai and C. Z. Wei, "Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems", *Annals of Statistics*, vol. 10, no. 1, pp. 154–166, 1982.

[90] L. Lennart and P. E. Caines, "Asymptotic normality of prediction error estimators for approximate system models", *Stochastics*, vol. 3, no. 1-4, pp. 29–46, 1980.

[91] L. Ljung and P. E. Caines, "Asymptotic normality of prediction error estimators for approximate system models", in *1978 IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes*, 1978, pp. 927–932. DOI: 10.1109/CDC.1978.268066.

[92] T. W. Anderson and N. Kunitomo, "Asymptotic distributions of regression and autoregression coefficients with martingale difference disturbances", *Journal of Multivariate Analysis*, vol. 40, no. 2, pp. 221–243, 1992.

[93] S. Borovkova, H. P. Lopuhaä, and B. N. Ruchjana, "Consistency and asymptotic normality of least squares estimators in generalized star models", *Statistica Neerlandica*, vol. 62, no. 4, pp. 482–508, 2008.

[94] V. Fabian, "On asymptotic normality in stochastic approximation", *The Annals of Mathematical Statistics*, pp. 1327–1332, 1968.

[95] J. Sacks, "Asymptotic distribution of stochastic approximation procedures", *The Annals of Mathematical Statistics*, vol. 29, no. 2, pp. 373–405, 1958.

[96] P. H. Algoet, "The strong law of large numbers for sequential decisions under uncertainty", *IEEE Transactions on Information Theory*, vol. 40, no. 3, pp. 609–633, 1994.

[97] R. van Handel, "Ergodicity, decisions, and partial information", *Séminaire de Probabilités XLVI*, pp. 411–459, 2014.

[98] B. Hajek, "Ergodic process selection", in *Open Problems in Communication and Computation*, Springer, 1987, pp. 199–203.

[99] M. Duflo, *Random Iterative Models.* Berlin-Heidelberg: Springer, 1997.

[100] N. Maigret, "Théorème de limite centrale fonctionnel pour une chaine de Markov récurrente au sens de Harris et positive", in *Annales de l'institut Henri Poincaré. Section B. Calcul des probabilités et statistiques*, vol. 14, 1978, pp. 425–440.

[101] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "Regret bounds for robust adaptive control of the linear quadratic regulator", *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[102] A. Cassel, A. Cohen, and T. Koren, "Logarithmic regret for learning linear quadratic regulators efficiently", in *International Conference on Machine Learning*, PMLR, 2020, pp. 1328–1337.

[103] F. Wang and L. Janson, "Exact asymptotics for linear quadratic adaptive control", *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 12 136–12 247, 2021.

[104] Y. Lu and Y. Mo, "Almost surely $\sqrt{T}$ regret bound for adaptive LQR", *arXiv preprint arXiv:2301.05537*, 2023.

[105] B. Sayedana, M. Afshari, P. E. Caines, and A. Mahajan, "Strong consistency and rate of convergence of switched least squares system identification for autonomous Markov jump linear systems", *IEEE Transactions on Automatic Control*, vol. 69, no. 6, pp. 3952–3959, 2024. DOI: 10.1109/TAC.2024.3351806.

[106] B. Sayedana, P. E. Caines, and A. Mahajan, "Asymptotic normality of cumulative cost in linear quadratic regulators", in *2024 IEEE 63rd Conference on Decision and Control (CDC)*, 2024, pp. 1856–1862. DOI: 10.1109/CDC56724.2024.10886506.

[107] B. Sayedana, M. Afshari, P. E. Caines, and A. Mahajan, "Relative almost sure regret bounds for certainty equivalence control of Markov jump systems", in *2023 IEEE 62nd Conference on Decision and Control (CDC)*, 2023, pp. 6629–6634. DOI: 10.1109/CDC49753.2023.10383246.

[108] B. Sayedana, M. Afshari, P. E. Caines, and A. Mahajan, "Consistency and rate of convergence of switched least squares system identification for autonomous Markov jump linear systems", in *2022 IEEE 61st Conference on Decision and Control (CDC)*, 2022, pp. 6678–6685. DOI: 10.1109/CDC51059.2022.9993169.

[109] H.-F. Chen and L. Guo, "Convergence rate of least-squares identification and adaptive control for stochastic systems", *International Journal of Control*, vol. 44, no. 5, pp. 1459–1476, 1986.

[110] H.-F. Chen and L. Guo, "Optimal adaptive control and consistent parameter estimates for armax model with quadratic cost", *SIAM Journal on Control and Optimization*, vol. 25, no. 4, pp. 845–867, 1987.

[111] T. E. Duncan and B. Pasik-Duncan, "Adaptive control of continuous-time linear stochastic systems", *Mathematics of Control, signals and systems*, vol. 3, no. 1, pp. 45–60, 1990.

[112] A. Zhang and M. Wang, "Spectral state compression of Markov processes", *IEEE Transactions on Information Theory*, vol. 66, no. 5, pp. 3202–3231, 2019.

[113] W. F. Stout, *Almost Sure Convergence.* Academic Press, 1974.

[114] P. Brémaud, *Markov chains: Gibbs fields, Monte Carlo simulation, and queues.* Springer Science & Business Media, 2013, vol. 31.

[115] Y. Sattar and S. Oymak, "Non-asymptotic and accurate learning of nonlinear dynamical systems", *arXiv preprint arXiv:2002.08538*, 2020.

[116] A. Czornik, *On control problems for jump linear systems.* Wydawn. Politechniki Śląskiej, 2003.

[117] K. J. Åström, *Introduction to Stochastic Control Theory.* Dover, 1970.

[118] B. Sayedana, P. E. Caines, and A. Mahajan, "Concentration of cumulative reward in Markov decision processes", *arXiv preprint arXiv:2411.18551*, 2024.

[119] D. P. Bertsekas, *Dynamic Programming and Optimal Control, Volume II*, 4th. Belmont, MA: Athena Scientific, 2012, ISBN: 978-1-886529-44-1.

[120] L. Kallenberg, "Classification problems in MDPs", in *Markov Processes and Controlled Markov Chains*, Boston, MA: Springer, 2002, pp. 151–165.

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

[122] D. P. Bertsekas, *Dynamic Programming and Optimal Control, Volume I*, 4th. Belmont, MA: Athena Scientific, 2012, ISBN: 978-1-886529-44-1.

[123] P. Billingsley, *Convergence of Probability Measures* (Wiley Series in Probability and Statistics), 2nd. Hoboken, NJ: John Wiley & Sons, 2013, ISBN: 978-1-118-12237-2.

[124] J. Neveu, *Discrete-Parameter Martingales* (North-Holland Mathematical Library), trans. by T. Speed. Amsterdam: North-Holland, 1975, vol. 10, ISBN: 978-0-7204-2830-5.

[125] M. Raginsky and I. Sason, *Concentration of Measure Inequalities in Information Theory, Communications, and Coding* (Foundations and Trends in Communications and Information Theory). Now Publishers Inc., 2014, vol. 10, pp. 1–246, ISBN: 978-1-60198-839-5.

[126] A. Balsubramani, "Sharp finite-time iterated-logarithm martingale concentration", *arXiv preprint arXiv:1405.2639*, 2014.

[127] R. B. Ash, B. Robert, C. A. Doleans-Dade, and A. Catherine, *Probability and measure theory*. Academic press, 2000.

[128] P. Billingsley, *Probability and measure*. John Wiley & Sons, 2017.

[129] K. J. Åström, *Introduction to stochastic control theory*. Courier Corporation, 2012.