1-1-2012

Large deviations of stochastic systems and applications

Qi He
Wayne State University,

LARGE DEVIATIONS OF STOCHASTIC SYSTEMS AND APPLICATIONS

by

QI HE

DISSERTATION

Submitted to the Graduate School,

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2012

MAJOR: MATHEMATICS

Approved by:

Advisor	Date

DEDICATION

To my parents

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincerest appreciation and deepest gratitude to my advisor, professor George Yin, for his endless care, constant support and inspiring guidance during the five-year pursuit of my PhD degree.

Next, I am taking this opportunity to thank Professor Tze Chien Sun, Professor Peiyong Wang, and Professor Leyi Wang for serving in my committee.

I am grateful to Professor Xianping Guo, who recommended me to Wayne State University and helped me a lot when I worked in China.

I owe many thanks to Professor Rafail Khasminskii, Professor Paul-Liu Chow, Professor Alex Korostelev and Professor Qing Zhang, who have contributed to my solid knowledge in Probability and Stochastic Processes Theory.

I would like to thank my parents for their unconditional and unlimited love and support.

Finally, I could not close this section before I express my appreciation to the entire Department of Mathematics for their hospitality, services and supports during my study at Wayne State University.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1: Introduction

My dissertation focuses on large deviations of stochastic systems with applications to optimal control and system identification. It encompasses analysis of two-time-scale Markov processes and system identification.

Applications of Markovian models have emerged from manufacturing systems, wireless communications, internet traffic modeling, and financial engineering in recent years. Many such problems involve large-scale systems. An effective way of modeling and computation is to use a two-time-scale formulation. Previous results in the literature show that under suitable conditions, one can obtain a limit system. Although the limit is much simpler to deal with, the results cannot provide a decisive answer to certain probabilistic error bound in the normal deviation range. The large deviations principles can provide very precise estimates, but they are mathematically very demanding. Obtaining the desired bounds turns out to be very challenging that requires delicate and detailed estimates. In addition, all of the existing results to date dealt with homogeneous Markov processes, whereas in this dissertation, we consider large deviations of systems driven by nonhomogeneous Markov processes. Our results help us to substantially reduce the computational effort providing a systematic approach for reducing a large-scale system to a system with much smaller dimension.

Traditional system identification concentrates on convergence and convergence rates of estimates in mean squares, or in distribution, or in a strong sense. For system diagnosis and complexity analysis, however, it is essential to understand the probability of identification errors over a finite data window. This dissertation investigates identification errors in a large deviations framework. By considering both space complexity in terms of quantization and time complexity with respect to data window sizes, our study provides a new perspective to understand the fundamental relationship between probabilistic errors and resources that represent data sizes in computer algorithms, sample sizes in statistical analysis, channel bandwidths in communications, etc. This relationship is derived by establishing the large deviations principle for quantized identification that links binary-valued data at one end and regular sensors at the other.

When the deviation is beyond the normal range but is not as large as in the "large deviation regime," we are in the so-called moderate deviations range. Some applications, for example observer design of involving stochastic systems need such estimates. Similar problems also arise in optimal control of stochastic systems. The required mathematical techniques are different from the large deviations estimates. Our current and future efforts are devoted to this topic.

The rest of the dissertation is arranged as follows. Chapter 2 give a brief introduction to large deviations principle. Chapter 3 discusses large deviations of two-time-scale Markovian switching system. Chapter 4 focuses on large deviations of identification systems with its applications. A few further remarks are made in Chapter 5.

Chapter 2: Preliminary of Large Deviations Principle

The theory of large deviations is concerned with the study of the estimates on probabilistic and moments that are associated with "rare" events. For example, the description of events where a sum of random variables deviates from its mean by more than a "normal" amount, which beyond the central limit theorem. It has applications in probability theory, statistics, operation research, communication networks, information theory, statistical physics, financial mathematics and queuing systems. The first rigorous large deviations results were obtained by Harald Cramer in the late 1930s, who applied them to model the insurance business. He gave a large deviation principle (LDP) result for a sum of i.i.d. random variables, where the rate function is expressed by a power series. Then S.R.S. Varadhan developed a general framework for the LDP in 1966. In the following, the large deviation studied by several people, including the work of Schilder (LDP for Brownian motion), Sanov (LDP for ergodic processes), and Freidlin and Wentzell (LDP for diffusions) with some abstract foundation of LDP. A significant step forward was achieved through a series of papers of Donsker and Varadhan, starting in the mid 1970s in which they developed a systematic large deviation theory for empirical measures in the i.i.d. and Markov processes, whose contribution is emphasized by recent award of the Abel prize.

Consider one simple example. Let $X_1, \dots X_n$ be i.i.d random variable with finite mean $E[X_1] = \mu$. Then the law of large number states that the sample mean approaches the true mean, if the sample size n goes to infinity. Define $S_n = \frac{\sum_{k=1}^{n} X_k}{n}$. Then by the law of large number,

$$S_n \to \mu, w.p.1.$$

It is of interest to obtain the rate of convergence. To make this precisely, let us fix a number $a > \mu$. By the law of large number, we know that $P(S_n > a) \to 0$, as n goes to infinity. In large deviations theory, we are interested in the rate at which the probablity $P(S_n > a)$ decay to 0.

To demonstrate the concepts, we define the following functions.

$$\begin{aligned} \varphi(t) &= E[e^{X_1 t}] \\ I(\beta) &= \sup_t [t\beta - \log \varphi(t)] \end{aligned} \tag{2.1}$$

In the above, we see that $\varphi(t)$ is the moment generating function. We assume that $M(t)$ exists in a neighborhood of 0. The function $I(\beta)$ is referred to as a Legendre-Fenchel transform and is also called rate function. Note that $I(\beta)$ is always non-negative.

Then we formulate the first basic result of large deviations theory, which goes back to Cramer (1938). This result identifies the large deviation behavior of the empirical average $\frac{1}{n}S_n$.

Theorem 2.1. *Let $\{X_n\}$ be i.i.d $\mathbb{R}$ valued random variable satisfying*

$$\varphi(t) = Ee^{tX_1} < \infty \quad \forall t \in \mathbb{R}$$

Then, for all $a > EX_1$,

$$\lim_{n\to\infty} \frac{1}{n} \log P(S_n \leq an) = -I(a).$$

For more general definition of large deviations principle. Let $\{X^\epsilon, \epsilon > 0\}$ be a collection of random variable defined on a Polish space (i.e., complete separable metric space) $(\Omega, \mathcal{F}, P)$ and taking values in a Polish space $\mathcal{E}$. Denote the metric on $\mathbb{E}$ as $d(\cdot,\cdot)$ and expectation with respect to P by $\mathbb{E}$. The theory of large deviations focuses on random variables $\{X^\epsilon\}$ for which the probabilities $P(X^\epsilon \in A)$ converge to 0 exponentially fast for a class of Borel sets A. The exponential decay rate of these probabilities is expressed in terms of a function I mapping $\mathcal{E}$ into $[0,\infty]$. This function is called a rate function if it has compact level sets, i.e, for each $M<\infty$ the level sets $\{x\in\mathcal{E} : I(x)\leq M\}$ is a compact subset in $\mathcal{E}$.

Definition 2.2. *(Large deviation principle) Let I be a rate function on $\mathcal{E}$. The sequence $\{X^\epsilon\}$ is said to satisfy the large deviations principle on $\mathcal{E}$, as $\epsilon\to 0$, with rate function I if for any Borel set B in $\mathcal{E}$,*

$$\begin{aligned} -\inf_{\beta\in B^\circ} I(\beta) &\leq \liminf_{\epsilon\to 0} \epsilon \log P\{X^\epsilon \in B\} \\ &\leq \limsup_{\epsilon\to 0} \epsilon \log P\{X^\epsilon \in B\} \\ &\leq -\inf_{\beta\in \overline{B}} I(\beta), \end{aligned} \tag{2.2}$$

where B° and $\overline{B}$ denote the interior of and closure of B.

Remark 2.3. *Instead of* (2.2), *we can say that*

1 (Large deviations upper bound.) For each closed subset F of $\mathcal{E}$

$$\limsup_{\epsilon\to 0} \epsilon \log P(X^\epsilon \in F) \leq -\inf_{x\in F} I(x).$$

2 (Large deviations lower bound.) For each open subset G of $\mathcal{E}$

$$\liminf_{\epsilon\to 0} \epsilon \log P(X^\epsilon \in G) \geq -\inf_{x\in G} I(x).$$

Next we introduce a couple of preliminary results on the LDP which we use often in the following chapters. Lemma 2.4 is a version of the Gärtner-Ellis Theorem (see [27, Lemma 1]), which states the LDP in $\mathbb{R}^k$ space. Here and hereafter, we use $\langle a,b\rangle$ to denote the usual inner product in $\mathbb{R}^k$ for $a,b\in\mathbb{R}^k$.

Lemma 2.4. *(Gärtner-Ellis Theorem) Let $\{X_n\}$ denote a sequence of $\mathbb{R}^k$-valued random vectors, for which the following limit exists*

$$H(\tau) = \lim_{n\to\infty} \frac{1}{n} \log E \exp\{n\langle \tau, X_n\rangle\},$$

where $\tau \in \mathbb{R}^k$ and $H(\cdot)$ is continuously differentiable. Define the dual function $I(\beta) = \sup_{\tau\in\mathbb{R}^k}[\langle \tau, \beta\rangle - H(\tau)]$. Then $\{X_n\}$ satisfies large deviations principle with rate function $I(\cdot)$.

Next we state the contraction principle , which states that the LDP is preserved by a continuous mapping, see [9, Theorem 4.2.1, p.126]. In fact, the results in [9] are more general and can be applied to mappings between Hausdorff topological spaces. But assertion (b) is sufficient for this brief.

Lemma 2.5. *(Contraction Principle) Let $f : \mathbb{R}^r \to \mathbb{R}^k$ be a continuous function. Assume that a family of random variable $\{X_n\}$ on $\mathbb{R}^r$ satisfies the LDP with rate function $I : \mathbb{R}^r \mapsto [0, +\infty]$. Then the family of random variable defined by $\{Y_n\}, Y_n = f(X_n)$ satisfies the LDP with rate function $\widetilde{I}(y) = \inf\{I(x) : x \in \mathbb{R}^r, y = f(x)\}$.*

Proof. The proofs of (a) can be found in [27, Lemma 1] and the proof of (b) is in [9, Theorem 4.2.1, p.126]. □

To understand large deviations principle, we give some examples in the following.

Example 2.6. *Consider a sequence of i.i.d random variable $\{X_n\}$, which follows standard normal distribution. By law of large number*

$$S_n = \frac{1}{n} \sum_{i=1}^{n} X_i \to 0.$$

Given any number $0 < a$, *the probability* $P(S_n > a) \to 0$. *Then it can be shown that*

$$\lim_{n\to\infty} \frac{1}{n} \log P(S_n > a) = -I(a),$$

where $I(a) = \dfrac{a^2}{2}$. *In fact, the rate function can be calculated directly by Gätner-Ellis theorem.*

Example 2.7. *(Schilder theorem) We consider a rescaled random process*

$$X^\epsilon(t) = \epsilon W(t), [0, T]$$

in $\mathbb{R}^n$. *Here* $W(t)$ *is a Wiener process in* $\mathbb{R}^n$. *As* $\epsilon \to 0$, *the trajectory* $X^\epsilon(t)$ *converges in probability to the solution* 0. *on every finite time interval. In the path space, the probability distribution of* $X^\epsilon(t)$ *is nearly degenerate at the path* $f(\cdot) \equiv 0$. *Any other event that does not include* 0 *and its neighborhood has very small probability. In fact, on the Banach space* $C_0 = C_0([0, T]; \mathbb{R}^n)$ *of continuous functions equipped with the supremum norm* $||\cdot||_\infty$, *the process satisfy the large deviations principle with rate function* $I : C_0 \mapsto [0, \infty]$ *given by*

$$I(\varphi) = \begin{cases} \dfrac{1}{2}\displaystyle\int_0^T \dot{\varphi}(s)ds, & \text{if } \varphi \in C_{0,T}(\mathbb{R}^n) \text{ is absolutely continuous,} \\ \infty & \text{otherwise,} \end{cases}$$

In other words, for every open set $G \subseteq C_0$ *and every closed set* $F \subseteq C_0$,

$$\limsup_{\epsilon\to 0} \epsilon \log P(X^\epsilon(\cdot) \in G) \leq -\inf_{\varphi\in G} I(\varphi),$$

and

$$\liminf_{\epsilon\to 0} \epsilon \log P(X^\epsilon(\cdot) \in F) \geq -\inf_{\varphi\in F} I(\varphi).$$

Chapter 3: LDP of Two-time-scale Markovian Switching System

3.1 Introduction and Motivation

Randomly varying discrete events exist in many problems arising in manufacturing systems, production planning, queueing networks, Monte Carlo simulation, and random environment. These discrete events are often modeled by pure jump processes with the use of Markovian models. Applications of Markovian models have emerged from wireless communications, internet traffic modeling, and financial engineering in recent years. The rapid progress in technology has opened up new domains and provided greater opportunities for further exploration.

To make the computation affordable and feasible, one often has to contend with finding approximate solutions. This is particularly true for many control and optimization problems. An effective modeling and computational step is to use a two-time-scale formulation. Time-scale separation is often inherent in the underlying problems, for instance, equity investors in a stock market can be classified as belonging to two categories, long-term investors and short-term investors. The long-term investors consider a relatively longtime horizon and make decisions based on weekly or monthly performance of the stock, whereas the short term investors (such as day traders) focus on returns in short-term, daily or an even shorter period. Their time scales are in sharp contrast. An effective way to delineate the distinct rates of changes is to introduce a small parameter $\epsilon > 0$ into the system. Note that ϵ is only used to separate different time scales so that we can provide asymptotic analysis for small ϵ. In the literature, Simon and Ando [33] used such an idea and introduced the so-called hierarchical decomposition and aggregation. Sethi and Zhang [32] initiated the study of nearly optimal controls for flexible manufacturing systems. To further investigate the underlying properties, Khamsminskii, Yin, and Zhang developed asymptotic expansions for the probability distribution vectors [23,24] using an analytic approach. Subsequently, a more comprehensive study was launched in [46], which contains scaled sequence of occupation measures, switching diffusion limits, near-optimal controls of Markovian systems, Markov decision processes, and numerical methods, among others.

In this dissertation, we furthered our investigation started in [46–48]. We aim at establishing large deviations principles for singularly perturbed systems involving rapidly fluctuating Markov chains. Let $\alpha^\epsilon(t)$ be a Markov chain with a finite space $\mathcal{M} = \{1, \ldots, m_0\}$ generated by $Q(t)/\epsilon$, where $Q(t) \in \mathbb{R}^{m_0 \times m_0}$ is a generator and ϵ is a small parameter. Recall that (see [23]) a Markov chain or its generator $Q(t)$ is irreducible, if the system of equations

$$\begin{cases} \nu(t)Q(t) = 0, \\ \displaystyle\sum_{i=1}^{m_0} \nu_i(t) = 1 \end{cases} \tag{3.1}$$

has a unique solution such that $\nu_i(t) > 0$ for each $i \in \mathcal{M}$. The unique solution of (3.1), namely, the row vector $\nu(t) = (\nu_1(t), \nu_2(t), \ldots, \nu_{m_0}(t))$ is termed a quasi-stationary distribution. When $Q(t) = Q$ independent of t, $\nu(t) = \nu$ becomes the stationary distribution. Throughout this dissertation, we assume that $Q(t)$ is irreducible for each $t \in [0, T]$.

In fact, many problems arising in manufacturing, production planning, and networked systems can be formulated as one that is driven by a two-time-scale Markovian chain, in which the state space $\mathcal{M}$ is large. As a motivational example, consider the following example.

Example 3.1. This example is motivated by the linear quadratic Gaussian (LQG) regulator problem considered in [49]; see also [5] and references therein. Let $\alpha^\epsilon(t)$ be a continuous-time Markov chain with state space $\mathcal{M}$ and generator $Q^\epsilon(t) = Q(t)/\epsilon$, where $Q(t)$ is irreducible. Consider

$$\begin{aligned} &\dot{x}^\epsilon(t) = [A(\alpha^\epsilon(t))x^\epsilon(t) + B(\alpha^\epsilon(t))u(t)], \\ &x^\epsilon(s) = x, \text{ for } s \le t \le T, \end{aligned} \tag{3.2}$$

where $x(t) \in \mathbb{R}^{n_1}$ is the state, $u(t) \in \mathbb{R}^{n_2}$ is the control, $A(i) \in \mathbb{R}^{n_1 \times n_1}$ and $B(i) \in \mathbb{R}^{n_1 \times n_2}$ are well defined and have finite values for each $i \in \mathcal{M}$. The objective is to find the optimal control $u(\cdot)$ so that the expected quadratic cost function

$$\begin{aligned} J(s,i,x,u(\cdot)) = E\bigg\{ \int_s^T & [x'(t)M(\alpha^\epsilon(t))x(t) \\ & +u'(t)N(\alpha^\epsilon(t))u(t)]dt + x'(T)Dx(T)\bigg\} \end{aligned} \tag{3.3}$$

is minimized, where E is the expectation given $\alpha^\epsilon(s) = \alpha$ and $x(s) = x$, $M(i)$, $i = 1, \ldots, m_0$, are symmetric nonnegative definite matrices, and $N(i)$, $i = 1, \ldots, m_0$, and D are symmetric positive definite matrices. Let $v^\epsilon(s,i,x) = \inf_{u(\cdot)} J^\epsilon(s,i,x,u(\cdot))$ be the value function. Then v^ϵ satisfies the following system of HJB equations: for $0 \le s \le T$ and $i \in \mathcal{M}$,

$$\begin{aligned} 0 = \frac{\partial v^\epsilon(s,i,x)}{\partial s} &+ \min_u \bigg\{ (A(i)x + B(i)u)' \frac{\partial v^\epsilon(s,i,x)}{\partial x} \\ &+x'M(i)x + u'N(i)u + Q^\epsilon(s)v^\epsilon(s,\cdot,x)(i)\bigg\}, \end{aligned} \tag{3.4}$$

with the boundary condition $v^\epsilon(T,i,x) = x'Dx$, where

$$Q^\epsilon(s)v^\epsilon(s,\cdot,x)(i) = \sum_{j \ne i} q^\epsilon_{ij}(s)(v^\epsilon(s,j,x) - v^\epsilon(s,i,x)).$$

Following the approach in [16] and [49], we write

$$v^\epsilon(s,i,x) = x'K^\epsilon(s,i)x + q^\epsilon(s,i), \tag{3.5}$$

for some $n_1 \times n_1$ matrix K^ϵ and a scalar function q^ϵ. Without loss of generality, we may assume K^ϵ to be symmetric. This yields the following system of Riccati equations for $K^\epsilon(s,i)$,

$$\begin{aligned}\dot{K}^\epsilon(s,i) = -K^\epsilon(s,i)A(i) - A'(i)K^\epsilon(s,i) - M(i) \\ +K^\epsilon(s,i)B(i)N^{-1}(i)B'(i)K^\epsilon(s,i) - Q^\epsilon(s)K^\epsilon(s,\cdot)(i),\end{aligned} \tag{3.6}$$

with $K^\epsilon(T,i) = D$, and the equations for q^ϵ,

$$\dot{q}^\epsilon(s,i) = -Q^\epsilon(s)q^\epsilon(s,\cdot)(i), \tag{3.7}$$

with $q^\epsilon(T,i) = 0$. Similar to [49], it can be shown that the optimal control $u^{\epsilon,*}$ has the form:

$$u^{\epsilon,*}(s,i,x) = -N^{-1}(i)B'(i)K^\epsilon(s,i)x. \tag{3.8}$$

To get the optimal control, one must solve the system of corresponding Riccati equations. The large dimensionality makes the computation difficult. It would be much better if we can treat a "reduced" system with much less effort. Using the weak convergence methods, $x^\epsilon(\cdot)$ converges weakly to $x(\cdot)$. Based on the averaged dynamic system, one can construct controls leading to near optimality. As ϵ is getting smaller and smaller, we expect that $x^\epsilon(t)$ will hover about $x(t)$. One immediate question is: What can we say about $P(|x^\epsilon(t) - x(t)| \geq a)$ for $a > 0$? Although our previous results delineate the limit diffusion, they cannot provide a decisive answer to the question above. We must resort to large deviations principle for a precise estimate.

Our study will be closely related to singularly perturbed Markov chains. To see this, let us begin with the following scenario. For a positive integer k, let $f : \mathcal{M} \to \mathbb{R}^k$ be a well-defined function, and consider the process $n^\epsilon(t) = \int_0^t f(\alpha^\epsilon(s))ds$. Assuming either $Q(\cdot)$ is smooth or measurable, it can be shown (see [46, p. 84 or p. 209]) that $n^\epsilon(t) \to \overline{n}(t) =$

$\int_0^t \sum_{i=1}^m f(i)\nu_i(s)ds$ in probability as $\epsilon \to 0$. The above is essentially a law of large numbers type result. When $Q(t) = Q$ is a constant matrix, the Markov chain $\alpha^\epsilon(\cdot)$ is homogeneous. The irreducibility of Q implies that the process is mixing with exponential mixing rate. The mixing property then implies the convergence is in the sense of almost surely. In addition, we can also consider $[n^\epsilon(t) - \overline{n}(t)]/\sqrt{\epsilon}$. We can prove that the scaled sequence converges weakly to a diffusion process.

Consider the difference $n^\epsilon(t) - \overline{n}(t)$. We ask ourself the question: What can we say about $P(|n^\epsilon(t) - \overline{n}(t)| > a)$ for some $a > 0$? Now it is clear that neither the law of large numbers nor the central limit theorem provides us with the desired answers. The law of large numbers and the central limit theorem only tell us that $\{|n^\epsilon(t) - \overline{n}(t)| > a\}$ is a event with small probability, but does not quantify how small it is. Concerning such rates of convergence problems, first, we consider $n^\epsilon(T)$ for a fixed T, which turns out to be a generalized occupation measure of the Markov chain.

Large deviations of occupation measures for Markov processes have been studied by many researchers. In the literature, Donsker and Varadan proved the large deviations principle (LDP) under the assumption of existence, continuity, and strictly positive transition density, together with exponential tightness [12]. Their result was improved by Deuschel and Strook [10] by applying dominating measure method. Subsequently, Ney and Nummelin [31], de Acosta [1, 2] and Jain [21] extended the results to general irreducible Markov processes, and obtained lower bounds of LDP in its full generality. In their case, the rate function may be different from the classical one given in [12], as shown in the works of Dinwoodie [11] and Dupuis and Zeitouni [14]. All the above papers dealt with homogeneous Markov processes, whereas we consider large deviations of systems driven by nonhomogeneous Markov chains.

If we take a dynamic system point of view, we may consider the convergence of the process $n^\epsilon(\cdot)$ on the interval $[0,T]$. In this regard, it is useful to use the tool of the action functional developed by Freidlin and Wentzell [17]. Here, the limit of $n^\epsilon(t)$ is the average of f with respect to the invariant measure ν. In [17], homogeneous Markov processes were considered, and in [13], Dupuis and Kushner considered large deviations of nonhomogeneous systems by using averaging principles, in which the systems are perturbed by Gaussian process.

Inspired by many existing works on large deviations, in this work, we use averaging method to study our problem. In contrast to the aforementioned works, we consider systems driven by two-time-scale Markov chains. A distinct feature is that we must handle nonhomogeneity. In addition, we use simple conditions. In fact, not more than the irreducibility of the generator $Q(t)$ is required. Under such simple conditions, we obtain large deviations upper and lower bounds. Then we use the large deviations results to treat LQ problem in Example 3.1.

Assumption A. The generator $Q(t)$ is irreducible for each $t \in [0,T]$, and $Q(\cdot)$ is twice continuously differentiable in $[0,T]$.

Let us begin with a continuous-time Markov chain with state space $\mathcal{M}$ and generator Q that is irreducible. Then we can obtain the following results. The proof can be found in [46]. In what follows, we use $O(y)$ to denote the function of y satisfying $\sup_y |O(y)|/|y| < \infty$. Likewise, $o(y)$ denotes the function of y satisfying $|o(y)|/|y| \to 0$. In particular, $g = O(1)$ denotes the boundedness of g and $g = o(1)$ indicates $g \to 0$.

Lemma 3.2. *Suppose that $X(t)$ is a homogeneous and irreducible Markov chain with gen-*

erator Q and state space $\mathcal{M} = \{1, 2, \dots, m\}$. Then for each $i, j \in \mathcal{M}$,

$$E^i\Big(\frac{1}{\sqrt{T}}\int_0^T [I_{\{X(s)=j\}} - \nu_j]c_j ds\Big)^2 = O\Big(1\Big), \tag{3.9}$$

where c_j is a real number, $\nu = (\nu_1, \dots, \nu_{m_0})$ is the stationary distribution corresponding to the generator Q, and E^i denotes the expectation with $X(0) = i$.

Lemma 3.3. *With c_i defined in Lemma 3.2, let*

$$\begin{aligned} &z_T(t) = (z_T(t,1), \dots, z_T(t,m_0))' \in \mathbb{R}_0^m, \\ &z_T(t,j) = \frac{1}{\sqrt{T}}\int_0^{Tt} [I_{\{X(s)=j\}} - \nu_j]c_j ds, \quad \textit{for} \;\; t \in [0,1]. \end{aligned} \tag{3.10}$$

Assume the conditions of Lemma 3.2. Then as $T \to \infty$, $z_T(\cdot)$ converges weakly to $z(\cdot)$, a Brownian motion with mean 0 and covariance Σt, where $\Sigma = (\sigma_{ij}) \in \mathbb{R}^{m_0 \times m_0}$ with

$$\sigma_{ij} = c_i c_j \{\nu_i \int_0^\infty [p_{ij}(s) - \nu_j]ds + \nu_j \int_0^\infty [p_{ji}(s) - \nu_i]ds\}, \tag{3.11}$$

and $p_{ij}(s)$ is the ijth entry of the transition matrix $P(s)$.

To proceed, let us define the indecomposability of an arbitrary matrix. Note that many books also use the term irreducibility. To distinguish it with irreducibility of generators of Markov chains, we will use the terminology indecomposable throughout.

Definition 3.4. *For $n \geq 2$, an $m_0 \times m_0$ matrix A is decomposable if there exists an $m_0 \times m_0$ permutation matrix P such that*

$$PAP' = \begin{bmatrix} E & G \\ 0 & F \end{bmatrix}, \tag{3.12}$$

where E and F are square matrices of appropriate dimensions, P is a permutation matrix, and P' denotes the transpose of P. If no such permutation matrix exists, then A is indecomposable.

In what follows, for a matrix $A = (a_{ij})$, by $A \geq 0$ (resp. $A > 0$) we mean $a_{ij} \geq 0$ (resp. $a_{ij} > 0$). Then by virtue of [45, p. 282, Theorem 8.2], the following lemma holds.

Lemma 3.5. *Let A be a real matrix. Then $e^{At} > 0$ for all $t > 0$ if and only if $a_{ij} \geq 0$ for all $i \neq j$ and A is indecomposable.*

Lemma 3.6. *Assume the generator Q is irreducible. Then, $e^{(Q+B)t} > 0$ for any $t > 0$ and any diagonal matrix $B = diag(b_1, b_2, \ldots, b_{m_0})$.*

Proof. By virtue of Lemma 4.20, we need only show that $Q + B$ is indecomposable for any $B = \mathrm{diag}(b_1, b_2, \ldots, b_{m_0})$. To this end, suppose the contrary. Assume $Q+B$ is decomposable. Then there exists permutation matrix P such that

$$P(Q+B)P' = \begin{bmatrix} E & G \\ 0 & F \end{bmatrix},$$

where $E \in \mathbb{R}^{m_1 \times m_1}$ and $F \in \mathbb{R}^{m_2 \times m_2}$ with $m_1 + m_2 = m_0$. Hence,

$$Q = P' \begin{bmatrix} E & G \\ 0 & F \end{bmatrix} P - B = P'\big(\begin{bmatrix} E & G \\ 0 & F \end{bmatrix} - PBP'\big)P.$$

Since PBP' is still a diagonal matrix, we can write $PBP' = \begin{bmatrix} B_1 & 0 \\ 0 & B_2 \end{bmatrix}$, where $B_1 \in \mathbb{R}^{m_1 \times m_1}$ and $B_2 \in \mathbb{R}^{m_2 \times m_2}$ are diagonal matrix. So

$$Q = P' \begin{bmatrix} E - B_1 & G \\ 0 & F - B_2 \end{bmatrix} P.$$

This contradicts the irreducibility of Q. $\square$

Consider a homogeneous Markov chain $\{X(t) : t \geq 0\}$ with a generator Q that is irreducible. We have the following result.

Lemma 3.7. *Suppose that $\tau \in \mathbb{R}^k$ is a constant vector and $X(t)$ is stationary Markov chain with generator Q that is irreducible. Then for each $i \in \mathcal{M}$, the limit*

$$\lim_{T\to\infty} \frac{1}{T} \log E^i \exp\Big(\int_0^T \langle f(X(s)), \tau\rangle ds\Big) \tag{3.13}$$

exists, which is denoted by $H(\tau)$. The function $H(\cdot)$ is differentiable and convex.

Proof. The proof is similar to [18, p. 235, Theorem 4.2]. However, in the aforementioned reference, it is required that all $q_{ij} > 0$ for $i \neq j$. By virtue of Lemma 4.21, we only need $q_{ij} \geq 0$ for $i \neq j$. Define $A(\tau)$ to be the matrix whose entries are given by $A_{ij}(\tau) = q_{ij} + \delta_{ij}\big\langle f(i), \tau\big\rangle$, where $\delta_{ij} = 1$ when $i = j$ and 0 otherwise. Then $H(\tau)$ exists and is a real eigenvalue of $A(\tau)$, which exceeds the real parts of all other eigenvalues of $A(\tau)$. Furthermore, there exists an eigenvector $u(\tau)$ of $A(\tau)$, which satisfies $u(\tau) = (u_1, \dots, u_m)$ and $0 < \min_{i\leq m} u_i \leq \max_{i\leq m} u_i \leq 1$. The differentiability of $H(\tau)$ with respect to τ follows from the differentiability of the entries of $A(\tau)$. Moreover, The convexity of $H(\tau)$ follows from the convexity of the exponential function and the monotonicity and concavity of the logarithmic function. □

Remark 3.8. In fact, the $H(\tau)$ in the lemma depends on Q, so it may be written as $H_Q(\tau)$ if we wish to emphasize the Q dependence. However, for notational simplicity, we suppress the Q dependence. In view of (3.13), the $H(\tau)$ is roughly the limit of the logarithm of the moment generating function of $\int_0^T f(X(s))ds$. This $H(\tau)$ is the so-called H-functional related to large deviations theory. For necessary background of large deviations, and related discussions on H-functionals and large deviations rate functions etc., we refer the reader to [10, 12, 18, 20] among others. The next section is devoted H-functional of two-time-scale systems.

3.2 H-Functionals for Two-time-scale Markov Chains

First, let us state our standing assumption to be used throughout the chapter.

Assumption A. The generator $Q(t)$ is irreducible for each $t \in [0,T]$, and $Q(\cdot)$ is twice continuously differentiable in $[0,T]$.

In large deviations theory, H-functional plays a crucial role. In our case, this is related to the following limit

$$\lim_{\epsilon\to 0} \epsilon \log E^i \exp\{\frac{1}{\epsilon}\int_0^T \big\langle f(\alpha^\epsilon(s)), \tau(s)\big\rangle ds\},$$

where $\tau(\cdot) : [0,T] \mapsto \mathbb{R}^k$ is a step function, $f(\cdot) : \mathcal{M} \mapsto \mathbb{R}^k$, $\big\langle \cdot, \cdot \big\rangle$ denotes inner product on $\mathbb{R}^k$, and E^i denotes the expectation with $\alpha^\epsilon(0) = i$.

We will use Lemma 3.7 to obtain the result for the nonhomogeneous case. For notational simplicity, we write it $H(\tau, t)$ in what follows. The dependence of t is because of $H(\tau, t)$ and in view of Remark 3.8, this could be written as $H_{Q(t)}(\tau)$.

Theorem 3.9. *Consider $\alpha^\epsilon(\cdot)$. For Δ sufficiently small, each t satisfying $[t, t+\Delta] \subset [0,T]$, and for each $i \in \mathcal{M}$,*

$$\lim_{\epsilon\to 0} \epsilon \log E^i \exp\Big(\frac{1}{\epsilon}\int_t^{t+\Delta} \big\langle f(\alpha^\epsilon(s)), \tau\big\rangle ds\Big) = H(\tau, t)\Delta + o(\Delta).$$

Proof. Let $r = s - t$ and define $\widetilde{\alpha}^\epsilon(r) = \alpha^\epsilon(r+t)$. Then

$$E^i \exp\Big(\frac{1}{\epsilon}\int_t^{t+\Delta} \big\langle f(\alpha^\epsilon(s)), \tau\big\rangle ds\Big) = E^i \exp\Big(\frac{1}{\epsilon}\int_0^{\Delta} \big\langle f(\widetilde{\alpha}^\epsilon(r)), \tau\big\rangle dr\Big).$$

Thus, it reduces to prove that for $t = 0$,

$$\lim_{\epsilon \to 0} \epsilon \log E^i \exp\Big(\frac{1}{\epsilon}\int_0^\Delta \big\langle f(\alpha^\epsilon(s)), \tau\big\rangle ds\Big) = H(\tau, 0)\Delta + o(\Delta).$$

Define a family of operators $T^\epsilon_{s,t}$, for $0 \le s \le t$ by

$$T^\epsilon_{s,t} w(i) = E^{\epsilon s, i} w(\alpha^\epsilon(\epsilon t)) \exp\Big(\int_s^t \big\langle f(\alpha^\epsilon(\epsilon u)), \tau\big\rangle du\Big),$$

where w is a vector in $\mathbb{R}^m$. It follows that $T^\epsilon_{s,t}$ has monotone and semigroup properties,

$$\begin{aligned} &T^\epsilon_{s,t} w_1 \le T^\epsilon_{s,t} w_2, \quad \text{if} \quad w_1 \le w_2, \\ &T^\epsilon_{s,t} T^\epsilon_{t,p} = T^\epsilon_{s,p}, \quad \text{for} \quad s \le t \le p. \end{aligned} \tag{3.14}$$

We obtain that

$$A^\epsilon_t(\tau) = \lim_{t \to 0} \frac{T^\epsilon_{t,t+h} - I}{h} = (q_{ij}(\epsilon t) + \delta_{ij}\big\langle f(i), \tau\big\rangle), \tag{3.15}$$

where $q_{ij}(\epsilon s)$ is the ijth entry of $Q(\epsilon s)$. Define $A(\tau) = (q_{ij}(0) + \delta_{ij}\big\langle f(i), \tau\big\rangle)$. From the result of Lemma 3.7, $H(\tau, 0)$ is the real, simple eigenvalue of $A(\tau)$ that exceeds the real parts of all other eigenvalues, and there exists a eigenvector $u(\tau)$ of $A(\tau)$ that satisfies $u(\tau) = (u_1, \ldots, u_m)$ and $0 < c < \min_{i \le m} u_i \le \max_{i \le m} u_i \le 1$. For notation brevity, we use A and A^ϵ_t represent $A(\tau)$ and $A^\epsilon_t(\tau)$ respectively. Then $Au(\tau) = H(\tau, 0)u(\tau)$ and $e^{At}u(\tau) = e^{H(\tau,0)t}u(\tau)$. Let $\mathbf{1}$ be the vector with all components being one. Then by the positivity of the vector $u(\tau)$

$$\begin{aligned} (A^\epsilon_t - A)u(\tau) &\le |(A^\epsilon_t - A)u(\tau)|\, \mathbf{1} \\ &\le |(A^\epsilon_t - A)||u(\tau)|\, \mathbf{1} \\ &\le \frac{1}{c}|(A^\epsilon_t - A)||u(\tau)|u(\tau). \end{aligned}$$

In the above and hereafter, for two vectors v and $\widetilde{v}$, by $v \le \widetilde{v}$, we mean each component of v is less than or equal to that of $\widetilde{v}$. In addition, $|A|$ and $|u(\tau)|$ denote the matrix and vector norms, respectively. Similarly, we have

$$\begin{aligned}(A_t^\epsilon - A)u(\tau) &\geq -|(A_t^\epsilon - A)u(\tau)|\,\mathbf{1}\\ &\geq -\frac{1}{c}|(A_t^\epsilon - A)||u(\tau)|u(\tau).\end{aligned}$$

Consider $T_{0,t}^\epsilon, t \in [0, \dfrac{\Delta}{\epsilon}]$, and let

$$K(\Delta) = \sup_{t\in[0,\frac{\Delta}{\epsilon}]} |A_t^\epsilon - A|\frac{|u(\tau)|}{c} = \sup_{t\in[0,\Delta]} |Q(t) - Q(0)|\frac{|u(\tau)|}{c}.$$

Then we have the following inequalities,

$$-K(\Delta)u(\tau) \leq (A_t^\epsilon - A)u(\tau) \leq K(\Delta)u(\tau).$$

Furthermore, by the smoothness of $Q(t)$, $\lim_{\Delta\to 0} K(\Delta) = 0$. By virtue of the semigroup property (3.14) and derivative (3.15),

$$\begin{aligned}\frac{d(T_{0,t}^\epsilon)}{dt} &= \lim_{h\to 0}\frac{T_{0,t+h}^\epsilon - T_{0,t}^\epsilon}{h}\\ &= T_{0,t}^\epsilon \lim_{h\to 0}\frac{T_{t,t+h}^\epsilon - T_{t,t}^\epsilon}{h}\\ &= T_{0,t}^\epsilon A_t^\epsilon.\end{aligned}$$

Since $T_{s,t}^\epsilon$ is monotonic and $u(\tau) > 0$, we have

$$\begin{aligned}(H(\tau,0) - K(\Delta))T_{0,t}^\epsilon u(\tau) &\leq T_{0,t}^\epsilon[A - K(\Delta)]u(\tau)\\ &\leq \frac{d(T_{0,t}^\epsilon u(\tau))}{dt} = T_{0,t}^\epsilon A_t^\epsilon u(\tau)\\ &= T_{0,t}^\epsilon[A + (A_t^\epsilon - A)]u(\tau)\\ &\leq T_{0,t}^\epsilon[A + K(\Delta)]u(\tau) = (H(\tau,0) + K(\Delta))T_{0,t}^\epsilon u(\tau).\end{aligned}$$

By applying Gronwall's inequality , we can conclude that

$$e^{(H(\tau,0)-K(\Delta))t}u(\tau) \leq T_{0,t}^\epsilon u(\tau) \leq e^{(H(\tau,0)+K(\Delta))t}u(\tau).$$

Again, by the monotonicity of $T_{0,t}^\epsilon$,

$$c(T_{0,t}^\epsilon \mathbf{1})(i) \leq (T_{0,t}^\epsilon u(\tau))(i) \leq e^{(H(\tau,0)+K(\Delta))t}u(\tau)(i),$$

where $\mathbf{l}$ is the vector with all component equals one. Note that

$$\begin{aligned}
&\limsup_{\epsilon\to 0}\epsilon\log E^i\exp\Big(\frac{1}{\epsilon}\int_0^{\Delta}\big\langle f(\alpha^\epsilon(s)),\tau\big\rangle ds\Big)\\
&=\limsup_{\epsilon\to 0}\epsilon\log E^i\exp\Big(\int_0^{\frac{\Delta}{\epsilon}}\big\langle f(\alpha^\epsilon(\epsilon s)),\tau\big\rangle ds\Big)\\
&=\limsup_{\epsilon\to 0}\epsilon\log(T^\epsilon_{0,\frac{\Delta}{\epsilon}}\mathbf{l})(i)\\
&\le\limsup_{\epsilon\to 0}\epsilon\log\frac{1}{c}(T^\epsilon_{0,\frac{\Delta}{\epsilon}}u(\tau))(i)\\
&\le\limsup_{\epsilon\to 0}\epsilon\log\frac{1}{c}e^{H(\tau,0)\frac{\Delta}{\epsilon}}+\limsup_{\epsilon\to 0}\epsilon\log(e^{\frac{\Delta}{\epsilon}K(\Delta)}u(\tau)(i))\\
&=H(\tau,0)\Delta+\Delta K(\Delta)\\
&=H(\tau,0)\Delta+o(\Delta).
\end{aligned}$$

Hence,

$$\limsup_{\epsilon\to 0}\epsilon\log E^i\exp\Big(\frac{1}{\epsilon}\int_0^{\Delta}\big\langle f(\alpha^\epsilon(s)),\tau\big\rangle ds\Big)\le H(\tau,0)\Delta+o(\Delta).$$

Similarly, by applying the fact that

$$(T^\epsilon_{0,t}\mathbf{l})(i)\ge(T^\epsilon_{0,t}u(\tau))(i)\ge e^{H(\tau,0)-K(\Delta)}u(\tau)(i),$$

we obtain that

$$\begin{aligned}
&\liminf_{\epsilon\to 0}\epsilon\log E^i\exp\Big(\frac{1}{\epsilon}\int_0^{\Delta}\big\langle f(\alpha^\epsilon(s)),\tau\big\rangle ds\Big)\\
&=\liminf_{\epsilon\to 0}\epsilon\log E^i\exp\Big(\int_0^{\frac{\Delta}{\epsilon}}\big\langle f(\alpha^\epsilon(\epsilon s)),\tau\big\rangle ds\Big)\\
&=\liminf_{\epsilon\to 0}\epsilon\log(T^\epsilon_{0,\frac{\Delta}{\epsilon}}\mathbf{l})(i)\\
&\ge\liminf_{\epsilon\to 0}\epsilon\log(T^\epsilon_{0,\frac{\Delta}{\epsilon}}u(\tau))(i)\\
&\ge\liminf_{\epsilon\to 0}\epsilon\log e^{H(\tau,0)\frac{\Delta}{\epsilon}}+\liminf_{\epsilon\to 0}\epsilon\log(e^{-\frac{\Delta}{\epsilon}K(\Delta)}u(\tau)(i))\\
&=H(\tau,0)\Delta+o(\Delta).
\end{aligned}$$

Thus this lemma is concluded. □

Theorem 3.10. *Assume that $f:\mathcal{M}\to\mathbb{R}^k$ is a well-defined function, $\tau(t)$ is a step function on $[0,T]$, and $\alpha^\epsilon(t)$ is a continuous-time Markov chain with state space $\mathcal{M}$ and generator*

$Q^\epsilon(t) = Q(t)/\epsilon$. *Then under Assumption A with* $H(\cdot,\cdot)$ *defined previously,*

$$\lim_{\epsilon\to 0} \epsilon \log E^i \exp\{\frac{1}{\epsilon}\int_0^T \big\langle f(\alpha^\epsilon(s)), \tau(s)\big\rangle ds\} = \int_0^T H(\tau(t), t)dt.$$

Proof. Partition $[0,T]$ into subintervals so that $0 = t_0 \le t_1 \le t_2 \le \cdots \le t_n = T$, where $t_k = k\Delta$ for $k < n$ and $n = \lfloor \frac{T}{\Delta} \rfloor + 1$. Let τ_k be the value of $\tau(t)$ on $[t_{k-1}, t_k)$. Then by the Markov property of $\alpha^\epsilon(t)$,

$$\begin{aligned}
&\lim_{\epsilon\to 0} \epsilon \log E^i \exp\{\frac{1}{\epsilon}\int_0^T \big\langle f(\alpha^\epsilon(s)), \tau(s)\big\rangle ds\} \\
&\quad = \lim_{\epsilon\to 0} \epsilon \log E^i \exp\{\frac{1}{\epsilon}\int_{t_0}^{t_1} \big\langle f(\alpha^\epsilon(s)), \tau_1\big\rangle ds\} E^{\alpha^\epsilon(t_1)} \exp\{\frac{1}{\epsilon}\int_{t_1}^{t_2} \big\langle f(\alpha^\epsilon(s)), \tau_2\big\rangle) ds\} \\
&\qquad \times \cdots \times E^{\alpha^\epsilon(t_{n-1})} \exp\{\frac{1}{\epsilon}\int_{t_{n-1}}^{t_n} \big\langle f(\alpha^\epsilon(s)), \tau_n\big\rangle ds\} \\
&\quad = \sum_{k=0}^{n-1} \lim_{\epsilon\to 0} \epsilon \log E^{\alpha^\epsilon(t_k)} \exp\{\frac{1}{\epsilon}\int_{t_k}^{t_{k+1}} \big\langle f(\alpha^\epsilon(s)), \tau_k\big\rangle ds\},
\end{aligned}$$

where $E^{\alpha^\epsilon(t_0)} = E^{\alpha^\epsilon(0)} = E^i$. By using Theorem 3.9,

$$\lim_{\epsilon\to 0} \epsilon \log E^{\alpha^\epsilon(t_k)} \exp\{\frac{1}{\epsilon}\int_{t_k}^{t_{k+1}} \big\langle f(\alpha^\epsilon(s)), \tau(s)\big\rangle ds\} = H(\tau_k, t_k)(t_{k+1} - t_k) + o(\Delta)$$

Letting $\epsilon \to 0$ and $\Delta \to 0$, we have

$$\begin{aligned}
&\lim_{\Delta\to 0}\lim_{\epsilon\to 0} \epsilon \log E^i \exp\{\frac{1}{\epsilon}\int_0^T \big\langle f(\alpha^\epsilon(s)) ds, \tau(s)\big\rangle\} \\
&\quad = \lim_{\Delta\to 0}[\sum_{k=0}^{n-1} H(\tau_k, t_k)(t_{k+1} - t_k) + \frac{T}{\Delta} o(\Delta)] \\
&\quad = \int_0^T H(\tau(t), t)dt.
\end{aligned}$$

The theorem is thus proved. $\square$

3.3 LDP: Fixed Terminal Time

Consider $n^\epsilon(T) = \int_0^T f(\alpha^\epsilon(t))dt$ for a fixed time T, a random vector in $\mathbb{R}^k$. We introduce the Legendre transform $I(\cdot)$ on $\mathbb{R}^k$:

$$I(\gamma) = \sup_{\tau \in \mathbb{R}^k} [\langle \gamma, \tau \rangle - \int_0^T H(\tau, t)dt]. \tag{3.16}$$

It is readily seen that $I(\cdot)$ is convex and lower semi-continuous.

By Theorem 3.10, all the conditions in the Gärtner-Ellis (2.2) are satisfied. The following is the main result.

Theorem 3.11. *Under the conditions of Theorem* 3.10*, we have the following large deviation estimate. That is, for any* $B \subset \mathbb{R}^k$

$$\begin{aligned} -\inf_{\gamma \in B^\circ} I(\gamma) &\leq \liminf_{\epsilon \to 0} \epsilon \log P\{n^\epsilon(T) \in B\} \\ &\leq \limsup_{\epsilon \to 0} \epsilon \log P\{n^\epsilon(T) \in B\} \\ &\leq -\inf_{\gamma \in \overline{B}} I(\gamma), \end{aligned}$$

where B° *and* $\overline{B}$ *are interior and closure of* B *in* $\mathbb{R}^k$*, respectively.*

Let $\delta(\alpha) = (I_{\{\alpha=1\}}, \ldots, I_{\{\alpha=m\}})$, and put $f = \delta(\alpha)$. Then $n^\epsilon(T)$ is just the occupation measure. The results can be strengthened when $Q(t) = Q$ is time homogeneous as following; see [20, p. 47].

Corollary 3.12. *When* $\alpha^\epsilon(t)$ *is homogeneous, that is* $Q(t) = Q$*, the rate function* S *which is defined in (4.13) can be simplified to*

$$I(\gamma) = \sup_{\tau > 0} [-T \sum_{k=1}^m \gamma_k \frac{(Q\tau)_k}{\tau_k}].$$

If Q *is symmetric, the supremum can be evaluated explicitly:*

$$I(\gamma) = -T \sum_{k,j=1}^m \sqrt{\gamma_k} Q_{kj} \sqrt{\gamma_j} = \langle \sqrt{\gamma}, -Q\sqrt{\gamma} \rangle.$$

3.4 LDP: Time-Varying Processes

For $t \in [0,T]$, $n^\epsilon(t)$ is a solution of the ordinary differential equation

$$\dot{n}^\epsilon(t) = f(\alpha^\epsilon(t)), \quad n^\epsilon(0) = 0.$$

Define $\overline{f}(t) = \sum_{i=1}^m f(i)\nu_i(t)$. Then $n^\epsilon(\cdot)$ converges in probability to $\overline{n}(\cdot)$, which is a solution of

$$\dot{\overline{n}}(t) = \overline{f}(t).$$

Inspired by [17], we can use the idea of averaging to treat $n^\epsilon(\cdot)$. However, we need some modification to fit our nonhomogeneous case. We state one of main results below. In what follows, denote $C_{0,T}(\mathbb{R}^k)$ the set of all the continuous functions $\varphi : [0,T] \to \mathbb{R}^k$, and $C^x_{0,T}(\mathbb{R}^k) = \{\varphi :\in C_{0,T}(\mathbb{R}^k), \varphi(0) = x\}$.

Theorem 3.13. *If assumption A holds, then for each set $B \subset C^0_{0,T}(\mathbb{R}^k)$,*

$$\begin{aligned} -\inf_{\varphi \in B^\circ} I(\varphi) &\le \liminf_{\epsilon \to 0} \epsilon \log P\{n^\epsilon \in B\} \\ &\le \limsup_{\epsilon \to 0} \epsilon \log P\{n^\epsilon \in B\} \\ &\le -\inf_{\varphi \in \overline{B}} I(\varphi), \end{aligned} \tag{3.17}$$

where

$$\begin{aligned} I(\varphi) &= \begin{cases} \displaystyle\int_0^T L(\dot{\varphi}(s), s)ds, & \text{if } \varphi \in C_{0,T}(\mathbb{R}^k) \text{ is absolutely continuous,} \\ \infty & \text{otherwise,} \end{cases} \\ L(\gamma, s) &= \sup_\tau [\langle \gamma, \tau \rangle - H(\tau, s)], \end{aligned}$$

B° and $\overline{B}$ denote interior and closure of B in $C^0_{0,T}(\mathbb{R}^k)$, respectively.

To proceed, we introduce the following equivalent statement of large deviations estimate (3.17).

Lemma 3.14. *Let $\Phi(s) = \{\varphi \in C^0_{0,T}(\mathbb{R}^k), I(\varphi) \leq s\}$. Then (3.17) in Theorem 3.13 is equivalent to the following statement :*

For each $\varphi \in C^0_{0,T}(\mathbb{R}^k)$, and each $s \geq 0, h > 0$ and $\delta > 0$, there is an $\epsilon_0 > 0$ such that for $\epsilon \leq \epsilon_0$,

$$\begin{aligned} P\{\rho_{0T}(n^\epsilon, \varphi) < \delta\} &\geq \exp(-\frac{1}{\epsilon}(I(\varphi) + h)), \\ P\{\rho_{0T}(n^\epsilon, \Phi(s)) > \delta\} &\leq \exp(-\frac{1}{\epsilon}(s - h)), \end{aligned} \tag{3.18}$$

where ρ_{0T} is the metric on $C_{0,T}(\mathbb{R}^k)$ defined by

$$\rho_{0T}(\varphi_1, \varphi_2) = \sup_{0 \leq t \leq T} |\varphi_1(t) - \varphi_2(t)|.$$

Proof. See [18, Theorem 3.3, p.85]. □

By this lemma, to prove Theorem 3.13, we need only show that (3.18) hold. Note that the functional $I(\varphi)$ is lower semi-continuous, and the set $\Phi(s)$ is compact in $C_{0,T}(\mathbb{R}^k)$ [18, Lemma 4.2, p.231].

Lemma 3.15. *Suppose Assumption A holds, then as $\epsilon \to 0$, for any $s, \delta, h > 0$ and $\varphi \in C^0_{0,T}(\mathbb{R}^k)$,*

$$P\{\rho_{0T}(n^\epsilon, \varphi) < \delta\} \geq \exp\{-\frac{1}{\epsilon}(I(\varphi) + h)\}, \tag{3.19}$$

and

$$P\{\rho_{0T}(n^\epsilon, \Phi(s)) > \delta\} \leq \exp\{-\frac{1}{\epsilon}(s - h)\}, \tag{3.20}$$

where $\Phi(s) = \{\varphi \in C_{0,T}(\mathbb{R}^k) : \varphi(0) = 0, I(\varphi) \leq s\}$.

Proof. We choose small number Δ such that $\frac{T}{\Delta} = n$ is integer. Let $\tau_1, \tau_2, \ldots, \tau_n \in \mathbb{R}^k$, and define $\tau(s)$ as a piecewise constant function on $[0, T]$ such that for $s \in [(i-1)\Delta, i\Delta)$, it takes value $\sum_{k=i}^{n} \tau_k, (i = 1, 2, 3, \ldots, n)$. Denoting the function $h_\epsilon(\tau_1, \ldots, \tau_n)$ by

$$h_\epsilon(\tau_1, \ldots, \tau_n) = \epsilon \log E \exp\{\frac{1}{\epsilon} \sum_{k=1}^{n} \big\langle \tau_k, n^\epsilon(k\Delta) \big\rangle \}.$$

Observe that

$$\begin{aligned} \sum_{k=1}^{n} \big\langle \tau_k, n^\epsilon(k\Delta) \big\rangle &= \big\langle \sum_{k=1}^{n} \tau_k, \int_0^{\Delta} f(\alpha^\epsilon(s))ds \big\rangle + \big\langle \sum_{k=2}^{n} \tau_k, \int_{\Delta}^{2\Delta} f(\alpha^\epsilon(s))ds \big\rangle \\ &\quad + \cdots + \big\langle \tau_n, \int_{(n-1)\Delta}^{n\Delta} f(\alpha^\epsilon(s))ds \big\rangle. \end{aligned}$$

Hence, by the definition of $\tau(s)$, we can write

$$h_\epsilon(\tau_1, \ldots, \tau_n) = \epsilon \log E \exp\{\frac{1}{\epsilon} \int_0^{T} \big\langle \tau(s), f(\alpha^\epsilon(s)) \big\rangle \} ds.$$

By virtue of Theorem 3.10, we have the limit

$$h(\tau_1, \ldots, \tau_n) = \lim_{\epsilon \to 0} h_\epsilon(\tau_1, \ldots, \tau_n) = \int_0^{T} H(\tau(s), s) ds.$$

It can be seen that $h(\tau_1, \ldots, \tau_n)$ is convex and differentiable in the variables $\tau_1, \ldots, \tau_n$.

Then we define the Legendre transform of $h(\tau_1, \ldots, \tau_n)$. Let $\gamma(s)$ be a piecewise linear function on $[0, T]$ having jumps at integer multiples of Δ and taking value γ_k, at $k\Delta$ and

$\gamma_0 = 0$. By definition of the Legendre transform and the mean-value theorem,

$$
\begin{aligned}
l(\gamma_1, \dots, \gamma_n) &= \sup_{\tau_1,\dots,\tau_n} [\sum_{k=1}^{n} \langle \gamma_k, \tau_k \rangle - \int_0^T H(\tau(s), s) ds] \\
&= \sup_{\tau_1,\dots,\tau_n} [\sum_{k=1}^{n} \langle \gamma_k, \tau_k \rangle - \sum_{k=1}^{n} \int_{(k-1)\Delta}^{k\Delta} H(\tau(s), s) ds] \\
&= \sup_{\tau_1,\dots,\tau_n} [\sum_{k=1}^{n} \langle \gamma_k, \tau_k \rangle - \Delta \sum_{k=1}^{n} H(\sum_{i=k}^{n} \tau_i, s_k) ds] \\
&= \Delta \sup_{\tau_2,\dots,\tau_n} \{ \sup_{\tau_1} [\langle \frac{\gamma_1}{\Delta}, \tau_1 + \sum_{i=2}^{n} \tau_i \rangle - H(\tau_1 + \sum_{i=2}^{n} \tau_i, s_1) ds] \\
&\quad + [\sum_{k=2}^{n} \langle \frac{\gamma_k - \gamma_1}{\Delta}, \tau_k \rangle - \sum_{k=2}^{n} H(\sum_{i=k}^{n} \tau_i, s_k) ds] \} \\
&= \Delta \sup_{\tau_2,\dots,\tau_n} \{ L(\frac{\gamma_1}{\Delta}, s_1) + [\sum_{k=2}^{n} \langle \frac{\gamma_k - \gamma_1}{\Delta}, \tau_k \rangle - \sum_{k=2}^{n} H(\sum_{i=k}^{n} \tau_i, s_k) ds] \} \\
&= \Delta L(\frac{\gamma_1}{\Delta}, s_1) + \Delta \sup_{\tau_2,\dots,\tau_n} [\sum_{k=2}^{n} \langle \frac{\gamma_k - \gamma_1}{\Delta}, \tau_k \rangle - \sum_{k=2}^{n} H(\sum_{i=k}^{n} \tau_i, s_k) ds] \\
&= \cdots \\
&= \sum_{k=1}^{n} L(\frac{\gamma_k - \gamma_{k-1}}{\Delta}, s_k),
\end{aligned}
\tag{3.21}
$$

where $s_k \in [(k-1)\Delta, k\Delta)$.

If we let $\eta^\epsilon = (n^\epsilon(\Delta), n^\epsilon(2\Delta), \dots, n^\epsilon(n\Delta))$, and $\tau = (\tau_1, \dots, \tau_n) \in (\mathbb{R}^k)^n$. Then

$$
h(\tau_1, \dots, \tau_n) = \lim_{\epsilon \to 0} \epsilon \log E \exp\{ \frac{1}{\epsilon} \langle \tau, \eta^\epsilon \rangle \}.
$$

For $e = (e_1, \dots, e_n), g = (g_1, \dots, g_n) \in (\mathbb{R}^k)^n$, define

$$
\begin{aligned}
&\Phi^\Delta(s) = \{ (e_1, \dots, e_n) : l(e_1, \dots, e_n) \le s \} \quad \text{for} \quad s < \infty, \\
&\overline{\rho}(e, g) = \max_{1 \le i \le n} |e_i - g_i|.
\end{aligned}
$$

Then by Lemma 2.4 and Lemma 3.14, it follows that for any $a, \delta, h > 0$ and any $\gamma \in (\mathbb{R}^k)^n$ and for sufficiently small ϵ the following inequalities hold

$$
\begin{aligned}
&P\{ \overline{\rho}(\eta^\epsilon, \gamma) < \delta \} > \exp\{ -\frac{1}{\epsilon} (l(\gamma_1, \dots, \gamma_n) + h) \}, \\
&P\{ \overline{\rho}(\eta^\epsilon, \Phi^\Delta(s)) > \delta \} < \exp\{ -\frac{1}{\epsilon} (s - h) \}.
\end{aligned}
\tag{3.22}
$$

Let $\varphi \in C_{0,T}(\mathbb{R}^k), I(\varphi) < \infty, \delta > 0$. We define $\overline{\varphi}^{\Delta} = (\varphi(\Delta), \varphi(2\Delta), \ldots, \varphi(n\Delta)) \in (\mathbb{R}^k)^n$. Relying on the fact that the trajectories of n^{ϵ} and φ for which $I(\varphi) < \infty$ are Lipschitz continuous, it follows that for sufficiently small $\widetilde{\delta}$ and Δ,

$$P\{\rho_{0,T}(n^{\epsilon}, \varphi) < \delta\} > P\{\overline{\rho}(\eta^{\epsilon}, \overline{\varphi}^{\Delta}) < \widetilde{\delta}\}. \tag{3.23}$$

By (3.21) and the continuity of $L(\varphi(s), s)$ in s, for each $\frac{h}{2} > 0$, we can choose sufficient small Δ such that

$$|l(\gamma_1, \ldots, \gamma_n) - \int_0^T L(\dot{\gamma}(s), s)| < \frac{h}{2},$$

where $\dot{\gamma}(s) = (d/ds)\gamma(s)$. Estimating the right-hand side of (3.23), for each $h > 0$ and for sufficient small ϵ, we have

$$\begin{aligned} P\{\rho_{0,T}(n^{\epsilon}, \varphi) < \delta\} &> P\{\overline{\rho}(\eta^{\epsilon}, \overline{\varphi}^{\Delta}) < \widetilde{\delta}\} \\ &\geq \exp\{-\frac{1}{\epsilon}(l(\varphi_{\Delta}, \ldots, \varphi_{n\Delta}) + \frac{h}{2})\} \\ &\geq \exp\{-\frac{1}{\epsilon}(\int_0^T L(\dot{\overline{\varphi}}_s, s)ds + h)\}, \end{aligned}$$

where $\overline{\varphi}_s$ is a piecewise linear function having jumps at integer multiples of Δ and being identical to $\varphi(s)$ at $s = k\Delta$ for $k = 1, \ldots, n$. Next, since φ is absolutely continuous and $L(\gamma, t)$ is convex with respect to γ, we obtain

$$\begin{aligned} \int_0^T L(\dot{\overline{\varphi}}(s), s)ds &= \sum_{k=1}^{n} \int_{(k-1)\Delta}^{k\Delta} L(\dot{\overline{\varphi}}(s), s)ds \\ &= \sum_{k=1}^{n} L(\frac{1}{\Delta}\int_{(k-1)\Delta}^{k\Delta} \dot{\varphi}(s)ds, s_k)\Delta \\ &\leq \sum_{k=1}^{n} \int_{(k-1)\Delta}^{k\Delta} L(\dot{\varphi}(s), s_k)ds, \quad \text{where } s_k \in [(k-1)\Delta, k\Delta). \end{aligned}$$

Again, by the continuity of $L(\dot{\varphi}(s), s)$, we can choose sufficient small Δ such that

$$\begin{aligned} \int_0^T L(\dot{\overline{\varphi}}(s), s)ds &\leq \sum_{k=1}^{n} \int_{(k-1)\Delta}^{k\Delta} L(\dot{\varphi}(s), s_k)ds \\ &\leq \int_0^T L(\dot{\varphi}(s), s)ds + h \\ &= I(\varphi) + h. \end{aligned}$$

Then (3.19) follows immediately. To get the second inequality we observe that for sufficiently small Δ and $\widetilde{\delta}$,

$$\{\rho_{0,T}(n^\epsilon, \Phi(s)) > \delta\} \subset \{\overline{\rho}(\eta^\epsilon, \Phi^\Delta(s)) > \widetilde{\delta}\},$$

From this inclusion and the estimate of (3.22), we obtain (3.20). The proof of this lemma is completed. □

Completion of Proof of Theorem 3.13**.** Theorem 3.13 follows immediately from the Lemma above and Lemma 3.14. □

3.5 LDP for ODEs with Markovian Switching

This section is devoted to dynamic systems represented by ordinary differential equations with Markovian switching. We consider nonlinear ordinary differential equations involving a Markov chain. The Markov chain $\alpha^\epsilon(\cdot)$ is fast varying and can be thought of as a noise process. The continuous state on the other hand, varies much slowly. As a result, an averaging takes place. The dynamic system is replaced by an average with respect to the quasi-stationary measure of the fast varying process. To be more specific, consider the system of ordinary differential equations

$$\dot{x}^\epsilon(t) = b(x^\epsilon(t), \alpha^\epsilon(t)), \quad x^\epsilon(0) = x. \tag{3.24}$$

Then under simple conditions, we obtain the following result, whose proof is based on the asymptotic properties of the two-time-scale Markov chains (see [46, Chapters 4 and 5]).

Lemma 3.16. *Suppose $b(x,i) : \mathbb{R}^k \times \mathcal{M} \mapsto \mathbb{R}^k$ such that for each $i \in \mathcal{M}$, $b(\cdot, i)$ grows at* *most linearly and satisfies the Lipschitz condition, i.e.,*

$$|b(x,i)| \leq K(1+|x|) \ \ \text{for each} \ \ x \in \mathbb{R}^k,$$
$$|b(x,i) - b(y,i)| \leq K|x-y| \ \ \text{for all} \ \ x, y \in \mathbb{R}^k.$$

Suppose also $\alpha^\epsilon(t) \sim Q(t)/\epsilon$, *where* $Q(t)$ *is irreducible. Then* $x^\epsilon(\cdot)$, *the solution of* (3.24), *converges weakly to* $x(\cdot)$ *such that for each* $T > 0$ *and for any* $t \in [0,T]$,

$$\dot{x}(t) = \overline{b}(x(t)), \ \ x(0) = x_0, \tag{3.25}$$

where

$$\overline{b}(x) = \sum_{j=1}^{m} b(x,i)\nu_i,$$

and $\nu(t) = (\nu_1(t), \ldots, \nu_m(t))$ *is the quasi-stationary distribution associated with* $Q(t)$.

Proof. Step 1. Show $\sup_{t\in[0,T_1]} E|x^\epsilon(t)|^2 < \infty$. We simply note that

$$|x^\epsilon(t)|^2 \leq 2|x_0|^2 + 2\Big|\int_0^t b(x^\epsilon(u), \alpha^\epsilon(u))du\Big|^2.$$

The desired result then follows from the Cauchy-Schwarz inequality, linear growth on $b(\cdot,\alpha)$ for each $\alpha \in \mathcal{M}$, and the Gronwall inequality.

Step 2. Show that $\{x^\epsilon(\cdot)\}$ is tight in $D([0,T];\mathbb{R}^k)$, the space of functions that are right continuous and have left-hand limits endowed with the Skorohod topology. For any $\delta > 0$, any $s, t \geq 0$ with $s \leq \delta$, the Cauchy-Schwarz inequality, the linear growth of $b(\cdot,\alpha)$ for each $\alpha \in \mathcal{M}$, and the moment bound in Step 1 lead to

$$\begin{aligned} E|x^\epsilon(t+s) - x^\epsilon(t)|^2 &= E\Big|\int_t^{t+s} b(x^\epsilon(u),\alpha^\epsilon(u))du\Big|^2 \\ &\leq sE\int_t^{t+s} |b(x^\epsilon(u),\alpha^\epsilon(u))|^2 du \leq Ks^2 \leq K\delta^2. \end{aligned}$$

Thus, $\lim_{\delta\to 0}\limsup_{\epsilon\to 0} E|x^\epsilon(t+s) - x^\epsilon(t)|^2 = 0$. Thus, $\{x^\epsilon(\cdot)\}$ is tight by [25, p. 47]. In addition, the limit process has continuous sample paths with probability one.

Step 3. Characterize the limit. Since $\{x^\epsilon(\cdot)\}$ is tight, by Prohorov's theorem, we can extract a weakly convergent subsequence. Extract such a sequence and still denote it by $\{x^\epsilon(\cdot)\}$ for notational simplicity. Denote the limit by $x(\cdot)$. By Skorohod representation with a slight abuse of notation, we may assume $x^\epsilon(\cdot)$ converges to $x(\cdot)$ w.p.1 and the convergence is uniform on any bounded interval. We proceed to characterize the limit process as a solution of a martingale problem with operator L, where the operator is defined as

$$Lf(x) = \left\langle \nabla f(x), \bar{b}(x) \right\rangle$$

for any $f \in C_0^1$ (collection of C_1 functions with compact support).

Note that the martingale problem mentioned above has a unique solution. This is mainly due to the fact that the associated ordinary differential equation has a unique solution owing to the Lipschitz continuity of $b(\cdot, \alpha)$ for each $\alpha \in \mathcal{M}$. For a detailed proof of a similar but more complex case, see [46, Lemma 7.18].

To characterize the limit, it suffices to show that

$$f(x(t)) - f(x(0)) - \int_0^t Lf(x(u))du \;\text{ is a martingale.} \tag{3.26}$$

To this end, let $h(\cdot)$ be any bounded and continuous function, $t, s \geq 0$, $\kappa > 0$ be an arbitrary positive integer, $t_l \leq t$ with $l \leq \kappa$. To verify (3.26), it suffices to show

$$Eh(x(t_l) : l \leq \kappa)[f(x(t+s)) - f(x(t)) - \int_0^t Lf(x(u))du] = 0. \tag{3.27}$$

We begin with the verification with the sequence indexed by ϵ. By the weak convergence and the Skorohod representation,

$$\begin{aligned} &Eh(x^\epsilon(t_l) : l \leq \kappa)[f(x^\epsilon(t+s)) - f(x^\epsilon(t))] \\ &\quad \to Eh(x(t_l) : l \leq \kappa)[f(x(t+s)) - f(x(t))] \;\text{ as }\; \epsilon \to 0. \end{aligned}$$

It is also easily seen that

$$Eh(x^\epsilon(t_l) : l \le \kappa)[f(x^\epsilon(t+s)) - f(x^\epsilon(t)) - \int_0^t L^\epsilon f(x^\epsilon(u))du] = 0,$$

where for each $V(\cdot, i) \in C_0^1$,

$$L^\epsilon V(x,i) = \big\langle \nabla V(x,i), b(x,i) \big\rangle + Q^\epsilon(t)V(x,\cdot)(i),\ i \in \mathcal{M},$$

with $Q^\epsilon(t) = (q_{ij}^\epsilon(t))$ and

$$Q^\epsilon(t)V(x,\cdot)(i) = \sum_{j=1}^m q_{ij}^\epsilon(t)V(x,j),\ i \in \mathcal{M},$$

for each $V(\cdot,i) \in C_0^1$ and $i \in \mathcal{M}$. Since the function $f(\cdot)$ is independent of $i \in \mathcal{M}$, $\sum_{j=1}^m q_{ij}^\epsilon(t) f(x) = 0$. Thus

$$\begin{aligned}
&Eh(x^\epsilon(t_l) : l \le \kappa)[\int_t^{t+s} Lf(x^\epsilon(u))du] \\
&\quad = Eh(x^\epsilon(t_l) : l \le \kappa)[\int_t^{t+s} \big\langle \nabla f(x^\epsilon(u)), b(x^\epsilon(u), \alpha^\epsilon(u)) \big\rangle du].
\end{aligned}$$

Note that

$$\begin{aligned}
&\int_t^{t+s} \big\langle \nabla f(x^\epsilon(u)), b(x^\epsilon(u), \alpha^\epsilon(u)) \big\rangle du \\
&\quad = \sum_{i=1}^m \int_t^{t+s} \big\langle \nabla f(x^\epsilon(u)), b(x^\epsilon(u), i) \big\rangle \nu_i(u) du \\
&\qquad + \sum_{i=1}^m \int_t^{t+s} \big\langle \nabla f(x^\epsilon(u)), b(x^\epsilon(u), i) \big\rangle [I_{\{\alpha^\epsilon(u)=i\}} - \nu_i(u)] du.
\end{aligned} \tag{3.28}$$

We shall show that the last term in (3.28) contributes to a limit zero. To proceed, we partition the interval $[t, t+s]$ as follows: For $0 < \Delta < 1$, let $t_0 = t < t_1 < t_2 < \cdots < t_N \le t+s$ such that $t_k = \epsilon^{1-\Delta}(k+1)$. For simplicity, we let $t_N = t+s$; the modification is straightforward otherwise. In addition, it suffices to work on a fixed $i \in \mathcal{M}$. Using the partition, and realize that $x^\epsilon(\cdot)$ changes relatively slowly compared to $\alpha^\epsilon(\cdot)$. In each subinterval $[t_k, t_{k+1}]$, we can approximate $\big\langle \nabla f(x^\epsilon(u)), b(x^\epsilon(u), \alpha^\epsilon(u)) \big\rangle$ by $\big\langle \nabla f(x^\epsilon(t_k)), b(x^\epsilon(t_k), \alpha^\epsilon(t_k)) \big\rangle$, a piecewise constant function. The difference of the about two goes to 0 owing to the smoothness of $f(\cdot)$ and $b(\cdot, i)$ and the selection of the partition $[t_k, t_{k+1}]$. Note that the total number of subintervals is $\lfloor s/\epsilon^{1-\Delta}\rfloor$, where $\lfloor z \rfloor$ is the usual floor function and gives the integer part of z for any $z \in \mathbb{R}$. Thus,

$$\begin{aligned}
&\lim_{\epsilon\to 0} Eh(x^\epsilon(t_l) : l \le \kappa) \sum_{k=1}^{\lfloor s/\epsilon^{1-\Delta}\rfloor} \int_t^{t+s} \big\langle \nabla f(x^\epsilon(u)), b(x^\epsilon(u), i)\big\rangle [I_{\{\alpha^\epsilon(u)=i\}} - \nu_i(u)]du \\
&\quad = \lim_{\epsilon\to 0} Eh(x^\epsilon(t_l) : l \le \kappa) \sum_{k=1}^{\lfloor s/\epsilon^{1-\Delta}\rfloor} \int_{t_k}^{t_{k+1}} \big\langle \nabla f(x^\epsilon(t_k)), b(x^\epsilon(t_k), i)\big\rangle [I_{\{\alpha^\epsilon(u)=i\}} - \nu_i(u)]du.
\end{aligned}$$

Detailed computation with the use of [46, Theorem 7.2] yields that the limit in the last expression is nothing but 0. Likewise it can be shown that

$$\begin{aligned}
&\lim_{\epsilon\to 0} Eh(x^\epsilon(t_l) : l \le \kappa) \sum_{k=1}^{\lfloor s/\epsilon^{1-\Delta}\rfloor} \int_{t_k}^{t_{k+1}} \big\langle \nabla f(x^\epsilon(u)), b(x^\epsilon(u), i)\big\rangle \nu_i(u)du \\
&\quad = \lim_{\epsilon\to 0} Eh(x^\epsilon(t_l) : l \le \kappa) \sum_{k=1}^{\lfloor s/\epsilon^{1-\Delta}\rfloor} \int_{t_k}^{t_{k+1}} \big\langle \nabla f(x^\epsilon(t_k)), b(x^\epsilon(t_k), i)\big\rangle \nu_i(u)du \\
&\quad = Eh(x(t_l) : l \le \kappa) \int_t^{t+s} \big\langle \nabla f(x(u)), b(x(u), i)\big\rangle \nu_i(u)du.
\end{aligned}$$

This leads to

$$\begin{aligned}
&\lim_{\epsilon\to 0} Eh(x^\epsilon(t_l) : l \le \kappa) \int_t^{t+s} \big\langle f(x^\epsilon(u)), b(x^\epsilon(u), \alpha^\epsilon(u))\big\rangle du \\
&\quad = Eh(x(t_l) : l \le \kappa) \int_t^{t+s} \big\langle \nabla f(x(u)), \overline{b}(x(u))\big\rangle du.
\end{aligned}$$

Piecing together the arguments used thus far, the desired limit is obtained. □

Remark 3.17. Strictly speaking, the $\overline{b}(x)$ above should be written as $\overline{b}(x, t)$ due to the t-dependence of the quasi-stationary distribution. Since $b(x, i)$ is not a function that depends on t explicitly, we use $\overline{b}(x)$ to represent the average.

Lemma 3.18. *Assume the conditions of Lemma* 3.16*. Then the set of solutions* $\{x^\epsilon(\cdot) : \epsilon > 0\}$ *is a compact subset of* $C_{0,T}(\mathbb{R}^k)$ *a.s.*

Proof. First note that

$$\begin{aligned}
|x^\epsilon(t)| &\le |x_0| + \int_0^t \sum_{i=1}^{m_0} |b(x^\epsilon(s), i) I_{\{\alpha^\epsilon(s)=i\}}| ds \\
&\le |x_0| + K \int_0^t (1 + |x^\epsilon(s)|) ds \\
&\le |x_0| + K \int_0^t (1 + \sup_{0\le u\le s} |x^\epsilon(u)|) ds.
\end{aligned}$$

Thus,

$$\sup_{0\le t\le T} |x^\epsilon(t)| \le |x_0| + K \int_0^T (1 + \sup_{0\le u\le s} |x^\epsilon(u)|) ds.$$

An application of Gronwall's inequality implies that $\sup_{0\le t\le T} |x^\epsilon(t)| \le K < \infty$ a.s. Thus $\{x^\epsilon(\cdot)\}$ is uniformly bounded. In the above and hereafter, K is a generic positive constant whose values may change for different appearances.

It follows that

$$\begin{aligned}
|x^\epsilon(t+s) - x^\epsilon(t)| &\le \int_t^{t+s} \sum_{i=1}^{m_0} \sup_{t\le u\le t+s} |b(x^\epsilon(u), i) I_{\{\alpha^\epsilon(u)=i\}}| du \\
&\le Ks \quad \text{a.s.}
\end{aligned}$$

Thus, $\{x^\epsilon(\cdot)\}$ is equicontinuous a.s. By the well-known Ascoli-Arzelá theorem, the desired compactness follows. □

We say that an H functional exist if there is a function $H(\cdot,\cdot,\cdot)$ such that

$$\lim_{\epsilon\to 0} \epsilon \log E^i \exp\{\frac{1}{\epsilon} \int_0^T \big\langle b(\varphi(s), \alpha^\epsilon(s)), \beta(s) \big\rangle ds\} = \int_0^T H(\varphi(s), \tau(s), s) ds,$$

for any step functions $\varphi(s)$ and $\tau(s)$ in $\mathbb{R}^k$. Similarly to Theorem 3.10, we have the following theorem to guarantee the existence of the H functional.

Theorem 3.19. *Under the conditions of Lemma* 3.16*, there exists a function* $H : \mathbb{R}^k \times \mathbb{R}^k \times [0,T] \to \mathbb{R}$ *such that*

$$\lim_{\epsilon\to 0} \epsilon \log E^i \exp\{\frac{1}{\epsilon} \int_0^T \big\langle b(\varphi(s), \alpha^\epsilon(s)), \tau(s) \big\rangle ds\} = \int_0^T H(\varphi(s), \tau(s), s) ds, \tag{3.29}$$

for any step functions $\varphi(s)$ and $\tau(s)$ on the interval $[0,T]$ with values in $\mathbb{R}^k$, and for any $i \in \mathcal{M}$. Moreover, H is jointly continuous in its variables and convex in the second argument.

Proof. Let $\varphi(s)$ and $\tau(s)$ be step functions and φ_k and τ_k be their values on interval $[t_{k-1}, t_k)$, for $0 \le t_0 < t_1 < \cdots < t_n = T$, where $t_k = k\Delta$ for $k < n$ and $n = \lfloor \frac{T}{\Delta} \rfloor + 1$. In view of the Markov property, we can write

$$\begin{aligned}
\lim\nolimits_{\epsilon\to 0}\ & \epsilon \log E^i \exp\{\frac{1}{\epsilon}\int_0^T \big\langle b(\varphi(s), \alpha^\epsilon(s)), \tau(s)\big\rangle ds\} \\
&= \lim_{\epsilon\to 0} \epsilon \log E^i \exp\{\frac{1}{\epsilon}\int_0^{t_1} \big\langle b(\varphi_1, \alpha^\epsilon(s)), \tau_1\big\rangle ds\} \\
&\quad \times E^{\alpha^\epsilon(t_1)} \exp\{\frac{1}{\epsilon}\int_{t_1}^{t_2} \big\langle b(\varphi_2, \alpha^\epsilon(s)), \tau_2\big\rangle ds\} \times \cdots \\
&\qquad \times E^{\alpha^\epsilon(t_{n-1})} \exp\{\frac{1}{\epsilon}\int_{t_{n-1}}^{t_n} \big\langle b(\varphi_n, \alpha^\epsilon(s)), \tau_n\big\rangle ds\}.
\end{aligned}$$

Applying Theorem 3.10, we obtain

$$\lim_{\epsilon\to 0} \epsilon \log E^i \exp\{\frac{1}{\epsilon}\int_{t_{k-1}}^{t_k} \big\langle b(\varphi_k, \alpha^\epsilon(s)), \tau_k\big\rangle ds\} = H(\varphi_k, \tau_k, t_{k-1})(t_k - t_{k-1}) + o(\Delta).$$

Estimating the same way on each interval $[t_{k-1}, t_k)$ gives us

$$\lim_{\epsilon\to 0} \epsilon \log E^i \exp\{\frac{1}{\epsilon}\int_0^T \big\langle b(\varphi_k, \alpha^\epsilon(s)), \tau_k\big\rangle ds\} = \sum_{k=1}^n H(\varphi_k, \tau_k, t_{k-1})(t_k - t_{k-1}) + \frac{T}{\Delta} o(\Delta).$$

Letting $\Delta \to 0$, (4.31) is proved. Similar to the proof of Theorem 3.10, it can be shown that H is jointly continuous and convex in τ. With the existence of H functional we are ready to present the following theorem. Recall that $C^x_{0,T}(\mathbb{R}^k) = \{\varphi :\in C_{0,T}(\mathbb{R}^k), \varphi(0) = x\}$.

Theorem 3.20. *Under the conditions of Theorem 3.19, as $\epsilon \to 0$, for any $s, \delta, h > 0$ and $\varphi \in C^x_{0,T}(\mathbb{R}^k)$,*

$$\begin{aligned}
P\{\rho_{0T}(x^\epsilon, \varphi) < \delta\} &\ge \exp\{-\frac{1}{\epsilon}(I(\varphi) + h)\}, \\
P\{\rho_{0T}(x^\epsilon, \Phi_x(s)) > \delta\} &\le \exp\{-\frac{1}{\epsilon}(s - h)\},
\end{aligned}$$

where $\Phi_x(s) = \{\varphi \in C^x_{0,T}(\mathbb{R}^k) : I(\varphi) \leq s\}$, *and*

$$I(\varphi) = \begin{cases} \displaystyle\int_0^T L(\varphi(s), \dot{\varphi}(s), s)ds, & \text{if } \varphi \in C_{0,T}(\mathbb{R}^k) \text{ is absolutely continuous,} \\ \infty & \text{otherwise,} \end{cases}$$

$$L(x, \gamma, s) = \sup_{\tau \in \mathbb{R}^k} [\langle \gamma, \tau \rangle - H(x, \tau, s)].$$

Proof. The proof uses similar approach as that of [18, Theorem 4.1]. The difference is that we need to deal with nonhomogenous Markov chain, i.e., the H functional and its Legendre transform are time dependent. Moreover, we treat the linear growth in x of $b(x, i)$ rather than bounded $b(x, i)$ in [18, Theorem 4.1]. Starting with Lemma 2.4, fix x and $\Delta > 0$. Let $\psi(\cdot)$ denote a function that is constant ψ_i on $[(i-1)\Delta, i\Delta)$. Define the function $x_t^{\psi,\epsilon}$ by

$$x_t^{\psi,\epsilon} = x + \int_0^t b(\psi(s), \alpha^\epsilon(s))ds.$$

Let $\tau_1, \tau_2, \ldots, \tau_n \in \mathbb{R}^k$. We define $\tau(s)$ as the piecewise constant function on $[0, T]$ that for $s \in [(i-1)\Delta, i\Delta)$ takes value $\sum_{k=i}^n \tau_k$, $(i = 1, 2, 3, \ldots, n)$. Denoting the function $h_\epsilon^x(\tau_1, \ldots, \tau_n)$ by

$$h_\epsilon^x(\tau_1, \ldots, \tau_n) = \epsilon \log E \exp\{\frac{1}{\epsilon} \sum_{k=1}^n \langle \tau_k, x_{k\Delta}^{\psi,\epsilon} \rangle\}.$$

Note that

$$\begin{aligned} \sum_{k=1}^n \langle \tau_k, x_{k\Delta}^{\psi,\epsilon} \rangle = {}& \langle \sum_{k=1}^n \tau_k, \int_0^\Delta b(\psi(s), \alpha^\epsilon(s))ds \rangle + \langle \sum_{k=2}^n \tau_k, \int_\Delta^{2\Delta} b(\psi(s), \alpha^\epsilon(s))ds \rangle \\ & + \cdots + \langle \tau_n, \int_{(n-1)\Delta}^{n\Delta} b(\psi(s), \alpha^\epsilon(s))ds \rangle + \langle \sum_{k=1}^n \tau_k, x \rangle. \end{aligned}$$

Hence, by the definition of $\tau(s)$, we can write

$$h_\epsilon^x(\tau_1, \ldots, \tau_n) = \epsilon \log E \exp\{\frac{1}{\epsilon} \int_0^T \langle \tau(s), b(\psi(s), \alpha^\epsilon(s)) \rangle\}ds + \langle \sum_{k=1}^n \tau_k, x \rangle.$$

By virtue of Theorem 3.19, we have the limit

$$\lim_{\epsilon \to 0} h_\epsilon^x(\tau_1, \ldots, \tau_n) = h^x(\tau_1, \ldots, \tau_n) + \langle \sum_{k=1}^n \tau_k, x \rangle,$$

and

$$h^x(\tau_1, \ldots, \tau_n) = \int_0^T H(\psi(s), \tau(s), s) ds.$$

We can show that $h^x(\tau_1, \ldots, \tau_n)$ is convex and differentiable in the variables $\tau_1, \ldots, \tau_n$.

Then we define the Legendre transform of $h^x(\tau_1, \ldots, \tau_n)$. Let $\gamma(s)$ be a piecewise linear function on $[0, T]$ having jumps at integer multiples of Δ and taking value $\gamma_k, \gamma_0 = x$, at $k\Delta$. If $x = 0$, then by definition of the Legendre transform and mean-value theorem,

$$\begin{aligned}
l^0(\gamma_1, \ldots, \gamma_n) &= \sup_{\tau_1, \ldots, \tau_n} [\sum_{k=1}^n \langle \gamma_k, \tau_k \rangle - \int_0^T H(\psi(s), \tau(s), s) ds] \\
&= \sup_{\tau_1, \ldots, \tau_n} [\sum_{k=1}^n \langle \gamma_k, \tau_k \rangle - \sum_{k=1}^n \int_{(k-1)\Delta}^{k\Delta} H(\psi(s), \tau(s), s) ds] \\
&= \sup_{\tau_1, \ldots, \tau_n} [\sum_{k=1}^n \langle \gamma_k, \tau_k \rangle - \Delta \sum_{k=1}^n H(\psi_k, \sum_{i=k}^n \tau_i, s_k) ds] \\
&= \Delta \sup_{\tau_2, \ldots, \tau_n} \{\sup_{\tau_1} [\langle \frac{\gamma_1}{\Delta}, \tau_1 + \sum_{i=2}^n \tau_i \rangle - H(\psi_1, \tau_1 + \sum_{i=2}^n \tau_i, s_1) ds] \\
&\quad + [\sum_{k=2}^n \langle \frac{\gamma_k - \gamma_1}{\Delta}, \tau_k \rangle - \sum_{k=2}^n H(\psi_k, \sum_{i=k}^n \tau_i, s_k) ds]\} \\
&= \Delta \sup_{\tau_2, \ldots, \tau_n} \{L(\psi_1, \frac{\gamma_1}{\Delta}, s_1) + [\sum_{k=2}^n \langle \frac{\gamma_k - \gamma_1}{\Delta}, \tau_k \rangle - \sum_{k=2}^n H(\psi_k, \sum_{i=k}^n \tau_i, s_k) ds]\} \\
&= \Delta L(\psi_1, \frac{\gamma_1}{\Delta}, s_1) + \Delta \sup_{\tau_2, \ldots, \tau_n} [\sum_{k=2}^n \langle \frac{\gamma_k - \gamma_1}{\Delta}, \tau_k \rangle - \sum_{k=2}^n H(\psi_k, \sum_{i=k}^n \tau_i, s_k) ds] \\
&= \cdots \\
&= \sum_{k=1}^n L(\psi_k, \frac{\gamma_k - \gamma_{k-1}}{\Delta}, s_k), \text{ where } s_k \in [(k-1)\Delta, k\Delta).
\end{aligned} \tag{3.30}$$

Observe that

$$l^x(\gamma_1, \ldots, \gamma_n) = l^0(\gamma_1 - x, \ldots, \gamma_n - x),$$

we have the same estimates (3.30) for arbitrary x. Furthermore, we can choose sufficient small Δ such that

$$|l^x(\gamma_1, \ldots, \gamma_n) - \int_0^T L(\psi(s), \dot{\gamma}(s), s) ds| < \delta, \tag{3.31}$$

for any $\delta > 0$. If we let $\eta^\epsilon = (x^{\psi,\epsilon}_{\Delta}, x^{\psi,\epsilon}_{2\Delta}, \ldots, x^{\psi,\epsilon}_{n\Delta})$, and $\tau = (\tau_1, \ldots, \tau_n)$. Then

$$h^x(\tau_1, \ldots, \tau_n) = \lim_{\epsilon \to 0} \epsilon \log E \exp\{\frac{1}{\epsilon}\langle \tau, \eta^\epsilon \rangle\}.$$

For e_i and $g_i \in \mathbb{R}^k$, define

$$\begin{aligned}
&\Phi^\Delta(s) = \{e = (e_1, \ldots, e_n) : l(e_1, \ldots, e_n) \le s\} \quad \text{for} \quad s < \infty, \\
&\bar{\rho}(e, g) = \max_{1 \le i \le n} |e_i - g_i|.
\end{aligned}$$

Let $\varphi(\cdot)$ denote a continuous function and φ_Δ denote the vector $\{\varphi(i\Delta), i \le n\}$. Define

$$\begin{aligned}
&S^\psi_T(\varphi) = \int_0^T L(\psi(s), \dot\varphi(s), s) ds, \\
&\Phi^\psi_x(s) = \{\varphi(\cdot) \in C_{0,T}(\mathbb{R}^k) : \varphi(0) = x, \ S^\psi_T(\varphi) \le s\},
\end{aligned}$$

respectively. It can be shown that $S^\psi_T(\varphi)$ is lower semi-continuous in both φ and ψ, and $\Phi^\psi_x(s)$ is compact. Applying Lemma 2.4 and Lemma 3.15, we obtain a large deviation estimate for the samples of $x^{\psi,\epsilon}$ and φ, with sampling interval Δ,

$$\begin{aligned}
&P\{\bar\rho(\eta^\epsilon, \varphi_\Delta) < \delta\} \ge \exp\{-\frac{1}{\epsilon}(l^x(\varphi(\Delta), \ldots, \varphi(n\Delta)) + h)\} \\
&P\{\bar\rho(\eta^\epsilon, \Phi^\Delta(s)) > \delta\} \le -\frac{(s-h)}{\epsilon}.
\end{aligned} \tag{3.32}$$

Then following the same proof as that of Lemma 3.15 and using the Lipschitz continuity of the trajectories of $x^{\epsilon,\psi}$ and φ for which $I(\varphi) < \infty$, we obtain that for each fix ψ and φ, $s \ge 0, h > 0$, and $\delta > 0$, there is an $\epsilon > 0$ such that for $\epsilon < \epsilon_0$

$$\begin{aligned}
&P\{\rho_{0T}(x^{\psi,\epsilon}, \varphi) < \delta\} \ge -\frac{(S^\psi_T(\varphi) + h)}{\epsilon}, \\
&P\{\rho_{0T}(x^{\psi,\epsilon}, \Phi_x(s)) > \delta\} \le -\frac{(s-h)}{\epsilon}.
\end{aligned} \tag{3.33}$$

Similar to [18, Lemma 5.2], we can prove that for any step function sequence ψ^n converging to some function $\varphi \in C_{0,T}(\mathbb{R}^k)$, there exists a sequence $\varphi^n \in C_{0,T}(\mathbb{R}^k)$, uniformly converging to φ such that

$$\limsup_{n\to\infty} \int_0^T L(\psi^n(s), \dot\varphi^n(s), s) \le S_T(\varphi).$$

Applying this fact, for any $\varphi \in C_{0,T}(\mathbb{R}^k)$ and $S_{0T}(\varphi) < \infty$, and any $\lambda > 0$, we can choose step function ψ^λ and $\varphi^\lambda \in C_{0,T}(\mathbb{R}^k)$ such that

$$\begin{aligned} &\rho_{0T}(\varphi^\lambda, \varphi) < \lambda, \\ &\sup_{0 \le t \le T} |\psi^\lambda(t) - \varphi(t)| < \lambda, \quad \text{and} \\ &\int_0^T L(\psi^\lambda(s), \dot{\varphi}^\lambda(s), s) ds < S_T(\varphi) + h. \end{aligned}$$

Then by the Lipschitz continuity of the function $b(x, i)$, we have

$$\begin{aligned} |x^\epsilon(t) - \varphi(t)| &\le |x^\epsilon(t) - x^{\varphi,\epsilon}(t)| + |x^{\varphi,\epsilon}(t) - x^{\psi^\lambda,\epsilon}(t)| \\ &\quad + |x^{\psi^\lambda,\epsilon}(t) - \varphi^\lambda| + |\varphi^\lambda - \varphi(t)| \\ &\le K \int_0^t |x^\epsilon(s) - \varphi(s)| ds + K \int_0^t |\varphi(s) - \psi^\lambda(s)| ds + \rho_{0T}(x^{\psi^\lambda,\epsilon}, \varphi^\lambda) + \lambda \\ &\le \rho_{0T}(x^{\psi^\lambda,\epsilon}, \varphi^\lambda) + \lambda + Kt\lambda + K \int_0^t |x^\epsilon(s) - \varphi(s)| ds. \end{aligned}$$

Applying Gronwall's inequality,

$$P\{\rho_{0T}(x^\epsilon, \varphi) < \delta\} \ge P\{\rho_{0T}(x^{\psi^\lambda,\epsilon}, \varphi^\lambda) < \widetilde{\delta}\}$$

provided that λ and $\widetilde{\delta}$ are sufficiently small. Directly computation and estimate (3.33) yield

$$\begin{aligned} P\{\rho_{0T}(x^\epsilon, \varphi) < \delta\} &\ge P\{\rho_{0T}(x^{\psi^\lambda,\epsilon}, \varphi^\lambda) < \widetilde{\delta}\} \\ &\ge \exp\{-\frac{S^{\psi^\lambda}(\varphi^\lambda) + h}{\epsilon}\} \\ &\ge \exp\{-\frac{I(\varphi) + 2h}{\epsilon}\} \end{aligned}$$

for sufficiently small ϵ. To get the second inequality, by means of Lemma 3.18, the trajectories of x^ϵ forms a compact set in $C_{0,T}(\mathbb{R}^k)$. Let us denote it by F. Let F_1 be the compact set obtained from F by omitting the $\delta/2$ neighborhood of the set $\Phi_x(s)$. Then applying the lower semi-continuity of $S_T^\psi(\varphi)$ in both ψ and φ, for any $h > 0$ there exists a δ_h such that $S_T^\psi(\varphi) > s - h/2$ if $\rho_{0T}(\varphi, \psi) < \delta_h$ and $I(\varphi) > s$ for any $\varphi \in F_1$. Denote $\widetilde{\delta} = \delta_h/(4KT + 2)$. By the compactness of F, choose finite $\widetilde{\delta}$ net of it and let $\varphi_1, \ldots, \varphi_n$ be the elements of this net which belonging to F_1. It can be seen that

$$P\{\rho_{0T}(x^\epsilon,\Phi_x(s))>\delta\}\le\sum_{i=1}^{n}P\{\rho_{0T}(x^\epsilon,\varphi_i)<\widetilde{\delta}\},\ \text{if }\widetilde{\delta}<\delta.$$

Again, relying on the Lipschitz continuity of $b(x,i)$, for $\rho_{0T}(\varphi,\psi)<\widetilde{\delta}$, we obtain that

$$P\{\rho_{0T}(x^\epsilon,\varphi)<\widetilde{\delta}\}\le P\{\rho_{0T}(x^{\epsilon,\psi},\varphi)<(2KT+1)\widetilde{\delta}\}.$$

We can choose step functions $\psi_1,\dots,\psi_n$ such that $\rho_{0T}(\psi_i,\varphi_i)<\frac{\widetilde{\delta}}{2}$. The above two inequalities imply that

$$P\{\rho_{0T}(x^\epsilon,\Phi_x(s))>\delta\}\le\sum_{i=1}^{n}P\{\rho_{0T}(x^{\epsilon,\psi_i},\varphi_i)\le(2KT+1)\widetilde{\delta}\}.$$

Estimating every terms on the right-hand side by (3.33), By Lemma 3.14 and the definition of δ_h

$$\begin{aligned}P\{\rho_{0T}(x^{\epsilon,\psi_i},\varphi_i)<(2KT+1)\widetilde{\delta}\}&=P\{\rho_{0T}(x^{\epsilon,\psi_i},\varphi_i)<\delta_h/2\}\\&\le\exp\{-\frac{\inf\{S_T^{\psi_i}(\varphi):\rho_{0T}(\varphi,\varphi_i)<\delta_h/2\}-h/4}{\epsilon}\}\\&\le\exp\{-\frac{s-h}{\epsilon}\}.\end{aligned}$$

Finally, we get the lower estimate

$$P\{\rho_{0T}(x^\epsilon,\Phi_x(s))>\delta\}\le\exp\{-\frac{s-h}{\epsilon}\}.$$

The desired result thus follows. □

Applying Lemma 3.14, we have the following large deviation estimate immediately.

Theorem 3.21. *Under assumption of Lemma* 3.16*, then for each set* $B\subset C_{0,T}^{x}(\mathbb{R}^k)$*,*

$$\begin{aligned}-\inf_{\varphi\in B^\circ}I(\varphi)&\le\liminf_{\epsilon\to0}\epsilon\log P\{x^\epsilon\in B\}\\&\le\limsup_{\epsilon\to0}\epsilon\log P\{x^\epsilon\in B\}\\&\le-\inf_{\varphi\in\overline{B}}I(\varphi),\end{aligned}$$

where

$$
\begin{aligned}
I(\varphi) \qquad &= \begin{cases} \displaystyle\int_0^T L(\varphi(s), \dot{\varphi}(s), s)ds, & \textit{if} \;\; \varphi \in C_{0,T}(\mathbb{R}^k) \textit{ is absolutely continuous,} \\ \infty \;\; \textit{otherwise,} \end{cases} \\
L(x, \gamma, s) \;\; &= \sup_{\tau \in \mathbb{R}^k} [\langle \gamma, \tau \rangle - H(x, \tau, s)],
\end{aligned}
$$

B° *and* $\overline{B}$ *denote interior and closure of* B *in* $C^x_{0,T}(\mathbb{R}^k)$, *respectively.*

3.6 Application of LDP in LQ Problem

This section is devoted to the application to the LQ problem given in Example 3.1. As in [49], we can show that $K^\epsilon(s,i) \to \overline{K}(s)$ uniformly on $[0,T]$ as $\epsilon \to 0$, where $\overline{K}(s)$ is the unique solutions to the following differential equation:

$$
\dot{\overline{K}}(s) = -\overline{K}(s)\overline{A}(s) - \overline{A}'(s)\overline{K}(s) - \overline{M}(s) + \overline{K}(s)\overline{BN^{-1}B'}(s)\overline{K}(s), \tag{3.34}
$$

with $\overline{K}(T) = D$, with

$$
\begin{aligned}
&\overline{A}(s) = \sum_{j=1}^m \nu_j(s) A(j), \;\; \overline{M}(s) = \sum_{j=1}^m \nu_j(s) M(j) \;\; \text{and} \\
&\overline{BN^{-1}B}(s) = \sum_{j=1}^m \nu_j(s) B(j) N^{-1}(j) B'(j).
\end{aligned}
$$

Using the idea presented in [49], we will be able to obtain near-optimal control of the system. The idea is to use the optimal control of the limit system to build controls and apply that to the original system.

Using the weak convergence method together with (3.8) and similar to Section 3.5, we are able to show that $x^\epsilon(\cdot)$ converges weakly to $x(\cdot)$, which is a solution of

$$
\dot{\overline{x}}(t) = \Big(\overline{A}(t)x(t) - \overline{BN^{-1}B'}(t)\overline{K}(s)\Big)\overline{x}(t). \tag{3.35}
$$

To see how good the approximation is, one can estimate the error $\rho_{0T}(x^\epsilon, \overline{x})$. To do this, we can introduce an intermediate process $\overline{x}^\epsilon(t)$ defined as follows:

$$\dot{\overline{x}}^\epsilon = \Big(A(\alpha^\epsilon(t)) - B(\alpha^\epsilon(t))N^{-1}(\alpha^\epsilon(t))B'(\alpha^\epsilon(t))\overline{K}(t)\Big)\overline{x}^\epsilon(t),$$

with $\overline{x}^\epsilon(s) = x$, Here we used the constructed near optimal control

$$\overline{u}^\epsilon(t,\alpha,x) = -N^{-1}(\alpha)B'(\alpha)\overline{K}(t)x.$$

By considering the difference $(x^\epsilon(t) - \overline{x}^\epsilon(t))^2$ and using Gronwall's inequality,

$$\rho_{0T}(x^\epsilon, \overline{x}^\epsilon) \le C\sqrt{\sup_{t\in[0,T]} \int_0^t \Big(K^\epsilon(\alpha^\epsilon(r), r) - \overline{K}(r)\Big)^2 dr} = C\sqrt{\int_0^T \Big(K^\epsilon(\alpha^\epsilon(r), r) - \overline{K}(r)\Big)^2 dr},$$

for some constant C. Note that K^ϵ and $\overline{K}$ are both deterministic functions, the error $(K^\epsilon - \overline{K})$ can be determined by simple numerical comparison. Using the triangle inequality

$$\rho_{0T}(x^\epsilon, \overline{x}) \le \rho_{0T}(x^\epsilon, \overline{x}^\epsilon) + \rho_{0T}(\overline{x}^\epsilon, \overline{x}),$$

one remains to estimate $\rho_{0T}(\overline{x}^\epsilon, \overline{x})$. This can be done with our large deviations result. By Theorem 3.21, for any $\delta > 0$ there exists $d_0 > 0$ such that

$$P(\rho_{0T}(\overline{x}^\epsilon, \overline{x}) \ge \delta) \le \exp(-\frac{d_0}{\epsilon}),$$

for sufficient small ϵ. To summarize the discussion up to now, we have the following proposition.

Proposition 3.22. *There exist positive constants C and d_0 such that*

$$\rho_{0T}(x^\epsilon, \overline{x}) \le C\sqrt{\int_0^T \Big(K^\epsilon(\alpha^\epsilon(r), r) - \overline{K}(r)\Big)^2 dr} + \exp(-\frac{d_0}{\epsilon}),$$

for sufficient small ϵ.

Chapter 4: LDP of System Identification

Traditional system identification concentrates on convergence and convergence rates of estimates in suitable senses, such as in mean squares, in distribution, or with probability one. Such asymptotical analysis is inadequate in applications that require a finite data analysis. Especially, for system diagnosis and its related complexity analysis, it is essential to understand probability of identification errors over a finite data window. For example, in real-time diagnosis, parameter values must be evaluated to judge if they belong to a “normal” region or a “fault” has occurred. This set-based identification amounts to a hypothesis testing, which relies on a probabilistic characterization of parameter estimates.

To address such applications, this work investigates identification errors in a probabilistic framework. By considering both space complexity in terms of signal quantization and time complexity with respect to data window sizes, this study provides a new perspective to understand the fundamental relationship between probabilistic errors and resources, which may represent data sizes in computer usage, computational complexity in algorithms, sample sizes in statistical analysis, channel bandwidths in communications, etc. This relationship is derived by establishing large deviations principles for quantized identification which links binary-valued data at one end and regular sensor at the other. Under some mild conditions, we obtain both large deviation upper and lower bounds. In addition, our results accommodate typical independent and identically distributed noise sequences, as well as more general classes of mixing-type noise sequences.

Traditional system identification under regular sensors is a mature field with many significant results concerning identification errors, convergence, convergence rates, under deterministic or stochastic noises, and their applications in model validation, adaptive control, diagnosis, signal processing, etc. Our study departs from the existing literature from several perspectives. First, rather than concentrating on convergence of the parameter estimates and/or treating asymptotic distributions of centered and suitably scaled sequence of the estimates, we investigate large deviations of the parameter estimators, which provide an accurate characterization on dependence of probabilistic identification errors on data window sizes. Second, this study requires different techniques from traditional tools for identification convergence analysis. Third, we are dealing with large deviation principles on quantized identification, which represents a first attempt in this direction. Finally, our results target different applications and delineate the rate functions of the estimators that resolve some intriguing questions defying clarifying answers under other identification frameworks.

To elaborate our motivations, consider the following scenario. Suppose a sequence of vector-valued estimates $\{\widehat{\theta}_N\}$ of the true parameter θ has been generated by an identification algorithm, and under suitable (and rather broad) conditions, the sequence is strongly consistent in the sense of convergence with probability one (w.p.1). We may further establish that $\sqrt{N}(\widehat{\theta}_N - \theta)$ converge in distribution to a normal random variable with mean zero and a derived covariance. Such estimates are termed as consistent (convergence to the true value) with the scaling factor and asymptotic normality providing rates of convergence. They, however, do not specify errors in probability at a finite time. Suppose that we are interested in the quality of estimation in terms of $P(|\widehat{\theta}_N - \theta| > a)$ for a given $a > 0$. The strong consistency or the asymptotic normality can affirm

$$P(|\widehat{\theta}_N - \theta| > a) \to 0 \quad \text{as} \quad n \to \infty, \tag{4.1}$$

but cannot give more precise bounds for a finite N. Large deviation principles represent probabilistic errors as a function of N by deriving a rate function. When both upper and lower bounds on rate functions are derived, they can be used to characterize guaranteed error bounds in probability using the upper bound and study complexity issues using the lower bound. As such they offer distinctive aspects of identification errors beyond conventional measures of identification accuracy.

In this dissertation, large deviation principles on identification accuracy are developed under both regular sensors and quantized observations. The setup for identification under regular sensors follows the persistent identification framework introduced in [38], whereas system identification with binary and quantized sensors is developed within the quantized identification setup of [44]. To accommodate common practical scenarios of correlated noises, this work deals with noises of mixing types. On the other hand, detailed treatments for uncorrelated noises are also presented separately, due to their relative simplicity and clarity of the results. Consequently, they allow a clear interpretation of the results and help in understanding the key issues involved.

4.1 System Setup

Consider a single-input-single-output (SISO) linear time-invariant (LTI) stable discrete-time system

$$y(t) = \sum_{i=0}^{\infty} a_i u(t-i) + d(t), \quad t = t_0 + 1, \ldots, \tag{4.2}$$

where $\{y(t)\}$ is the noise corrupted observation, $\{d(t)\}$ is the disturbance, $\{u(t)\}$ is the input with $u(t) = 0, t < 0$, and $a = \{a_i, i = 0, 1, \ldots\}$, satisfying $\|a\|_1 = \sum_{i=0}^{\infty} |a_i| < \infty$. To proceed, we define

$$\theta = (a_0, a_1, \ldots, a_{m_0-1})' \in \mathbb{R}^{m_0}, \quad \widetilde{\theta} = (a_{m_0}, a_{m_0+1}, \ldots)', \tag{4.3}$$

where z' denotes the transpose of z, θ is the vector-valued modeled part of the parameters, and $\widetilde{\theta}$ is known as the unmodeled dynamics. Separation of the modeled part and unmodeled dynamics is a standard modeling practice to limit model complexity [37, 38, 50], which enables us to treat parameters within a finite dimensional space; see also related work [?] and references therein. This model complexity reduction produces model errors, due to the "truncation."

Throughout this chapter, we assume that the input u is uniformly bounded $\|u\|_\infty \leq u_{\max}$. After applying u to the system and taking N output observations in the time interval $t_0, \ldots, t_0 + N - 1$, the observation can be rewritten as

$$y(t) = \varphi'(t)\theta + \widetilde{\varphi}'(t)\widetilde{\theta} + d(t), \tag{4.4}$$

where

$$\begin{aligned} &\varphi(t) = (u(t), u(t-1), \ldots, u(t-m_0+1))', \quad \text{and} \\ &\widetilde{\varphi}(t) = (u(t-m_0), u(t-m_0-1), \ldots)', \end{aligned} \tag{4.5}$$

or, in a vector form

$$Y_N(t_0) = \Phi_N(t_0)\theta + \widetilde{\Phi}_N(t_0)\widetilde{\theta} + D_N(t_0),$$

where $\varpi_N(t_0) = (\varpi(t_0), \ldots, \varpi(t_0+N-1))'$ for $\varpi_N(t_0)$ being $Y_N(t_0)$, or $D_N(t_0)$, or $\Phi_N(t_0)$, or $\widetilde{\Phi}_N(t_0)$. Estimates will be derived from this relationship, depending on the sensing schemes used for y. Large deviations of the estimates will be investigated accordingly.

Remark 4.1. Typical linear finite-dimensional stable systems have rational transfer functions. When represented them by their impulse responses, they are always IIR (infinite impulse response), with a decaying tail. Consequently, it is essential that the model structure (4.2) starts with an IIR expression.

Note a truncation on the IIR is used to reach an FIR (finite impulse response) model plus an unmodeled dynamics. There are many other approaches to approximate a higher-order or infinite-dimensional system, including more general base functions, state space model reduction, Hankel-norm reduction, etc. It turns out that the FIR model allows a much simpler algorithm development than other approaches. Consequently, we have adopted it here.

4.2 LDP of System Identification under I.I.D. Noise

We first consider system identification under i.i.d. noises. Extension to correlated noises will be treated later. We begin by making the following assumptions. We should emphasize here that since we consider open-loop identification problems, the input signal u is part of experimental design and can be selected to enhance the identification process.

(A1) (a) $\{d(t)\}$ is a sequence of independent and identically distributed zero-mean random variables. Its moment generating function exists and is denoted by $g(t)$. (b) $\Phi_N(t_0)$ has full column rank. (c) The input signal $\{u(t)\}$ is periodic with period m_0.

Remark 4.2. The condition (b) is a persistent excitation condition, ensuring that the input is sufficiently rich for parameter estimation. When the input signal is periodic, this condition can be substantially simplified. The condition (c) is quite unique. In our previous work [38, 41, 44], we have demonstrated some key desirable features of periodic inputs: (1) Under a periodic full-rank input, a complicated identification problem can be decomposed into a finite number of very simple identification problems; (2) Under a periodic full-rank input, we can identify a rational transfer function under quantized observations without essential difficulties; (3) Under an externally applied periodic full-rank input, we can identify a system in the closed-loop setting with guaranteed persistent excitation; (4) For applications of laws of large numbers, under periodic inputs, estimation errors can be written as direct averages, leading to possibility of deriving not only upper bounds, but also CR lower bounds. However, non-periodic signals can certainly be used. In this case, without the above benefits, we can still derive upper bounds on estimation errors under the persistent excitation condition (b). But the results will be more conservative and lower bounds are harder to obtain. Since we aim to investigate complexity issues, tight error bounds are essential. Consequently, this chapter focuses on periodic full-rank inputs.

4.2.1 LDP of System Identification with Regular Sensors

Consider an m_0-periodic signal u, and denote $N = km_0$ with an integer k. To simplify the expression, we write $\Phi_{m_0}(t_0)$ as Φ_0. We can write $\Phi_N(t_0)$ and $\widetilde{\Phi}_N(t_0)$ in (4.5) as

$$\begin{aligned}\Phi_N(t_0) &= (I_{m_0}, \ldots, I_{m_0})'\Phi_0,\\ \widetilde{\Phi}_N(t_0) &= (\Phi_N(t_0), \Phi_N(t_0), \ldots),\end{aligned}$$

where I_{m_0} denotes the $m_0 \times m_0$ identity matrix. In what follows, we apply the standard least squares estimation method. Denote $L(t_0) = (\Phi'_N(t_0)\Phi_N(t_0))^{-1}\Phi'_N(t_0)$. Then

$$L(t_0) = \frac{1}{k}\Phi_0^{-1}(I_{m_0}, \ldots, I_{m_0}). \tag{4.6}$$

Define the estimator $\widehat{\theta}_k = L(t_0)Y_N(t_0)$. Then

$$\widehat{\theta}_k = \theta + L(t_0)\widetilde{\Phi}_N(t_0)\widetilde{\theta} + L(t_0)D_N(t_0). \tag{4.7}$$

It follows that the deterministic part of the identification error becomes

$$\eta_k^d = (I_{m_0}, I_{m_0}, \ldots)\widetilde{\theta}. \tag{4.8}$$

Since η_k^d is independent of k, we write it as η^d. It is easily seen from (4.8) that $\|\eta^d\|_1 \le \|\widetilde{\theta}\|_1$. The stochastic part of the identification error is

$$\begin{aligned} \eta_k^s &= \Phi_0^{-1}(U_k^1, \ldots, U_k^{m_0})', \quad \text{where} \\ U_k^i &= \frac{1}{k}\sum_{l=0}^{k-1} d(t_0 + lm_0 + i), \quad \text{for} \quad i = 1, \ldots, m_0. \end{aligned} \tag{4.9}$$

Since $\{d(t)\}$ is an i.i.d sequence, η_k^s tends to 0 as $k \to \infty$, w.p.1. Hence, $\lim_{k\to\infty} \widehat{\theta}_k = \theta + \eta^d$, w.p.1., where $\|\eta^d\|_1 \le \|\widetilde{\theta}\|_1$. In Wang and Yin [38], identification error bounds were obtained by using a combined approach of stochastic averaging and worst-case identification methods. To proceed, our task here is to establish probabilistic error bounds on η_k^s by the large deviations principle.

Remark 4.3. Typical probabilistic errors of identification problems only consider the form $P(|\widehat{\theta}_k - \theta| \ge a)$ for some $a > 0$. The LDP is a more general and refined property in that it permits probabilistic characterization of the estimates on any open or closed sets. For ease of presentation, the statement of the above theorem is concerned with an open set B. It can be stated in terms of a closed set B. If a closed set B is used, on the left-hand side, replace B by its interior B^0 and on the right-hand side replace $\overline{B}$ by B. In a more general set, we can state the result using a Borel set B together with B^0 and $\overline{B}$ used on the left side and right side of the inequality, respectively.

To proceed, consider $U_k = (U_k^1, \ldots, U_k^{m_0})$ defined in (4.9) and let G be a linear function on $\mathbb{R}^{m_0}$ defined by $G(x) = \theta + \eta^d + \Phi_0^{-1}x,\ x \in \mathbb{R}^{m_0}$, where η^d comes from the unmodeled dynamics term. Note that $\widehat{\theta}_k = G(U_k)$ and Φ_0 is full rank. To establish the LDP on $\widehat{\theta}_k$, we will first find the rate function of $\{U_k\}$ and then apply the contraction principle Lemma 2.5 to derive the rate function of $\widehat{\theta}_k$. For $\tau = (\tau_1, \dots, \tau_{m_0})'$ the H-functional of $\{U_k\}$ is

$$\begin{aligned} H(\tau) &= \lim_{k\to\infty} \frac{1}{k} \log E \exp\{k\langle U_k, \tau\rangle\} \\ &= \lim_{k\to\infty} \frac{1}{k} \log E \exp\{\sum_{l=0}^{k-1}\sum_{i=1}^{m_0} d(t_0 + lm_0 + i)\tau_i\} \\ &= \log E \exp\{\sum_{i=1}^{m_0} d(t_0+i)\tau_i\} = \sum_{i=1}^{m_0} \log g(\tau_i). \end{aligned}$$

The Legendre transform of H is

$$\begin{aligned} I(\beta) &= \sup_{\tau\in\mathbb{R}^{m_0}} [\langle \beta, \tau\rangle - H(\tau)] \\ &= \sup_{\tau_1,\dots,\tau_{m_0}} [\sum_{i=1}^{m_0} (\beta_i\tau_i - \log g(\tau_i))]. \end{aligned} \tag{4.10}$$

Hence, by Lemma 2.4, $I(\beta)$ is the rate function of $\{U_k\}$.

Theorem 4.4. *Under Assumption* (A1), $\widetilde{I}(\widetilde{\beta}) = I(G^{-1}(\widetilde{\beta})) = I(\Phi_0(\widetilde{\beta} - \theta - \eta^d))$ *is the rate function for* $\{\widehat{\theta}_k\}$. *That is, for any open set* B *in* $\mathbb{R}^{m_0}$, (2.2) *holds with* $I(\beta)$ *replaced by* $\widetilde{I}(\widetilde{\beta})$, $X_k = \widehat{\theta}_k$, *and* $\lambda_k = (1/k)$, *respectively.*

Proof. Since $\widehat{\theta}_k = G(U_k)$ and G is bijection and continuous, the result is followed by the contraction principle Lemma 2.5 . □

Remark 4.5. If the i.i.d. noise $\{d(l)\}$ has the standard normal distribution, then the rate function $\widetilde{I}(\widetilde{\beta})$ has a simple form $\widetilde{I}(\widetilde{\beta}) = \dfrac{|\Phi_0(\widetilde{\beta} - \theta - \eta^d)|^2}{2}$, where $|\cdot|$ is the Euclidian norm.

4.2.2 LDP of System Identification with Binary Sensors

Suppose that the output y is measured by a binary sensor with a known threshold C. That is, we observe only $s(t) = \chi_{\{y(t)\le C\}} = \begin{cases} 1, & \text{if } y(t) \le C \\ 0, & \text{otherwise.} \end{cases}$

(A2) (a) $\{d(t)\}$ is a sequence of independent and identically distributed zero-mean random variables with distribution function $F(x)$, which is a bijection, continuous function whose inverse F^{-1} exists and is continuous. The moment generating function of $d(t)$ exists. (b) $\|\widetilde{\theta}\|_1 \le \widetilde{\eta}$.

In [44], Wang, Zhang, and Yin proposed an algorithm to determine θ, obtained its convergence, and studied the corresponding asymptotic distribution of normalized errors. Using the setup in (4.4), we recall the algorithm.

Algorithm.

Step 1. Define $Z_k = (Z_k^1, \dots, Z_k^{m_0})'$ with

$$Z_k^i = \frac{1}{k}\sum_{l=0}^{k-1} s(t_0 + lm_0 + i), \ \ i = 1, \dots, m_0. \tag{4.11}$$

Note that the event $\{y(t_0+lm_0+i) \le C\}$ is the same as the event $\{d(t_0+lm_0+i) \le \widetilde{c}_i\}$, where $\widetilde{c}_i = C - \widetilde{C}_i$ and $\widetilde{C}_i$ is the i-th component of $\Phi_0\theta + \widetilde{\Phi}_0\widetilde{\theta}$. Denote $\widetilde{c} = (\widetilde{c}_1, \dots, \widetilde{c}_{m_0})'$. Then, Z_k^i is the value of the k-sample empirical distribution.

Step 2. Since F is invertible, we can define

$$\begin{aligned} &\gamma_k^i = F^{-1}(Z_k^i), \ \ i = 1, \dots, m_0 \ \text{ and} \\ &\gamma_k = (\gamma_k^1, \dots, \gamma_k^{m_0})' = F^{-1}(Z_k), \\ &L_k = C\mathbf{1} - \gamma_k, \ \text{where } \mathbf{1} = (1, \dots, 1)'. \end{aligned} \tag{4.12}$$

Step 3. When the input u is m_0-periodic and Φ_0 is invertible, we define the estimate by $\widehat{\theta}_k = \Phi_0^{-1} L_k$.

Remark 4.6. In [44], the following result was proved. Under Assumption (A2), if the input u is m_0-periodic and Φ_0 is invertible, $\widehat{\theta}_k \to \widehat{\theta}$ w.p.1 as $k \to \infty$. Furthermore, $\|\widehat{\theta} - \theta\|_1 \le \widetilde{\eta}$, where θ is the true vector-valued parameter, and $\widetilde{\eta} > 0$ is given in Assumption (A2)(2), which represents the size of the unmodeled dynamics.

To proceed, define functions F and F^{-1} on $\mathbb{R}^{m_0}$ by $F(v) = (F(v_1), \ldots, F(v_{m_0}))'$ and $F^{-1}(v) = (F^{-1}(v_1), \ldots, F^{-1}(v_{m_0}))'$ for $v \in \mathbb{R}^{m_0}$. We now derive the large deviations principle for the identification problem. Define the function $\widehat{G} : \mathbb{R}^{m_0} \to \mathbb{R}^{m_0}$ by $\widehat{G}(x) = \Phi_0^{-1}[C\mathbf{1} - F^{-1}(x)]$ and write $\widehat{\theta}_k = \widehat{G}(Z_k)$. Since F and F^{-1} are bijection and continuous, we first study convergence rates of the sequence $\{Z_k\}$ and then apply the contraction principle, Lemma 2.5 , to derive the rate function of the estimates $\{\widehat{\theta}_k\}$. First, we find the H-functional of $\{Z_k\}$

$$\begin{aligned}
H(\tau) &= \lim_{k\to\infty} \frac{1}{k} \log E \exp\{k\langle Z_k, \tau\rangle\} \\
&= \lim_{k\to\infty} \frac{1}{k} \log E \exp\{\sum_{l=0}^{k-1}\sum_{i=1}^{m_0} \chi_{\{d(t_0+lm_0+i)\le \widetilde{c}_i\}}\tau_i\} \\
&= \log E \exp\{\sum_{i=1}^{m_0} \chi_{\{d(t_0+i)\le \widetilde{c}_i\}}\tau_i\} \\
&= \sum_{i=1}^{m_0} \log[e^{\tau_i} b_i + (1-b_i)],
\end{aligned}$$

where $b_i = P\{d(t_0) \le \widetilde{c}_i\}$. The Legendre transform of H is given by

$$I(\beta) = \sup_{\tau_1,\ldots,\tau_{m_0}} [\sum_{i=1}^{m_0} (\beta_i\tau_i - \log(e^{\tau_i} b_i + 1 - b_i))]. \tag{4.13}$$

It is easily seen that $I(\beta) = \infty$ if there exists $1 \le i \le m_0$ such that $\beta_i < 0$ or $\beta_i > 1$. For $0 \le \beta_i \le 1$, $i = 1, 2, \ldots, m_0$, denote $\overline{H}(\tau_1, \ldots, \tau_{m_0}) = \sum_{i=1}^{m_0} (\beta_i\tau_i - \log(e^{\tau_i} b_i + 1 - b_i))$. To find $I(\beta)$, setting $\frac{\partial \overline{H}(\tau_1,\ldots,\tau_{m_0})}{\partial \tau_i} = \beta_i - \frac{e^{\tau_i} b_i}{e^{\tau_i} b_i + 1 - b_i} = 0$, for $i = 1, \ldots, m_0$ leads to

$$\tau_i^* = \log \frac{\beta_i(1-b_i)}{b_i(1-\beta_i)},\ 0 \le \beta_i \le 1. \tag{4.14}$$

Substituting (4.14) into (4.13),

$$I(\beta) = \sum_{i=1}^{m_0} \log \frac{\beta_i^{\beta_i}(1-b_i)^{\beta_i-1}}{b_i^{\beta_i}(1-\beta_i)^{\beta_i-1}},\ 0 \le \beta_i \le 1.$$

To summarize,

$$I(\beta) = \begin{cases} \displaystyle\sum_{i=1}^{m_0} \log \frac{\beta_i^{\beta_i}(1-b_i)^{\beta_i-1}}{b_i^{\beta_i}(1-\beta_i)^{\beta_i-1}}, \ 0 \le \beta_i \le 1, \ i = 1, \ldots, m_0, \\ \infty, \ \text{otherwise.} \end{cases}$$

Theorem 4.7. *The above $I(\beta)$ is the rate function for the sequence $\{Z_k\}$ defined in* (4.11). *For any given open set B in $\mathbb{R}^{m_0}$,* (2.2) *holds with $\lambda_k = (1/k)$ and $X_k = Z_k$, respectively.*

Proof. Follows from a direct application of Lemma 2.4, we obtain the result. □

We now derive explicit solutions to $\inf_{\beta \in \overline{B}} I(\beta)$. Without loss of generality, we assume $i = 1$. Consider the typical application in which we are interested in $P\{|Z_k - b_1| \ge \epsilon\}$ for some small ϵ, namely, $\overline{B}_1 = (-\infty, b_1 - \epsilon] \cup [b_1 + \epsilon, \infty)$. Assume that $0 < b_1 - \epsilon < b_1 + \epsilon < 1$. Since the function

$$I(\beta_1) = \begin{cases} \log \frac{\beta_1^{\beta_1}(1-b_1)^{\beta_1-1}}{b_1^{\beta_1}(1-\beta_1)^{\beta_1-1}}, \ 0 \le \beta_1 \le 1, \\ \infty, \quad \text{otherwise} \end{cases}$$

reaches the minimum at b_1 and is monotone decreasing in $[0, b_1]$ and monotone increasing in $[b_1, 1]$, we have

$$\inf_{\beta_1 \in \overline{B_1}} I(\beta_1) = \begin{cases} \log \frac{(b_1+\epsilon)^{(b_1+\epsilon)}(1-b_1)^{b_1+\epsilon-1}}{b_1^{(b_1+\epsilon)}(1-(b_1+\epsilon))^{(b_1+\epsilon)-1}}, & b_1 \le 0.5, \\ \log \frac{(b_1-\epsilon)^{(b_1-\epsilon)}(1-b_1)^{b_1-\epsilon-1}}{b_1^{(b_1-\epsilon)}(1-(b_1-\epsilon))^{(b_1-\epsilon)-1}}, & b_1 > 0.5. \end{cases}$$

For small $\epsilon > 0$ and both $b_1 < 0.5$ and $b_1 > 0.5$, $\inf_{\beta_1 \in \overline{B}_1} I(\beta)$ can be approximated using the Taylor expansion (with respect to ϵ) by the same expression $\inf_{\beta_1 \in \overline{B_1}} I(\beta_1) = \frac{\epsilon^2}{2b_1(1-b_1)} + o(\epsilon^2)$. As a result, for small ϵ, the tail probability is dominated by $P\{|Z_k - b_1| \ge \epsilon\} \le K e^{-\frac{\epsilon^2}{2b_1(1-b_1)}k}$ for some $K > 0$. We point out that $E(s(1) - b_1)^2 = b_1(1 - b_1)$ is the variance of the binary sequence and $b_1(1-b_1)/k$ is the Cramér-Rao lower bound in terms of mean squares estimation errors of the empirical measure (see [39]). Figure 1 delineates variations on the rate function $I(\beta)$ under different values of p.

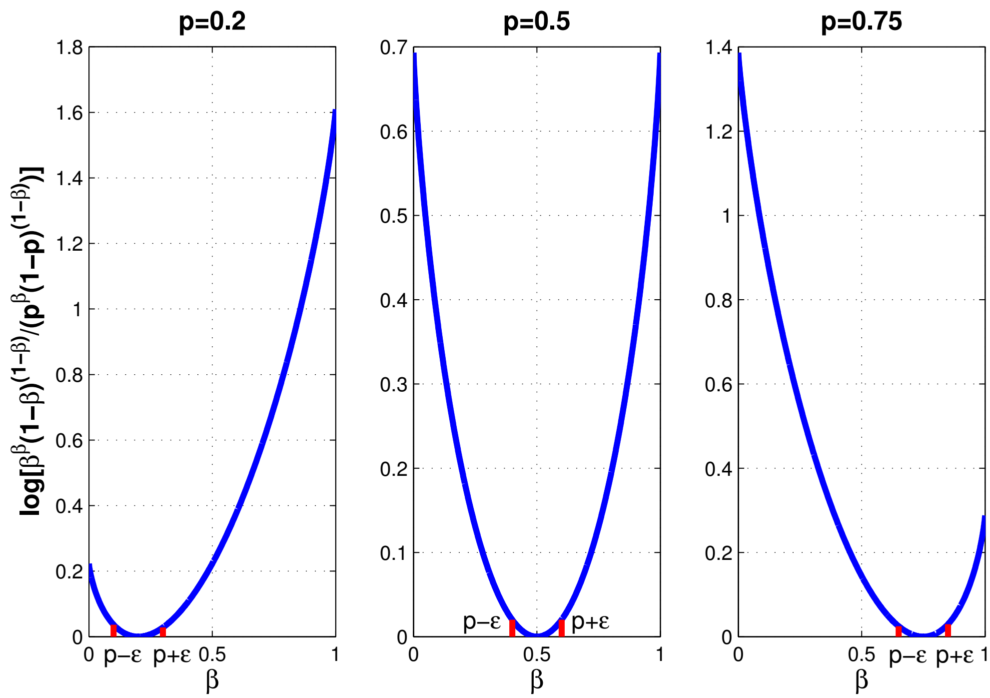

Figure 1: Variations of the rate function $I(\beta)$ with respect to different values of p

Theorem 4.8. *Under Assumption* (A2), $\widehat{I}(\widehat{\beta}) = I(\widehat{G}^{-1}(\widehat{\beta})) = I(F(C\mathbb{1} - \Phi_0\widehat{\beta}))$ *is the rate function for* $\{\widehat{\theta}_k\}$*. That is, for any open set* B *in* $\mathbb{R}^{m_0}$, (2.2) *holds with* $I(\beta)$ *replaced by* $\widehat{I}(\widehat{\beta})$, $X_k = \widehat{\theta}_k$, *and* $\lambda_k = (1/k)$, *respectively.*

Proof. Note that $\widehat{\theta}_k = \widehat{G}(Z_k)$, and $\widehat{G}$ and $\widehat{G}^{-1}$ are bijection and continuous. The result is obtained by the contraction principle Lemma 2.5.

Remark 4.9. In fact, we can write the rate function $\widehat{I}(\cdot)$ explicitly as

$$\widehat{I}(\widehat{\beta}) = \sum_{i=1}^{m_0} \log \frac{F(C\mathbb{1} - \Phi_0\widehat{\beta})_i^{F(C\mathbf{1}-\Phi_0\widehat{\beta})_i}(1-b_i)^{F(C\mathbf{1}-\Phi_0\widehat{\beta})_i - 1}}{b_i^{F(C\mathbf{1}-\Phi_0\widehat{\beta})_i}(1 - F(C\mathbb{1} - \Phi_0\widehat{\beta})_i)^{F(C\mathbf{1}-\Phi_0\widehat{\beta})_i - 1}},$$

where $F(C\mathbb{1} - \Phi_0\widehat{\beta})_i$ is the ith component of $F(C\mathbb{1} - \Phi_0\widehat{\beta})$.

To illustrate, consider the scaler case $m_0 = 1$. In this case, $F(C - \Phi_0\widehat{\beta})$ is reduced to $F(C - \widetilde{u}\widehat{\beta})$ where $\widetilde{u} \neq 0$ is a constant. Without loss of generality, assume $\widetilde{u} = 1$. Now, to obtain $P\{|\widehat{\theta}_k - \theta| \geq \epsilon\}$, we select for some small ϵ, $\overline{B}_1 = (-\infty, \theta - \epsilon] \cup [\theta + \epsilon, \infty)$. Let $\lambda = F(C - \widehat{\beta})$. Since $F(\cdot)$ is a strictly monotone increasing function, $\widehat{\beta} \leq \theta - \epsilon$ if and only if $\lambda \geq F(C - \theta + \epsilon)$ and $\widehat{\beta} \geq \theta + \epsilon$ if and only if $\lambda \leq F(C - \theta - \epsilon)$. Denote $\widetilde{B} = (-\infty, F(C - \theta - \epsilon)] \cup [F(C - \theta + \epsilon), \infty)$. It follows that

$$\begin{aligned}
&\inf_{\widehat{\beta} \in \overline{B}} \widehat{I}(\widehat{\beta}) = \inf_{\lambda \in \widetilde{B}} I(\lambda) \\
&= \begin{cases} \log \frac{F(C-\theta+\epsilon)^{F(C-\theta+\epsilon)}(1-F(C-\theta))^{F(C-\theta+\epsilon)-1}}{F(C-\theta)^{F(C-\theta+\epsilon)}(1-F(C-\theta+\epsilon))^{F(C-\theta+\epsilon)-1}}, & C - \theta \leq 0, \\ \log \frac{F(C-\theta-\epsilon)^{F(C-\theta-\epsilon)}(1-F(C-\theta))^{F(C-\theta-\epsilon)-1}}{F(C-\theta)^{F(C-\theta-\epsilon)}(1-F(C-\theta-\epsilon))^{F(C-\theta-\epsilon)-1}}, & C - \theta > 0. \end{cases}
\end{aligned}$$

For small ϵ, $g(\epsilon) := \inf_{\widehat{\beta} \in \overline{B}} \widehat{I}(\widehat{\beta})$ has $g(0) = 0$, $\dot{g}(0) = 0$, $\ddot{g}(0) = \frac{f^2(C-\theta)}{2F(C-\theta)(1-F(C-\theta))}$, where $f(x) = dF(x)/dx$ is the probability density function. As a result, it can be approximated by

$$\inf_{\widehat{\beta} \in \overline{B}} \widehat{I}(\widehat{\beta}) = \frac{f^2(C-\theta)\epsilon^2}{2F(C-\theta)(1-F(C-\theta))} + o(\epsilon^2).$$

Hence, asymptotically, the tail probability of the estimation error is dominated by

$$P\{|\theta_k - \theta| \geq \epsilon\} \leq K e^{-\frac{\epsilon^2 f^2(C-\theta)}{2F(C-\theta)(1-F(C-\theta))} k}$$

for some $K > 0$. It is noted that $F(C-\theta)(1-F(C-\theta))/(kf^2(C-\theta))$ is the Cramér-Rao lower bound (see [39]) in terms of mean squares estimation errors of $\widehat{\theta}_k$.

4.2.3 LDP of System Identification with Quantized Data

In this section, we study system identification under quantized observations or equivalently sensors with multiple thresholds. For clarity and notational simplicity, we develop our results for the case $m_0 = 1$, namely a gain system. This assumption is not restrictive. General identification problems can be reduced to a set of identification problems for gains under periodic full-rank inputs; see [44] for details. Consider the gain system given by

$$y(l) = u(l)\theta + d(l), \quad l = 1, 2, \ldots,$$

where $u(l)$ is the input and $d(l)$ is the noise. The output $y(l)$ is measured by a sensor of m thresholds $-\infty < C_1 < \cdots < C_m < \infty$. The sensor can be represented by the indicator function $s(l) = (s^1(l), \ldots, s^m(l))'$ where $s^i(l) = \chi_{\{C_{i-1}<y(l)\le C_i\}}, i = 1, \ldots, m$ with $C_0 = -\infty$ and χ_A the indicator of the set A. Without loss of generality, assume $u(l) \equiv 1$ for all l. Then $y(l) = \theta + d(l)$. Under Assumption (A2)(1), $\{y(l)\}$ is an i.i.d sequence that has the accumulative distribution function $F(\cdot)$. Let $p_i = P(C_{i-1} < y(l) \le C_i) = F(C_i-\theta)-F(C_{i-1}-\theta) := F_i(\theta)$. Consider k measurements on $s(l)$. Then $\xi_k^i = \dfrac{1}{k}\displaystyle\sum_{l=1}^{k} s^i(l)$, for $i = 1, \ldots, m$ is the sample relative frequency of $y(l)$ taking values in $(C_{i-1}, C_i]$. It follows that ξ_k^i is an unbiased estimator of p_i for each k. An estimator θ_k^i of θ can be derived from $\xi_k^i = F_i(\theta_k^i)$. Denote $G_i(x) = F_i^{-1}(x)$. Consequently, $\theta_k^i = G(\xi_k^i)$ is an estimator for θ. Define $\Theta_k = (\theta_k^1, \ldots, \theta_k^m)'$, $\xi_k = (\xi_k^1, \ldots, \xi_k^m)'$, and $G(v) = (G_1(v_1), \ldots, G_m(v_m))'$ for $v \in \mathbb{R}^m$. It was shown in [39] that $\Theta_k = G(\xi_k)$ is an asymptotically unbiased estimator of $\theta\mathbf{1}$.

We are interested in the LDP on $\Theta_k \to \theta\mathbf{1}$. The H-functional of the sequence $\{\xi_k\}$ is

$$H(\tau) = \lim_{k\to\infty} \frac{1}{k} \log E \exp\{k\langle \xi_k, \tau\rangle\}$$
$$\textstyle\sum_{l=1}^{k} s^i(l) = \log[e^{\tau_1}p_1 + e^{\tau_2}p_2 + \cdots + e^{\tau_m}p_m + 1 - \displaystyle\sum_{i=1}^{m} p_i].$$

Define $\tau = (\tau_1, \ldots, \tau_m)'$, $q(\tau) = (e^{\tau_1}, \ldots, e^{\tau_m})'$, $p = (p_1, \ldots, p_m)'$, $\beta = (\beta_1, \ldots, \beta_m)'$. Assume $p_i > 0$ for $i = 1, \ldots, m$. Also, define $p_{m+1} = 1 - \mathbf{1}'p$ and $\beta_{m+1} = 1 - \mathbf{1}'\beta$. Then $H(\tau) = \log(p'q(\tau)+p_{m+1})$. The Legendre transform of H is given by $I(\beta) = \sup_{\tau\in\mathbb{R}^m}[\langle\beta, \tau\rangle - H(\tau)] = \sup_{\tau\in\mathbb{R}^m}[\beta'\tau - \log(p'q(\tau) + p_{m+1})]$. Let $D = \{\beta : 0 \le \beta_i \le 1, \sum_{i=1}^{m} \beta_i \le 1\}$. If $\beta \notin D$, then $I(\beta) = \infty$. To see this, note that $\beta \notin D$ implies that either β has negative component or component greater than 1. Without loss of generality, let $\beta_1 < 0$. Then $I(\beta) = \infty$ by letting $\tau_i = -\infty$ for $i = 1, \ldots, m$. If $\beta_1 1$, we also get $I(\beta) = \infty$ by letting $\tau_1 = \infty$.

Next we consider the case of $\beta \in D$. To solve for the optimal τ^*, we consider the equation

$\beta - \frac{\mathcal{P}q(\tau)}{p'q(\tau)+p_{m+1}} = 0$, where $\mathcal{P} = \mathrm{diag}(p)$. Note that $\mathbf{1}'\mathcal{P} = p'$ and $p'\mathcal{P}^{-1} = \mathbf{1}'$. It follows that $q(\tau^*) = (\mathcal{P} - \beta p')^{-1}p_{m+1}\beta$. By the matrix inversion lemma,

$$\begin{aligned}(\mathcal{P} - \beta p')^{-1} &= \mathcal{P}^{-1} + \mathcal{P}^{-1}\beta(1 - p'\mathcal{P}^{-1}\beta)^{-1}p'\mathcal{P}^{-1}\\ &= \frac{1}{\beta_{m+1}}\mathcal{P}^{-1}(\beta_{m+1}I + \beta\mathbf{1}').\end{aligned}$$

This implies

$$q(\tau^*) = \frac{1}{\beta_{m+1}}\mathcal{P}^{-1}(\beta_{m+1}I + \beta\mathbf{1}')p_{m+1}\beta = \frac{p_{m+1}}{\beta_{m+1}}\mathcal{P}^{-1}\beta,$$

and $\tau^* = \log\frac{p_{m+1}}{\beta_{m+1}}\mathbf{1} + \log(\beta_1/p_1, \ldots, \beta_m/p_m)'$. In addition, $p'q(\tau^*) + p_{m+1} = \frac{p_{m+1}}{\beta_{m+1}}$. Consequently, $I(\beta)$ has the explicit expression

$$\begin{aligned}I(\beta) &= \beta'\log\frac{p_{m+1}}{\beta_{m+1}}(\beta_1/p_1, \ldots, \beta_m/p_m)' - \log\frac{p_{m+1}}{\beta_{m+1}}\\ &= \mathbf{1}'\beta\log\frac{p_{m+1}}{\beta_{m+1}} - \log\frac{p_{m+1}}{\beta_{m+1}} + \sum_{i=1}^{m}\beta_i\log\frac{\beta_i}{p_i}\\ &= \sum_{i=1}^{m+1}\beta_i\log\frac{\beta_i}{p_i}.\end{aligned}$$

To summarize,

$$I(\beta) = \begin{cases}\displaystyle\sum_{i=1}^{m+1}\beta_i\log\frac{\beta_i}{p_i}, & \beta \in D\\ \infty, & \text{otherwise.}\end{cases} \tag{4.15}$$

Remark 4.10. Note that the binary-sensor case and quantized-sensor case can also be derived from Sanov's theorem of empirical measures (see [9, Theorem 2.1.10, p.16]). The Sanov theorem states that the rate function of empirical measures is the relative entropy. In our case, it coincides with our calculation. For any probability vectors $v, \widetilde{v} \in \mathbb{R}^{m+1}$ (i.e., $v_i, \widetilde{v}_i \geq 0$ and $\sum_{i=1}^{m+1} v_i = 1$, $\sum_{i=1}^{m+1}\widetilde{v}_i = 1$), the relative entropy is defined by $H(v|\widetilde{v}) = \displaystyle\sum_{i=1}^{m+1} v_i\log(\frac{v_i}{\widetilde{v}_i})$. By the entropy form, the rate function is $I(\bar{\beta}) = H(\bar{\beta}|\bar{p})$, if we define $\bar{\beta} =$

$(\beta_1, \ldots, \beta_{m+1})'$ and $\bar{p} = (p_1, \ldots, p_{m+1})'$. In fact, for the binary case, we can write

$$\begin{aligned} I(\beta) &= \sum_{i=1}^{m_0} \log \frac{\beta_i^{\beta_i}(1-b_i)^{\beta_i - 1}}{b_i^{\beta_i}(1-\beta_i)^{\beta_i-1}} \\ &= \sum_{i=1}^{m_0} (\beta_i \log \frac{\beta_i}{b_i} + (1-\beta_i) \log \frac{1-\beta_i}{1-b_i}) = \sum_{i=1}^{m_0} H(\widetilde{\beta}_i | \widetilde{b}_i), \end{aligned}$$

if we define $\widetilde{\beta}_i = (\beta_i, 1-\beta_i)', \widetilde{b}_i = (b_i, 1-b_i)'$.

We now establish a monotonicity property in terms of the number of thresholds in quantized observations. This may be viewed as a partition property of relative entropies. Suppose that starting from the existing thresholds $\{C_1, \ldots, C_m\}$ one additional threshold $\widetilde{C}$ is added. Without loss of generality, assume $C_{i-1} < \widetilde{C} < C_i$. As a result, $p_i > 0$ is decomposed into two probabilities $p_i^1 = P(C_{i-1} < x \le \widetilde{C})$, $p_i^2 = P(\widetilde{C} < x \le C_i)$ with $p_i^1 + p_i^2 = p_i$. Excluding the trivial cases, we assume $0 < p_i^1 < p_i$. Similarly, β_i is decomposed into $\beta_i = \beta_i^1 + \beta_i^2$, $0 < \beta_i^1 < \beta_i$. This refinement on the threshold set expands $\beta = (\beta_1, \ldots, \beta_i, \ldots, \beta_{m+1})'$, $p = (p_1, \ldots, p_i, \ldots, p_{m+1})'$. to $\widetilde{\beta} = (\beta_1, \ldots, \beta_i^1, \beta_i^2, \ldots, \beta_{m+1})'$, $\widetilde{p} = (p_1, \ldots, p_i^1, p_i^2, \ldots, p_{m+1})'$.

Lemma 4.11. $I(\widetilde{\beta}) \ge I(\beta)$.

Proof: From (4.15), we need only show $f(\beta_i^1) := \beta_i^1 \log \frac{\beta_i^1}{p_i^1} + (\beta_i - \beta_i^1) \log \frac{\beta_i - \beta_i^1}{p_i - p_i^1} \ge \beta_i \log \frac{\beta_i}{p_i}$, for $0 < \beta_i^1 < \beta_i$. First, on the boundaries of $(0, \beta_i)$, $f(0) = \beta_i \log \frac{\beta_i}{p_i - p_i^1} \ge \beta_i \log \frac{\beta_i}{p_i}$; $f(\beta_i) = \beta_i \log \frac{\beta_i}{p_i^1} \ge \beta_i \log \frac{\beta_i}{p_i}$. In the interior, the stationarity condition $\frac{df(\beta_i^1)}{d\beta_i^1} = \log \frac{\beta_i^1}{p_i^1} - \log \frac{\beta_i - \beta_i^1}{p_i - {}_i^1} = 0$ leads to the stationary point $\frac{\widehat{\beta}_i^1}{p_i^1} = \frac{\beta_i}{p_i}, \frac{\beta_i - \widehat{\beta}_i^1}{p_i - p_i^1} = \frac{\beta_i}{p_i}$ and $f(\widehat{\beta}_i^1) = \beta_i \log \frac{\beta_i}{p_i}$. Since $\frac{d^2}{d\beta^2} f(\widehat{\beta}_i^1) = \frac{1}{\beta_i^1} + \frac{1}{\beta_i - \beta_i^1} > 0$, $\widehat{\beta}_i^1$ is indeed a minimum point. As a result, $\inf_{0 < \beta_i^1 < \beta_i} f(\beta_i^1) = \beta_i \log \frac{\beta_i}{p_i}$, which implies the desired inequality. $\square$

Applying the Gärtner-Ellis Theorem (Lemma 2.4 (a)), we obtain that $I(\cdot)$ is the rate function for the sequence $\{\xi_k\}$. Define $C = (C_1, \ldots, C_m)'$ and $\widehat{F}(v) = (F_1(v_1), \ldots, F_m(v_m))'$

for $v \in \mathbb{R}^m$. By virtue of the contraction principle Lemma 2.5 , the rate function of Θ_k is

$$\mathbb{I}(\widehat{\beta}) = \inf\{I(\beta) : G(\beta) = \widehat{\beta}\} = I(\widehat{F}(\widehat{\beta})). \tag{4.16}$$

Since Θ_k converges to the vector $\theta\mathbf{1}$, we may construct an estimator $\widehat{\theta}_k$ of θ by $\widehat{\theta}_k = \sum_{i=1}^m \gamma_i \theta_k^i = \gamma' \Theta_k$, where $\gamma = (\gamma_1, \ldots, \gamma_m)$ with $\gamma_1 + \cdots + \gamma_m = 1$. $\widehat{\theta}_k$ is asymptotically unbiased.

To find the LDP for the estimator $\widehat{\theta}_k$, we apply the contraction principle Lemma 2.5 to derive the rate function $I_\gamma(\beta) = \inf\{\mathbb{I}(\widehat{\beta}) : \gamma' \widehat{F}(\widehat{\beta}) = \beta\}$. For the typical case of characterizing the error probability $P\{|\widehat{\theta}_k - \theta| \geq \epsilon\}$, the set of interest is $B = (-\infty, \theta - \epsilon] \cup [\theta + \epsilon, \infty)$. To calculate $\widehat{I}(\epsilon) = \inf_{\beta \in \overline{B}} I_\gamma(\beta)$, we solve two constrained optimization problems

$$\begin{aligned}
\widehat{I}^+(\epsilon) &= \inf \sum_{i=1}^{m+1} \beta_i \log \frac{\beta_i}{p_i} \\
\text{s.t.} \quad & 0 < \beta_i < 1, i = 1, \ldots, m \\
& \sum_{i=1}^{m+1} \beta_i = 1 \quad \text{and} \quad \sum_{i=1}^{m} \gamma_i F_i(\beta_i) \geq \theta + \epsilon
\end{aligned}$$

$$\begin{aligned}
\widehat{I}^-(\epsilon) &= \inf \sum_{i=1}^{m+1} \beta_i \log \frac{\beta_i}{p_i} \\
\text{s.t.} \quad & 0 < \beta_i < 1, i = 1, \ldots, m \\
& \sum_{i=1}^{m+1} \beta_i = 1 \quad \text{and} \quad \sum_{i=1}^{m} \gamma_i F_i(\beta_i) \leq \theta - \epsilon.
\end{aligned}$$

Then, $\widehat{I}(\epsilon) = \min\{\widehat{I}^+(\epsilon), \widehat{I}^-(\epsilon)\}$. Although we do not expect to obtain closed-form solutions generally, the above constrained optimization problem characterizes the desired solution.

4.3 Examples and Discussions

We will present several simulation examples to demonstrate the large deviations principle for the estimates and illustrate complexity relationship among binary, quantized, and regular sensors by showing the monotonicity property of the corresponding rate functions.

Binary Sensors. We shall start with system identification under binary sensors. In this example, $\theta' = (1.75, 1.75, 2.75)$ is the true parameter vector and for simplicity assume $\widetilde{\theta} = 0$, i.e., no unmodelled dynamics. Select a 3-periodic input with one period values $(u(1), u(2), u(3)) = (3, 4, 5)$, which is full rank, and $\Phi_0 = \begin{pmatrix} 3 & 4 & 5 \\ 4 & 5 & 3 \\ 5 & 3 & 4 \end{pmatrix}$. The noise is an i.i.d. sequence of random variables with the standard normal distribution. The binary sensor has a threshold $C = 25$. For $\epsilon = 1$, we compare empirical measures of $P(|\widehat{\theta}_k - \theta| > \epsilon)$, with the calculated large deviations principle bound, $\exp\left(-k \inf_{|x-\theta|>\epsilon} I(x)\right)$.

Step 1. For each $k = 5, \ldots, 50$, the identification algorithm is repeated 1000 times and the sample frequencies of the event $|\widehat{\theta}_k - \theta| > \epsilon$ are then calculated as an approximation to $P(|\widehat{\theta}_k - \theta| > \epsilon)$.

Step 2. From Remark 4.9, compute

$$\widehat{I}(\widehat{\beta}) = \sum_{i=1}^{m_0} \log \frac{F(C - \Phi_0\widehat{\beta})_i^{F(C-\Phi_0\widehat{\beta})_i} (b_i - 1)^{F(C-\Phi_0\widehat{\beta})_i - 1}}{b_i^{F(C-\Phi_0\widehat{\beta})_i} (F(C - \Phi_0\widehat{\beta})_i - 1)^{1-F(C-\Phi_0\widehat{\beta})_i}},$$

where $F(C - \Phi_0\widehat{\beta})_i$ is the ith component of $F(C - \Phi_0\widehat{\beta})$ and $b_i = P\{d(l) \le C - (\Phi_0\theta)_i\}$. It is calculated $\inf_{|x-\theta|>\epsilon} I(x) = -0.16$, and the LDP exponential decaying curve is $\exp(-0.16k)$. These are shown in Figure 2. For sufficiently large k, these two curve match very well.

Regular Sensors. We consider the same example of system identification under regular sensor with $\epsilon = 0.5$. In this case we aim to find the convergence rate of

$$\widehat{\theta}_k = \theta + \eta_s^k = \theta + \Phi_0^{-1}(U_k^1, U_k^2, U_k^3)',$$

where $U_k^j = \dfrac{1}{k}\sum_{l=0}^{k-1} d(t_0 + lm_0 + j)$, for $j = 1, 2, 3$ and Φ_0 is as in the last example.

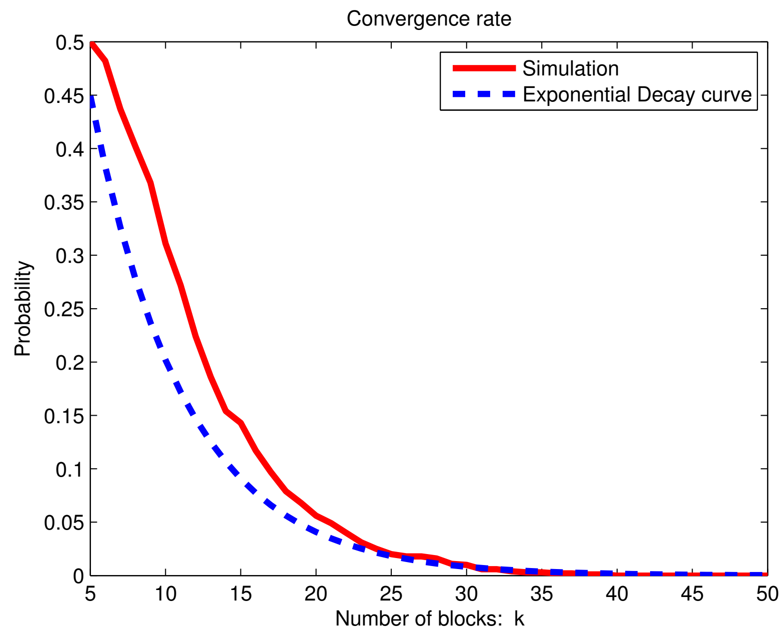

Figure 2: Comparison of the empirical errors and the LDP bound under a binary sensor.

By virtue of Theorem 4.4, the rate function is $\widetilde{I}(\widetilde{\beta}) = I(G^{-1}(\widetilde{\beta})) = I(\Phi_0\widetilde{\beta} - \theta)$, where $I(\beta) = \sup_{\tau_1,\ldots,\tau_{m_0}} [\sum_{i=1}^{m_0}(\beta_i\tau_i - \log g(\tau_i))]$, where $g(\cdot)$ is the moment generating function given in (A1). Since the noise d_l is a standard normal random variable, we get $I(\beta) = \dfrac{|\beta-\theta|^2}{2}$ by Remark 4.5. Hence, $\widetilde{I}(\widetilde{\beta}) = I(G^{-1}(\widetilde{\beta})) = \dfrac{|\Phi_0\widetilde{\beta}-\theta|^2}{2}$. We simulate the probability $P(|\widehat{\theta}_k - \theta| > 0.5)$, and then compare with the large deviations result, in which this probability is approximated by $K\exp\left(-k \inf_{|x|>1} \widetilde{I}(x)\right)$ for some $K > 0$.

Step 1. For each $k = 5, \ldots, 100$, taking 1000 samples and use the proportion of the number of times of $|\widehat{\theta}_k - \theta| > 0.5$ to approximate $P(|\widehat{\theta}_k - \theta| > 0.5)$.

Step 2. By calculating the value $\inf_{|x-\theta|>0.5} \widetilde{I}(x) = \inf_{|x-\theta|>0.5} \dfrac{|\Phi_0(x-\theta)|^2}{2} = 0.375$, the exponential decay curve is $\exp(-0.375k)$.

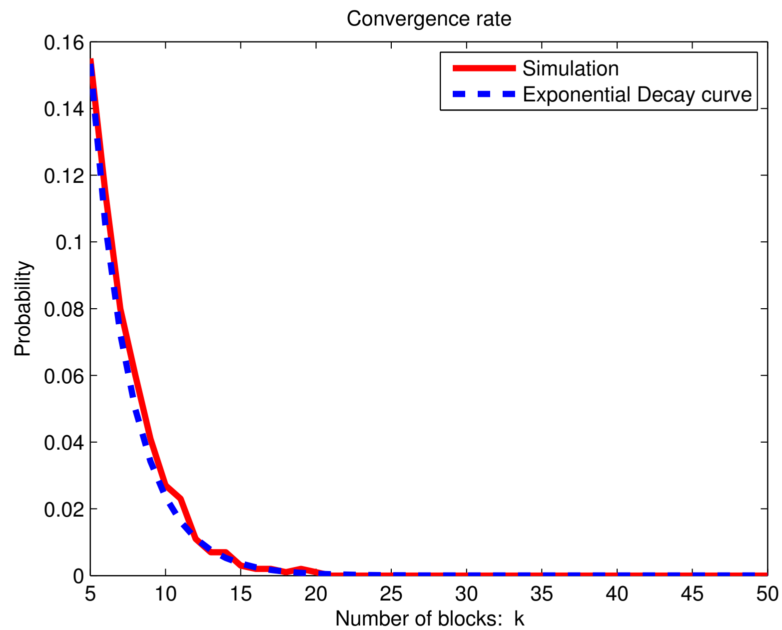

Figure 3: Comparison of the empirical errors and the LDP bound under a regular sensor.

4.3.1 Space Complexity: Monotonicity of Rate Functions with respect to Numbers of Sensor Thresholds.

The LDP indicates that $P(|\widehat{\theta}_k - \theta| \geq \epsilon) \leq K \exp(-\inf_{|\beta-\theta|\geq\epsilon} I_0(\beta)k)$, for some $K > 0$ where I_0 is the rate function depending on the sensor types. Comparison of the rates functions under different sensor types will demonstrate complexity and benefits relationships in sensor design. Let $y(l) = \theta + d(l)$ for $l = 1, 2, \ldots,$ where $d(l)$ is the i.i.d. noise with the standard normal distribution. We assume that $\theta = 3$, $C_1 = 2$, and $C_2 = 4$. We want to find the probability $P(|\widehat{\theta}_k - \theta| > 1)$.

a. Observations under Binary Sensors. Under a binary sensor of threshold $C_1 = 2$, we have the estimator $\widehat{\theta}_k^b = C_1 - F^{-1}(\xi_k^1)$, where $\xi_k^1 = \dfrac{\sum_{l=1}^k \chi_{\{d(l)\leq C_1-\theta\}}}{k}$. By Remark 4.9, $\widehat{\theta}_k^b$ has the rate function

$$I^b(\beta) = \log \frac{F(C_1-\beta)^{F(C_1-\beta)}(b_1-1)^{F(C_1-\beta)-1}}{b_1^{F(C_1-\beta)}(F(C_1-\beta)-1)^{1-F(C_1-\beta)}},$$

where $b_1 = P(d(l) \le C_1 - \theta)$. Hence the estimator $\Theta_k^b = (\widehat{\theta}_k^b, \widehat{\theta}_k^b)'$ has the rate function $\mathbb{I}^b(\widehat{\beta}) = \inf\{I^b(\beta) : (\beta, \beta)' = \widehat{\beta}\}$. Applying the LDP and the contraction principle Lemma 2.5 , with $\Theta = (\theta, \theta)'$, we obtain

$$\begin{aligned}\lim_{k\to\infty} \frac{1}{k} \log P(|\Theta_k^b - \Theta| > 1) &= -\inf_{|\widehat{\beta}-\Theta|>1} \mathbb{I}^b(\widehat{\beta}) \\ &= -\inf_{(\beta-3)^2>\frac{1}{2}} I^b(\beta) = -0.0658.\end{aligned}$$

b. Observations under Quantized Sensors. Consider a quantized sensor with two thresholds $C_1 = 2$ and $C_2 = 4$. The estimator for $\Theta = (3,3)'$ is $\Theta_k^q = (C_1 - F^{-1}(\xi_k^1), C_2 - F^{-1}(\xi_k^2))'$, where $\xi_k^i = \sum_{l=1}^k \chi_{\{d(l)\le C_i-\theta\}}/k$ for $i = 1, 2$. By (4.16), the rate function for Θ_k^q is

$$\begin{aligned}\mathbb{I}^q((\beta_1, \beta_2)') = \sup_{\tau_1,\tau_2} [\tau_1 F(C_1 - \beta_1) \\ +\tau_2 F(C_2 - \beta_2) - H(\tau_1, \tau_2)],\end{aligned}$$

where $H(\tau_1, \tau_2) = \log[e^{\tau_1+\tau_2} p_1 + e^{\tau_2}(p_2 - p_1) + (1 - p_2)]$, and $p_i = P\{d(l) \le C_i - \theta\}, i = 1, 2$. Applying the large deviations principle, we obtain $\lim_{k\to\infty} \log P(|\Theta_k^q - \Theta| > 1)/k = -\inf_{|\widehat{\beta}-\Theta|>1} \mathbb{I}^q(\widehat{\beta}) = -0.1014$.

c. Observations under Regular Sensors. When we use regular sensors, the estimator of θ is given by $\widehat{\theta}_k^r = \frac{\sum_{l=1}^k y_l}{k}$. By Remark 4.5, the rate function of $\widehat{\theta}_k^r$ is $I^r(\beta) = \dfrac{(\beta-3)^2}{2}$. Hence the rate function of the two dimensional estimator $\Theta_k^r = (\widehat{\theta}_k^r, \widehat{\theta}_k^r)'$ is $\mathbb{I}^r(\widehat{\beta}) = \inf\{I^r(\beta) : (\beta, \beta)' = \widehat{\beta}\}$. Applying the LDP and contraction principle Lemma 2.5 , we have

$$\begin{aligned}\lim_{k\to\infty} \frac{1}{k} \log P(|\Theta_k^r - \Theta| > 1) &= -\inf_{|\widehat{\beta}-\Theta|>1} \mathbb{I}^r(\widehat{\beta}) \\ &= -\inf_{(\beta-3)^2>\frac{1}{2}} \frac{(\beta-3)^2}{2} = -0.25.\end{aligned}$$

From the above discussions on the three different sensors, it is clear that there is a monotonicity of the rate functions when the sensor complexity increases, which can be summarized as

$$P(|\theta_k - \theta| > 1) = \begin{cases} \exp(-0.0658k) \text{ binary sensor,} \\ \exp(-0.1014k) \text{ quantized sensor,} \\ \exp(-0.25k) \text{ regular sensor,} \end{cases} \tag{4.17}$$

where the quantized sensor has two threshold values. Figure 4 displays the comparison results. In view of the study in [42], the two-threshold quantized sensor design is a refinement

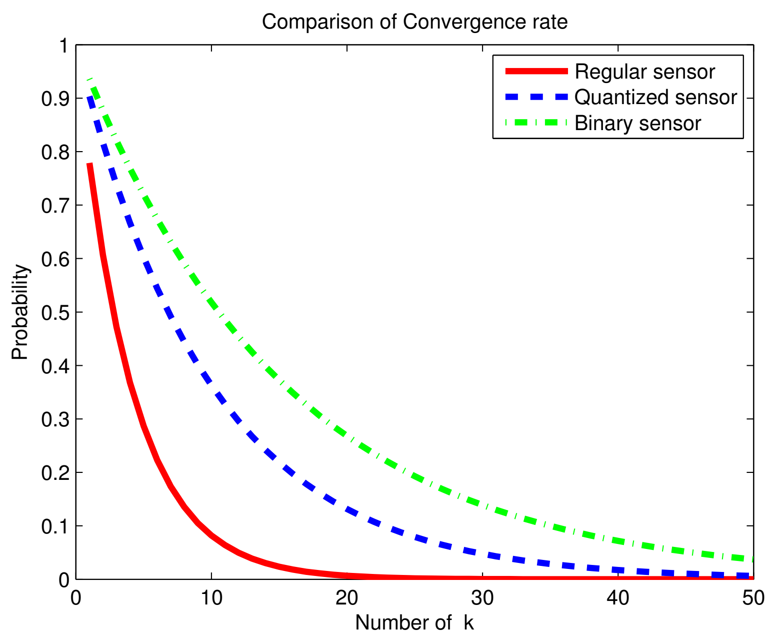

Figure 4: Comparison of convergence rates among different sensors.

of the binary sensor case, and the regular sensor is an infinite refinement of quantized sensors. The large deviations rate functions give precise descriptions on convergence rates, hence can be used in selecting sensor complexity levels.

4.4 LDP of System Identification under Mixing Noises

Up to this point, the random sequences considered are uncorrelated, i.e., white noises. In this section, we demonstrate that a much larger class of noise processes can be treated.

LDP for Empirical Means under ϕ-Mixing Conditions. It is natural to consider noise sequence $\{d(l)\}$ under mixing conditions. In this section we consider stationary ϕ-mixing random sequences with sufficiently fast convergence rates. Let $\{X_k\}$ be a stationary sequence. Denote by $\mathcal{F}_a^b$ the σ-algebra generated by $\{X_k : a \leq k \leq b\}$, and let $\phi(k) = \sup\{|P(B|A) - P(B)| : A \in \mathcal{F}_0^l, P(A) > 0, B \in \mathcal{F}_{k+l}^\infty, l \in Z_+\}$. $\{X_k\}$ is said to be ϕ-mixing if $\phi(k) \to 0$ as $k \to \infty$. Throughout this section we will need the following assumption.

(A3) (a) $\{X_k\}$ is a stationary ϕ-mixing random sequence and $\phi(k)$ satisfies $\phi(k) \leq \exp(-kr(k))$, where $r(k) \to \infty$ as $k \to \infty$, and $\displaystyle\sum_{k=1}^{\infty} \frac{r(k)}{k(k+1)} < \infty$. (b) $\{X_k\}$ takes values in a compact set $K_c \subset \mathbb{R}^d$.

Mixing processes are those whose remote past and distant future are asymptotically independent. For general reference on mixing processes, we refer the reader to [**?**]. Under Assumption (A3), we derive the following Theorem.

Theorem 4.12. *If $\{X_k\}$ is a stationary ϕ-mixing sequence satisfying* (A3)*, then*

$$\widehat{S}_k = \frac{X_1 + X_2 + \cdots + X_k}{k}, \;\; k \geq 1,$$

satisfies the LDP. That is, there is a rate function $I : \mathbb{R}^d \to [0, \infty]$ that is convex and lower semicontinuous, and that for any $B \subset \mathbb{R}^d$,

$$\begin{aligned} -\inf_{\gamma \in B^\circ} I(\gamma) &\leq \liminf_{k \to \infty} \frac{1}{n} \log P\{\widehat{S}_k \in B\} \\ &\leq \limsup_{k \to \infty} \frac{1}{k} \log P\{\widehat{S}_k \in B\} \leq -\inf_{\gamma \in \overline{B}} I(\gamma), \end{aligned}$$

where B° *and* $\overline{B}$ *denote the interior and closure of* B*, respectively. Moreover, the rate function is given as*

$$I(\gamma) = \sup_{\beta \in \mathbb{R}^d} [\langle \gamma, \beta \rangle - \Lambda(\beta)], \tag{4.18}$$

where $\Lambda(\beta) = \lim_{k\to\infty} k^{-1} \log E \exp(k\langle \beta, \widehat{S}_k \rangle)$.

Remark 4.13. Dealing with empirical measures, a large deviations principle for a class of stationary processes under certain mixing conditions was proved in [9]. Their condition is different from ours and is implied by the so-called ψ-mixing processes. The large deviations principle for arithmetic means of a ϕ-mixing process under Assumption (A3) without $\displaystyle\sum_{k=1}^{\infty} \frac{r(k)}{k(k+1)} < \infty$ was proved in [6]. However, the proof is a bit complicated. Here, we use a similar approach as that of [9] to derive an alternative proof of the LDP for a sample mean under ϕ-mixing assumptions. Our large deviations for the identification is then based on the large deviations for the sample mean. In order to prove the desired result, we need a number of preparatory results. They are stated in the following proposition. The first part in the proposition is the approximate sub-additivity ([9, Lemma 6.4.10, p.282]), and the second and the third parts are in [9, Lemma 4.4.8, Theorem 4.4.10, p.143], respectively.

Proposition 4.14. *The following results hold.*

(a) *Let* $h : \mathbb{N} \to \mathbb{R}$ *and assume that all* $k, l \geq 1$*,* $h(k+l) \leq h(k) + h(l) + \epsilon(k+l)$*, where* $\epsilon(k)$ *is a non-decreasing sequence satisfying* $\sum_{k=1}^{\infty} \frac{\epsilon(k)}{k(k+1)} < \infty$*. Then* $\displaystyle\lim_{k\to\infty} \frac{h(k)}{k}$ *exists and is finite.*

(b) *Let* $\{X_k\}$ *be a sequence of random variables taking values in a compact subset* $K_c \subset \mathbb{R}^d$*, and for any concave, bounded above, and continuous function* g*,* $\Lambda_g = \lim_{k\to\infty} \log E e^{kg(X_k)}$.

Then Λ_f *exists for all* f *belongs to* $C_b(\mathbb{R}^d)$*, which is the space of all bounded and continuous function on* $\mathbb{R}^d$*. Furthermore,* $\{X_k\}$ *satisfies the LDP with the rate function* $I(x) = \sup_{f\in C_b(\mathbb{R}^d)}[f(x) - \Lambda_f]$.

(c) *Assume that* $\{X_k\}$ *satisfies the LDP with a rate function* $I(\cdot)$ *that is convex, and* $\limsup_{k\to\infty} \frac{1}{k} \log Ee^{k\langle \lambda, X_k\rangle} < \infty, \forall\lambda \in \mathbb{R}^d$*. Then* $I(x) = \sup_{\lambda\in\mathbb{R}^d}[\langle\lambda, x\rangle - \Lambda(\lambda)]$*, where* $\Lambda(\lambda) = \lim_{k\to\infty} \frac{1}{k} \log Ee^{k\langle \lambda, X_k\rangle}$.

With the proposition above, we are ready to prove Theorem 4.12.

Proof of Theorem 4.12**.** We carry out the proof by adapting and modifying the proof of [9, Theorem 6.4.4,p.279]. Choose concave continuous function $g : \mathbb{R}^d \to [-B, 0]$. Since X_k take values on a compact subset $K_c \subset \mathbb{R}^d$, g is Lipschitz continuous on K_c. Assume $|g(x) - g(y)| < G|x - y|$ for all $x, y \in K$. Denote $C = \sup_{x\in K_c} g(x)$, and

$$\widehat{S}_k^m = \frac{X_{m+1} + \cdots + X_k}{k - m}.$$

Observe that

$$|\widehat{S}_{k+m} - (\frac{k}{k+m}\widehat{S}_k + \frac{m}{k+m}\widehat{S}_{k+m+l}^{k+l})| \le \frac{2lC}{k+m}.$$

By Lipschitz continuity of g,

$$|g(\widehat{S}_{k+m}) - g(\frac{k}{k+m}\widehat{S}_{m_0} + \frac{m}{k+m}\widehat{S}_{k+m+l}^{k+l})| \le \frac{2lCG}{k+m}.$$

Since g is concave,

$$(m + k)g(\widehat{S}_{m+k}) \ge kg(\widehat{S}_k) + mg(\widehat{S}_{k+m+l}^{k+l}) - 2lCG. \tag{4.19}$$

Denote $W = e^{kg(\widehat{S}_k)}, Z = e^{mg(\widehat{S}_{k+m+l}^{k+l})}$. By virtue of Assumption (A3),

$$EWEZ - EWZ \leq e^{-r(l)l}. \tag{4.20}$$

Noting that $EWEZ \geq -(m+k)B$ and let $l = m+k$, from (4.20),

$$\log EWZ \geq \log EWEZ + \log(e^{(B-r(m+k))(m+k)}). \tag{4.21}$$

Since $r(k+m) \to \infty$ as $m, k \to \infty$, we can choose m, k large enough such that $\log EWZ \geq \log EWEZ + \frac{1}{2}$. Let $h(k) = \log Ee^{kg(\widehat{S}_k)}$, by the above inequality and (4.19)

$$h(m+k) \leq h(m) + h(k) + \frac{1}{2} + 2r(m+k)CG. \tag{4.22}$$

Applying Assumption (A3), $\Lambda_g = \lim_{k\to\infty} \frac{1}{k} \log Ee^{kg(X_k)}$ exists and hence by Proposition 4.14 (a), $\{\widehat{S}_k\}$ satisfies the LDP with the rate function $I(x) = \sup_{f \in C_b(\mathbb{R}^d)}[f(x) - \Lambda_f]$. Next, we need to show that $I(\cdot)$ is convex to reduce the form of $I(\cdot)$ to (4.18) by means of Proposition 4.14 (c). To show that $I(\cdot)$ is convex, we need only show that if for some $M < \infty$ and fix open sets G and $\widetilde{G}$ $P(\widehat{S}_k \in G)P(\widehat{S}_k \in \widetilde{G}) \geq \exp(-Mk)$ for all k large enough, then

$$\liminf_{\eta \downarrow 0} \liminf_{k\to\infty} \rho(k, \eta) \geq 0, \tag{4.23}$$

where

$$\rho(k, \eta) = \frac{1}{k} \log \frac{P(\widehat{S}_k \in G, \widehat{S}^{k+\eta k}_{2k+\eta k} \in \widetilde{G})}{P(\widehat{S}_k \in G)P(\widehat{S}^{k+\eta k}_{2k+\eta k} \in \widetilde{G})}.$$

Applying Assumption (A3) again,

$$\frac{P(\widehat{S}_k \in G, \widehat{S}^{k+\eta k}_{2k+\eta k} \in \widetilde{G})}{P(\widehat{S}_k \in G)P(\widehat{S}^{k+\eta k}_{2k+\eta k} \in \widetilde{G})} \geq 1 - e^{(M - r(k\eta)k\eta)}. \tag{4.24}$$

Thus

$$\begin{aligned} &\liminf_{\eta \downarrow 0} \liminf_{k\to\infty} \rho(k, \eta) \\ &\quad \geq \liminf_{\eta \downarrow 0} \liminf_{k\to\infty} \frac{1}{k} \log(1 - e^{(M - r(k\eta)k\eta)}) = 0. \end{aligned} \tag{4.25}$$

The proof of the theorem is completed. □

LDP for System Identification with Regular Sensors under Mixing Noises. Applying Theorem 4.12 directly to the identification with regular sensors.

Theorem 4.15. *Assume that the noise sequence* $\{d(l)\}$ *satisfies Assumption* (A3). *Then* $\{\eta_k^s\}$ *satisfies the LDP with the rate function* $\widetilde{I}(\widetilde{\beta}) = I(G^{-1}(\widetilde{\beta})) = I(\Phi_{m_0}(t_0)\widetilde{\beta}))$, *where*

$$\begin{aligned} I(\beta) &= \sup_{\lambda \in \mathbb{R}^d} [\langle \lambda, \beta \rangle - \Lambda(\lambda)], \\ \Lambda(\lambda) &= \lim_{k\to\infty} \log Ee^{k\langle \lambda, X_k \rangle}. \end{aligned} \tag{4.26}$$

Proof. Let $Y_i = (d(t_0+(i-1)m_0+1), \dots, d(t_0+(i-1)m_0+m_0))'$. Then $X_k = (Y_1+\cdots+Y_k)/k$. Since $\{d_i\}$ satisfies Assumption (A3), so is $\{Y_i\}$. By theorem 4.12, $\{X_k\}$ satisfies the LDP with rate function I. Then by the contraction principle Lemma 2.5, $\{\eta_k^s\}$ satisfies the LDP with the rate function $\widetilde{I}(\widetilde{\beta}) = I(G^{-1}(\widetilde{\beta})) = I(\Phi_0\widetilde{\beta})$. The proof is completed. □

LDP for Identification with Binary Sensors under Mixing Conditions. Recall the notation in Section 4.1 and apply theorem 4.12.

Theorem 4.16. *Assume that the noise sequence* $\{d(l)\}$ *satisfies the first part of Assumption* (A3). *Then* $\{\widehat{\theta}_k\}$ *satisfies the LDP with the rate function* $\widehat{I}(\widehat{\beta}) = I(F(C - \Phi_0\widehat{\beta}))$, *where* $I(\beta)$ *and* $\Lambda(\lambda)$ *are given by* (4.26).

Proof. Let

$$Y_i = (\chi_{\{d(t_0+(i-1)m_0+1)\}}, \dots, \chi_{\{d(t_0+(i-1)m_0+m_0)\}})'.$$

Then $X_k = (Y_1 + \cdots + Y_k)/k$. Since $\{d(l)\}$ satisfies the first part of Assumption (A3) and $|Y_i| \le m_0$, $\{Y_i\}$ satisfies Assumption (A3). By Theorem 4.12 and the fact that $\{X_k\}$ satisfies the LDP with rate function I. By virtue of the contraction principle Lemma 2.5, $\{\widehat{\theta}_k\}$ satisfies the LDP with rate function $\widehat{I}(\widehat{\beta}) = I(F(C - \Phi_0\widehat{\beta}))$. The proof is concluded. $\square$

4.5 Discussions on Non-Periodic Inputs

For clarity, this brief assumes that the input is designed to be periodic. Advantages of using full rank and periodic inputs in quantized identification problems have been extensively discussed in [43]. However, such a choice is not a fundamental limitation and non-periodic inputs can also be used. As a first attempt of using LDPs in complexity analysis, we will not explore this direction in detail here, but only highlight the main steps. To illustrate the basic ideas and computational processes for quantized identification under non-periodic inputs, we should use the basic case of a binary sensor with threshold C and gain identification problems. In this case, the system is $y(t) = \theta u(t) + d(t)$, for $t = 0, 1, \ldots,$ and $s(t) = I_{\{y(t) \le C\}}$, where θ is to be identified. The empirical measure on the basis of k measurements is

$$\xi_k = \frac{1}{k}\sum_{t=0}^{k-1} s(t).$$

Denote

$$\begin{aligned} G_k(\theta) &= E\xi_k = \frac{1}{k}\sum_{t=0}^{k-1} \\ Es(t) &= \frac{1}{k}\sum_{t=0}^{k-1} F(C - \theta u(t)). \end{aligned}$$

Then, the estimate θ_k of θ is $\theta_k = G_k^{-1}(\xi_k)$. Although explicit expressions for $G(\cdot)$ and $G^{-1}(\cdot)$ may be difficult to obtain, its numerical solutions are straightforward. Note that

$$\begin{aligned} \epsilon_k &= \xi_k - E\xi_k = \frac{1}{k}\sum_{t=0}^{k-1}(s(t) - F(C - \theta u(t))) \quad \text{and} \\ e_k &= \theta_k - \theta = G^{-1}(\xi_k) - G^{-1}(E\xi_k), \end{aligned}$$

which may be approximated for small ϵ_k by

$$\begin{aligned}
e_k &\approx \left(\frac{\partial G(\theta)}{\partial \theta}\right)^{-1} \epsilon_k := g_k(\theta)\epsilon_k, \quad \text{where} \\
\frac{\partial G_k(\theta)}{\partial \theta} &= \frac{1}{k}\sum_{t=0}^{k-1} \frac{\partial F(C-\theta u(t))}{\partial \theta} \\
&= \frac{1}{k}\sum_{t=0}^{k-1} \widetilde{v}(y) \\
\widetilde{v}(t) &= -u(t) f(C-\theta u(t)).
\end{aligned}$$

Since the LDP is preserved by a continuous mapping, we need only concentrate on ϵ_k (and whose rate function can be easily modified by $g_k(\theta)$ to obtain the rate function of e_k). However, ϵ_k is a convex combination of independent variables. Hence, the rate function of ϵ_k can be obtained from these of $s(t) - F(C-\theta u(t))$. When $u(t)$ is periodic (in this special case, it is a constant), these variables become also identically distributed, rendering a substantial simplification on computation. This, however, does not alter the fundamental properties of LDPs and rate function expressions. In the special case of uniform distributions, more explicit expressions and much easier computations can be obtained. Indeed, if $d(t)$ is i.i.d., uniformly distributed in $[-\delta, \delta]$, we have $F(x) = \frac{1}{2\delta}(x+\delta)$ and $f(x) = \frac{1}{2\delta}, x \in [-\delta, \delta]$. Consequently, assuming $C - \theta u(t) \in [-\delta, \delta]$ for all t, we have with $\overline{u}_k = \sum_{t=0}^{k-1} u(t)/k$,

$$\begin{aligned}
G_k(\theta) &= \frac{1}{k}\sum_{t=0}^{k-1} \frac{1}{2\delta}(C-\theta u(t)+\delta) \\
&= \frac{C+\delta}{2\delta} - \frac{\frac{1}{k}\sum_{t=0}^{k-1} u(t)\overline{u}_k}{2\delta}\theta, \\
\theta_k &= \frac{\frac{C+\delta}{2\delta} - \xi_k}{\frac{\overline{u}_k \frac{1}{k}\sum_{t=0}^{k-1} u(t)}{2\delta}}, \quad \frac{\partial G_k(\theta)}{\partial \theta} \\
&= -\frac{1}{2\delta}\frac{1}{k}\sum_{t=0}^{k-1} u(t)\overline{u}_k, \\
e_k &\approx -\frac{2\delta}{\overline{u}_k \frac{1}{k}\sum_{t=0}^{k-1} u(t)}\epsilon_k.
\end{aligned}$$

Consequently, the rate function computation becomes straightforward.

4.6 Escape from A Domain

In this section, we concentrate on the identification problem with binary observations. Different from the study in the previous sections, rather than dealing with the discrete processes directly, we take a continuous-time interpolation. Thus here we provide an alternative view point for the study of large deviations on the parameter estimates. Consider the algorithm

$$Z_k^i = \frac{1}{k} \sum_{l=0}^{k-1} \chi_{\{y(lm_0+i) \le C\}}. \tag{4.27}$$

For ease of discussion, assume that there is no unmodeled dynamics. Thus, $Y_l = \Phi_0 \theta + D_l$. Denote $\chi_k^i = \chi_{\{d(km_0+i) \le C - (\Phi_0\theta)_i\}}$ and

$$\begin{aligned} S_k &= (\chi_k^1, \ldots, \chi_k^{m_0})', \\ Z^* &= (F(C - (\Phi_0\theta)_1), \ldots, F(C - (\Phi_0\theta)_{m_0}))', \end{aligned} \tag{4.28}$$

and define $Z_k = (Z_k^1, \ldots, Z_k^{m_0})'$. We can then write (4.27) recursively as

$$Z_{k+1} = Z_k - \frac{1}{k+1} Z_k + \frac{1}{k+1} S_k \begin{bmatrix} \chi_{\{d(km_0+1) \le C - (\Phi_0\theta)_1\}} \\ \vdots \\ \chi_{\{d(km_0+m_0) \le C - (\Phi_0\theta)_{m_0}\}} \end{bmatrix}. \tag{4.29}$$

Rather than directly analyzing the sequence of iterates, we use the method of ordinary differential equations for stochastic approximation; see Kushner and Yin [28, Chapters 5 and 6]. Define $t_k = \sum_{l=0}^{k-1} 1/(l+1)$ and $m(t) = \max\{k : t_k \le t\}$, where t_k connect the discrete iteration number with continuous time, and $m(t)$ serves as its inverse. Define the piecewise constant interpolation

$$Z^{\{0\}}(t) = Z_k \quad \text{for} \quad t \in [t_k, t_{k+1}) \quad \text{and} \quad Z^{\{k\}}(t) = Z^{\{0\}}(t + t_k).$$

Suppose for simplicity, $\{d_k\}$ is an i.i.d. sequence with zero mean and finite variance. We can verify that w.p.1, $Z^{\{k\}}(\cdot)$ is uniformly bounded and equicontinuous in the extended sense (see

[28, p. 102]). The Arzela-Ascoli Theorem [28, p. 102] yields that any convergent subsequence of $Z^{\{k\}}(\cdot)$ has a limit $Z(\cdot)$, which is a solution of the ordinary differential equation

$$\dot{Z}(t) = -Z(t) + Z^* \begin{bmatrix} F(C - (\Phi_0\theta)_1) \\ \vdots \\ F(C - (\Phi_0\theta)_{m_0}) \end{bmatrix}.. \tag{4.30}$$

Moreover, as $k \to \infty$, $Z^{\{k\}}(\cdot + T_k) \to Z^*$ w.p.1, where $\{T_k\}$ is any sequence of positive real numbers satisfying $T_k \to \infty$.

We proceed to study the asymptotic properties of $Z^{\{k\}}(\cdot)$. Of particular interest are estimates of the probabilities of $Z^{\{k\}}(\cdot)$ escape from a fixed neighborhood of the stable point Z^* of (4.30). To be precise, let G be a neighborhood of Z^* and define

$$\widetilde{\tau}_G^k = \min\{t : Z^{\{k\}}(t) \notin G\}.$$

By the w.p.1 convergence, the probability $P_x^k\{\widetilde{\tau}_G^k \le T\}$ tends to zero as $k \to \infty$, and it is natural to look for the rate of convergence. In particular, we seek a sequence $\lambda_k \to 0$ and $0 < V_2 < V_1 < \infty$ such that the following limit exists,

$$\begin{aligned} -V_1 &\le \liminf_{k\to\infty} \lambda_k \log P_x^k\{\widetilde{\tau}_G^k \le T\} \\ &\le \limsup_{k\to\infty} \lambda_k \log P_x^k\{\widetilde{\tau}_G^k \le T\} \\ &\le -V_2, \end{aligned}$$

where P_x^k denotes the probability conditioned on the event that $Z^{\{k\}}(0) = x \in G$. To this end, we need the following assumption.

Assume that there exists a function $H(\cdot,\cdot,\cdot) : \mathbb{R}^{m_0} \times \mathbb{R}^{m_0} \times [0,T] \to \mathbb{R}$ and a sequence $\lambda_k \to 0$ such that for any $x \in \mathbb{R}^{m_0}$ and piecewise constant function $\alpha(\cdot) : [0,T] \to \mathbb{R}^{m_0}$,

$$\int_0^T H(x, \alpha(s), s)ds = \lim_{k\to\infty} \lambda_k \log E \exp(\frac{1}{\lambda_k}\widetilde{\Pi}_k), \tag{4.31}$$

where

$$\widetilde{\Pi}_k = \sum_{i=0}^{N-1} \left\langle \alpha'(i\Delta), \sum_{j=m(t_k+t)}^{m(t_k+i\Delta+\Delta-1)} \frac{1}{j+1}(-x+S_j) \right\rangle$$

with S_j defined in (4.28) and $T = N\Delta$, and $\alpha(\cdot)$ is constant on the interval $[k\Delta, (k+1)\Delta)$. Let $C_x[0,T]$ denote the space of $\mathbb{R}^{m_0}$-valued continuous function on $[0,T]$ with initial value x, and with uniform convergence topology. Denote the interior and closure of set $A \subset C_x[0,T]$ by A° and $\overline{A}$, respectively. Then by [27, Theorem 1] we have the large deviations estimate.

Theorem 4.17. *Under the above assumption, for each set $A \subset C_x[0,T]$,*

$$\begin{aligned} -\inf\nolimits_{\varphi \in A^\circ} S(T,\varphi) &\leq \liminf_{k\to\infty} \lambda_k \log P_x\{Z^{\{k\}}(\cdot) \in A\} \\ &\leq \limsup_{k\to\infty} \lambda_k \log P_x\{Z^{\{k\}}(\cdot) \in A\} \\ &\leq -\inf_{\varphi \in \overline{A}} S(T,\varphi), \end{aligned}$$

where

$$\begin{aligned} S(T,\varphi) &= \int_0^T L(\varphi(s), \dot{\varphi}(s), s)ds, \\ L(x,\beta,s) &= \sup_{\alpha \in \mathbb{R}^{m_0}} [\langle \alpha, \beta \rangle - H(x,\alpha,s)]. \end{aligned}$$

To get the escape time estimate, set

$$A = \{\varphi(\cdot) : \varphi(0) = x, \varphi(t) \notin G, \text{some } t \leq T\}.$$

Then by Theorem 4.17,

$$\begin{aligned} -\inf\nolimits_{\varphi \in A^\circ} S(T,\varphi) &\leq \liminf_{k\to\infty} \lambda_k \log P_x\{\widetilde{\tau}_G^k \leq T\} \\ &\leq \limsup_{k\to\infty} \lambda_k \log P_x\{\widetilde{\tau}_G^k \leq T\} \\ &\leq -\inf_{\varphi \in \overline{A}} S(T,\varphi). \end{aligned}$$

To bring out the dynamic system aspect of the problem, we further examine the escape probability of the iterates away from a neighborhood of the true parameter θ. To do so, we define

$$\widehat{Z}^{\{k\}}(t) = \Phi_0^{-1} C \mathbb{1} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} - \Phi_0^{-1} F^{-1}(Z^{\{k\}}(t)),$$

where

$$F^{-1}(Z^{\{k\}}(t)) = \begin{bmatrix} F^{-1}(Z_1^{\{k\}}(t)) \\ \vdots \\ F^{-1}(Z_1^{\{k\}}(t)) \end{bmatrix},$$

where $F^{-1}(\cdot)$ is the inverse of $F(\cdot)$. It is easy to see that $\widehat{Z}^{\{k\}} \to \theta$ as $k \to \infty$. Given a neighborhood $\widehat{G}$ of θ, define

$$\widehat{\tau}^k_{\widehat{G}} = \min\{t : \widehat{Z}^{\{k\}}(t) \notin \widehat{G}\},$$

then

$$P_x\{\widehat{\tau}^k_{\widehat{G}} \leq T\} \to 0.$$

Recall that

$$A = \{\varphi(\cdot) : \varphi(0) = x, \varphi(t) \notin G, \text{some } t \leq T\},$$

and define

$$F(v) = \begin{bmatrix} F(v_1) \\ \vdots \\ F(v_{m_0}) \end{bmatrix}$$

for any for $v \in \mathbb{R}^{m_0}$. By applying the contraction principle Lemma 2.5 , we have the following estimates of rate of convergence for the escape from a domain problem.

$$\begin{aligned} -\inf_{\varphi \in B^\circ} S(T, \varphi) &\leq \liminf_{k \to \infty} \lambda_k \log P_x\{\widehat{\tau}^k_{\widehat{G}} \leq T\} \\ &\leq \limsup_{k \to \infty} \lambda_k \log P_x\{\widehat{\tau}^k_{\widehat{G}} \leq T\} \\ &\leq -\inf_{\varphi \in \overline{B}} S(T, \varphi), \end{aligned}$$

where

$$B = \{F(C - \Phi_0\varphi(\cdot)), \varphi(\cdot) \in A\}.$$

In the previous sections, we have taken a direct approach for getting the large deviations bounds. This section provides an alternative for the large deviations study.

Chapter 5: Further Remarks

This dissertation derived large deviations for two-time-scale Markov chains and examined the associated LQ control problems. For future study, an interesting and important problem is to examine the large deviations of two-time-scale Markov chains with generator $Q^\epsilon(t) = \widetilde{Q}(t)/\epsilon + \widehat{Q}(t)$, where $\widetilde{Q}(t) = \mathrm{diag}(\widetilde{Q}^1(t), \ldots, \widetilde{Q}^l(t))$ such that each $\widetilde{Q}^i(t)$ is an irreducible generator and $\widehat{Q}(t)$ is another generator; see [46, Chapter 6]. This is known as a nearly decomposable model. One of the main difficulties is that we no longer have a mixing process and the associated limit is not a diffusion but a switching diffusion. It deserves further investigation.

Another important problem is concerned with moderate deviations. Consider the system of ordinary differential equations

$$\dot{x}^\epsilon(t) = b(x^\epsilon(t), \alpha^\epsilon(t)), \quad x^\epsilon(0) = x,$$

where $\alpha^\epsilon(t) \sim Q(t)/\epsilon$, $Q(t)$ is irreducible. Then $x^\epsilon(\cdot)$, converges weakly to $x(\cdot)$ such that for each $T > 0$ and for any $t \in [0, T]$,

$$\dot{x}(t) = \overline{b}(x(t)), \;\; x(0) = x_0,$$

where

$$\overline{b}(x) = \sum_{j=1}^{m} b(x, i)\nu_i,$$

and $\nu(t) = (\nu_1(t), \ldots, \nu_m(t))$ is the quasi-stationary distribution associated with $Q(t)$. Define $y^\epsilon(t) = \dfrac{1}{\epsilon^\gamma}(x^\epsilon(t) - x(t))$. When $\gamma = 0$, this is the Large deviation problem. When $\gamma = \frac{1}{2}$, this becomes the Central limit problem. What will happen if $\gamma \in [0, \frac{1}{2}]$? This is in the framework of moderate deviations. We have already established the large deviations and central limit type results. To fill the gap between large deviation and central limit, it is natural to the consider moderate deviations problem.

For system identification problem, there is also an issue concerning moderate deviations problem. Recently, Thanh, Yin and Wang [36] considered the state observers with random sampling under double-indexed and randomly weighted sums of mixing process. They considered a multi-input-single-output linear-time-invariant system

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t), \\ y(t) = Cx(t), \end{cases}$$

where A, B, C are known system matrices. Using stochastic analysis, the study of this system reduces to the convergence analysis of the following sequence,

$$\frac{1}{n^r} \sum_{i=1}^{n} \alpha_j(t_i - t_n) d_i,$$

where $\alpha_j(t_i - t_n)$ is a random process and $1/2 < r < 1$, rendering a randomly weighted triangular array of noise driven by mixing process. Thanh, Yin and Wang [36] proved the strong laws of large number and ascertained the convergence rate. On the other hand, analysis from the aspect of moderate deviation may provide a tighter error bounds.

REFERENCES

[1] A. de Acosta, Large deviations for vector-valued additive functionals of a Markov process: Lower bound, *Ann. Probab.*, **16** (1988), pp. 925-960.

[2] A. de Acosta, Large deviations for empirical measures of Markov chains, *J. Theoret. Probab.* **3** (1990), pp. 395-431.

[3] K. Aström and B. Wittenmark, *Adaptive Control*, Addison-Wesley, 1989.

[4] P. Billingsley, *Convergence of Probability Measures*, J. Wiley, New York, NY, 1968.

[5] W.P. Blair and D.D. Sworder, Feedback control of a class of linear discrete systems with jump parameters and quadratic cost criteria, *Internat. J. Control*, **21** (1986), 833–841.

[6] W. Bryc and A. Dembo. *Large Deviations and strong mixing*, *Ann. Inst. H. Poincare Probab. Stat.*, **32** (1996), 549–569.

[7] P. E. Caines, *Linear Stochastic Systems*, Wiley, New York, 1988.

[8] H.-F. Chen and L. Guo, *Identification and Stochastic Adaptive Control*, Birkhäuser, Boston, 1991.

[9] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd Edition, Springer-Verlag, New York, 1998.

[10] J.D. Deuschel and D.W. Stroock, Large deviations, Academic Press, San Diego, 1989.

[11] I.H. Dinwoodie, Identifying a large deviation rate function, *Ann. Probab*, **21** (1993), pp. 216-223

[12] M.D. Donsker and S.R.S. Varadhan, Asymptotic evaluation of certain Markov process expectations for large time, ICIV. *Comm. Pure Appl. Math.* **28** (1975), pp.1-47,279-301;**29** (1976), pp.389-461;**36** (1983), pp.183-212;

[13] P. Dupuis and H.J. Kushner, Stochastic approximations via large deviations: Asymptotic properties, *SIAM J. Control Optim.* **23** (1985), pp. 675-696.

[14] P. Dupuis and O. Zeitouni, Nonstandard form of the rate function for the occupation measure of a Markov chain, *Stochastic Process. Appl.* **61** (1996), pp. 249-261.

[15] S.N. Ethier and T.G. Kurtz, *Markov Processes: Characterization and Convergence*, J. Wiley, New York, NY, 1986.

[16] W.H. Fleming and R.W. Rishel, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, NY, 1975.

[17] M.I. Freidlin and A.D. Wentzell, On small random perturbations of dynamical system, *Russian Math. Surveys*, **25** (1970), pp. 1-55.

[18] M.I. Friedlin and A.D. Wentzel, *Random Perturbations of Dynamical Systems*, Springer-Verlag, New York, 1984.

[19] J.Gartner, On large deviations from the invariant measure, *Theory Probabil. Appl.*, **22**(1977), 24-39.

[20] F. Hollander, *Large deviations*, Amer. Math. Soc., 2008.

[21] N.C. Jain, Large deviation lower bounds for additive functionals of Markov processes. *Ann. Probab.*, (1990), pp. 1071–1098.

[22] R. Z. Khasminskii, On stochastic processes defined by differential equations with a small parameter, *Theory Probab. Appl.* **11** (1966), 211–228.

[23] R.Z. Khasminskii, G. Yin, and Q. Zhang, Asymptotic expansions of singularly perturbed systems involving rapidly fluctuating Markov chains, *SIAM J. Appl. Math.*, **56** (1996), 277–293.

[24] R.Z. Khasminskii, G. Yin, and Q. Zhang, Constructing asymptotic series for probability distribution of Markov chains with weak and strong interactions, *Quart. Appl. Math.*, **LV** (1997), 177–200.

[25] H.J. Kushner, *Approximation and Weak Convergence Methods for Random Processes, with Applications to Stochastic Systems Theory*, MIT Press, Cambridge, MA, 1984.

[26] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification and Adaptive Control*, Prentice-Hall, Englewood Cliffs, NJ, 1986.

[27] H.J. Kushner, Asymptotic Behavior of Stochstic Approximation and Large deviation, *IEEE Trans. Automat. Control*, **29** (1984), 984–990.

[28] H.J. Kushner and G. Yin, *Stochastic Approximation Algorithms and Applications*, Springer-Verlag, New York, 2nd Ed., 2003.

[29] L. Ljung and T. Söderström, *Theory and Practice of Recursive Identification*, MIT Press, Cambridge, MA, 1983.

[30] M. Mariton and P. Bertrand, Robust jump linear quadratic control: A mode stabilizing solution, *IEEE Trans. Automat. Control*, **30** (1985), 1145–1147.

[31] P. Ney and E. Nummelin, Markov additive processes (I) Eigenvalue properties and limit theorems, (II) Large deviations, *Ann. Probab.* **15** (1987), pp. 561-592.

[32] S.P. Sethi and Q. Zhang, *Hierarchical Decision Making in Stochastic Manufacturing Systems*, Birkhäuser, Boston, MA, 1994.

[33] H.A. Simon and A. Ando, Aggregation of variables in dynamic systems, *Econometrica*, **29** (1961), 111–138.

[34] A.V. Skorohod, *Studies in the Theory of Random Processes*, Dover, New York, 1982.

[35] V. Solo and X. Kong, *Adaptive Signal Processing Algorithms*, Prentice-Hall, Englewood Cliffs, NJ, 1995.

[36] L.V. Thanh, G. Yin, and L.Y. Wang, State observers with random sampling times and convergence analysis of double-indexed and randomly-weighted sums of mixing processes, *SIAM Journal on Control and Optimization*, 49 (2011), 106-124.

[37] S. R. Venkatesh and M. A. Dahleh, "Identification in the presence of classes of unmodelled dynamics and noise", *IEEE Trans. Automatic Control*, vol. 42, pp. 1620-1635, 1997.

[38] L.Y. Wang and G. Yin, Persistent Identification of Systems with Unmodeled Dynamics and Exogenous Disturbances, *IEEE Trans. Automat. Control*, **45** (2000), 1246–1256.

[39] L.Y. Wang and G. Yin, Asymptotically efficient parameter estimation using quantized output observations, *Automatica*, **43** (2007), 1178–1191.

[40] L.Y. Wang and G. Yin, Quantized identification with dependent noise and Fisher information ratio of communication channels, *IEEE Trans. Automat. Control*, **53** (2010), 674–690.

[41] L.Y. Wang, G. Yin, and J.F. Zhang, Joint identification of plant rational models and noise distribution functions using binary-valued observations, *Automatica*, **42** (2006), 535–547.

[42] L.Y. Wang, G. Yin, J.F. Zhang, and Y.L. Zhao, Space and time complexities and sensor threshold selection in quantized identification, *Automatica*, **44** (2008), 3014–3024.

[43] L.Y. Wang, G. Yin, J.-F. Zhang, and Y.L. Zhao, *System Identification with Quantized Observations: Theory and Applications*, Birkhäuser, Boston, 2010

[44] L.Y. Wang, J.-F. Zhang, and G. Yin, System Identification Using Binary Sensors, *IEEE Trans. Automat. Control*, **48** (2003), 1892–1907.

[45] R. Varga, *Matrix Iterative Analysis*, 2nd ed., Springer, 2000.

[46] G. Yin and Q. Zhang, *Continuous-time Markov Chains and Applications: A Singular Perturbations Approach*, Springer-Verlag, New York, NY, 1998.

[47] G. Yin, Q. Zhang, and G. Badowski, Asymptotic properties of a singularly perturbed Markov chain with inclusion of transient states, *Ann. Appl. Probab.*, **10** (2000), 549–572.

[48] G. Yin and H.Q. Zhang, Singularly perturbed Markov chains: Limit results and applications, *Ann. Appl. Probab.*, **17** (2007), 207–229.

[49] Q. Zhang and G. Yin, On nearly optimal controls of hybrid LQG problems, *IEEE Trans. Automat. Control*, **44** (1999), 2271–2282.

[50] G. Zames, "On the metric complexity of causal linear systems: ε-entropy and ε-dimension for continuous time", *IEEE Trans. Automatic Control*, vol. 24, pp. 222-230, 1979.

ABSTRACT

LARGE DEVIATIONS OF STOCHASTIC SYSTEMS AND APPLICATIONS

by

QI HE

AUGUST 2012

Advisor: Dr. G. George Yin

Major: Mathematics (Applied)

Degree: Doctor of Philosophy

This dissertation focuses on large deviations of stochastic systems with applications to optimal control and system identification. It encompasses analysis of two-time-scale Markov processes and system identification with regular and quantized data. First, we develops large deviations principles for systems driven by continuous-time Markov chains with two-time scales and related optimal control problems. A distinct feature of our setup is that the Markov chain under consideration is time dependent or inhomogeneous. The use of two-time-scale formulation stems from the effort of reducing computational complexity in a wide variety of applications in control, optimization, and systems theory. Starting with a rapidly fluctuating Markovian system, under irreducibility conditions, both large deviations upper and lower bounds are established first for a fixed terminal time and then for time-varying dynamic systems. Then the results are applied to certain dynamic systems and LQ control problems.

Second, we studied large deviations for identifications systems. Traditional system identification concentrates on convergence and convergence rates of estimates in mean squares, in distribution, or in a strong sense. For system diagnosis and complexity analysis, however, it is essential to understand the probabilities of identification errors over a finite data window. This paper investigates identification errors in a large deviations framework. By considering both space complexity in terms of quantization levels and time complexity with respect to data window sizes, this study provides a new perspective to understand the fundamental relationship between probabilistic errors and resources that represent data sizes in computer algorithms, sample sizes in statistical analysis, channel bandwidths in communications, etc. This relationship is derived by establishing the large deviations principle for quantized identification that links binary-valued data at one end and regular sensors at the other. Under some mild conditions, we obtain large deviations upper and lower bounds. Our results accommodate independent and identically distributed noise sequences, as well as more general classes of mixing-type noise sequences. Numerical examples are provided to illustrate the theoretical results.

AUTOBIOGRAPHICAL STATEMENT

QI HE

Education

- Ph.D. in Applied Mathematics, AUGUST, 2012 (expected)
  Wayne State University, Detroit, Michigan
- M.A. in Mathematical Statistics, May, 2011
  Wayne State University, Detroit, Michigan
- M.A. in Mathematical Statistics, May, 2007
  Sun Yat-Sen University, Guangzhou, China
- B.S. in Mathematics, June 2004
  Sun Yat-Sen University, Guangzhou, China

Awards

1. The Alfred L. Nelson Award in Recognition of Outstanding Achievement, Department of Mathematics, Wayne State University, April 2012.
2. University Dissertation Research Fellowship (UGRF), 2011-2012.
3. The Maurice J. Zelonka Endowed Mathematics Scholarship, Department of Mathematics, Wayne State University, April 2011.
4. SIAM Student Travel Award to attend the SIAM Conference on Control and Its Applications (CT11), 2011
5. Graduate Student Professional Travel Award, Department of Mathematics, Wayne State University, August 2011.

List of Publications and Preprints

1. Qi He, G. Yin, and Q. Zhang, Large Deviations for Two-Time-Scale Systems Driven by Nonhomogeneous Markov Chains and LQ Control Problems, *SIAM Journal on Control and Optimization*, **49**, (2011), 1737-1765
2. Qi He and G. Yin, Invariant Density, Liapunov Exponent, and Almost Sure Stability of Markovian-regime-switching Linear Systems, *Journal of Systems Science and Complexity*, **24**, (2011), 79-92
3. Qi He, G.Yin, and L.Wang, Large Deviations for System Identification with Applications, Submitted
4. Qi He and G.Yin, Moderate Deviations for Nonhomogeneous Two-Time-Scale Markovian Systems, Submitted