21st International Symposium on
Mathematical Theory of Networks and Systems
July 7-11, 2014. Groningen, The Netherlands

Moderate Deviations for System Identification: Error Bounds from a Probability Concentration Perspective

Qi He[1], George Yin[2], and Le Yi Wang[3]

Abstract— **This paper is devoted to moderate deviations for characterizing parameter estimation errors in system identification. Moderate deviations for system identification provide probabilistic error bounds that are beyond laws of large numbers and central limit theorems, and cannot be expressed in terms of large deviations bounds. Such error bounds are crucial in complexity analysis for system identification. Explicit error bounds are derived and their relations to identification complexity analysis are explored. Examples are included to illustrate the ideas and results of this paper.**

Key Words. Moderate deviations, probability concentration, identification error bound, complexity.

I. Introduction

This paper concentrates on probabilistic characterization of parameter estimation errors in the framework of moderate deviations. Departing from the traditional pathway of convergence and rates of convergence, we study the problem of concentration probability, namely the concentration of random estimation errors within certain bounds. Due to new technology advancement in communications, cloud computing, traffic conditions, concentration of probability, or more generally concentration of measures, has recently drawn increased attention; see [8] and also [14]. Talagrad noted in [19] that "the idea of concentration of measure (which was discovered by V. Milman) is arguably one of the great ideas of analysis in our times. While its impact on Probability is only a small part of the whole picture, this impact should not be ignored." The same can be said for the impact on system identification. This paper is an effort in applying this concept to system identification. System identification under regular sensors has been well studied with many significant results concerning identification errors, convergence, convergence rates, and applications. For related results and literature, we refer the reader to the books [1], [3], [11], [16]. For a general stochastic approximation framework of analyzing recursive identification algorithms, we refer to [13] and the references therein. System identification under quantized observations is relatively new but has achieved substantial progress in the past decade [22]–[25]. A recent monograph covers some key results in this direction [26].

This research was supported in part by the National Science Foundation under DMS-1207667.

[1]Qi He is with the Department of Mathematics, University of California, Irvine, CA 92604, qhe@math.uci.edu.

[2]George Yin is with the Department of Mathematics, Wayne State University, Detroit, MI 48202, gyin@math.wayne.edu.

[3]Le Yi Wang is with the Department of Electrical and Computer Engineering, Wayne State University, Detroit, MI 48202. lywang@wayne.edu.

Although there have been much work on the convergence and rate of convergence on system identification, investigation on detailed study of probability deviations is scarce. To understand the main issues, suppose that a sequence of vector-valued estimates from system identification $\{\widehat{\theta}_k\}$ of the true parameter θ has been generated by an identification algorithm. Under suitable conditions, the sequence is strongly consistent in the sense of convergence with probability one (w.p.1). We may further establish that $\sqrt{n}(\widehat{\theta}_k - \theta)$ converges in distribution to a normal random variable with mean zero and appropriate asymptotic covariance. The strong consistency or the asymptotic normality can affirm $P(|\widehat{\theta}_k - \theta| > \varepsilon) \to 0$ as $n \to \infty$ for any given $\varepsilon > 0$, without more precise error bounds under a finite n. In [9], estimation error bounds in terms of the large deviations principle (LDP) were developed. For n sufficiently large, we have

$$P(|\widehat{\theta}_k - \theta| > \varepsilon) \le K \exp(-c_0(\varepsilon)k),$$

where $c_0(\varepsilon)$ is known as the rate function and K is a constant depending on noise characteristics. It is noted that probabilistic upper and lower bounds offer distinctive aspects of identification errors and both are important in decision making. Large deviations upper bounds can be used to characterize guaranteed error bounds in probability and lower bounds can be used to study identification complexity. In [9], LDP bounds were applied to application case studies for battery diagnosis, medical signal processing, and modeling of electric machines.

Suppose that the estimation errors are neither in the LDP ranges, nor characterizable using laws of large numbers (LLN) or central limit theorems (CLT). To be more specific, consider the estimation error bounds of the form

$$P\Big(|\widehat{\theta}_k - \theta| > \frac{\varepsilon}{k^\alpha}\Big), \quad \text{for} \quad \alpha \in (0, \frac{1}{2}). \tag{1}$$

When $\alpha = 0$, it is in the LDP region; the corresponding error bounds have been derived in [9]. When $\alpha = 1/2$, it is in the CLT scaling. In this paper, we consider $\alpha \in (0, \frac{1}{2})$. Such estimates are known to be in the moderate deviations range. There are several reasons why moderate deviations are of importance even though LDP estimates are available.

(a) LDP estimates provide tight probability bounds on the probability concentration of estimation errors over a fix region $[\theta - \varepsilon, \theta + \varepsilon]$. Nevertheless, in many applications, we need estimation errors over variable intervals for which the LDP bounds are not valid. In particular, when the lengths of the intervals diminish to 0 in the scale of $k^{-\alpha}$ such as $[\theta - \frac{1}{k^\alpha}\varepsilon, \theta + \frac{1}{k^\alpha}\varepsilon], \alpha \in (0, 1/2)$, both

ISBN: 978-90-367-6321-9

the intervals and estimators depend on sample sizes. For example, in reliability studies of medical diagnosis [9, pp. 50-51], high accuracy is usually required. In lieu of a LDP range, probabilistic characterization of type $P(|\widehat{Q}_k - Q^*| > (\delta/k^\alpha))$ is desirable, for some $\alpha \in (0, 1/2)$, with $\widehat{Q}_k$ and Q^* given in [9, p.51, (6.18)].

(b) Because of the scaling factor $0 < \alpha < (1/2)$, behavior of estimation errors is quite different from the LDP bounds. It is necessary to display more precise asymptotic properties of the corresponding error bounds.

(c) By using moderate deviations estimates, we can further study complexity issues in system identification.

Moderate deviations principles have been extensively studied. In [7] moderate deviations bounds were derived in conjunction with averaging principles by imposing some abstract conditions. Subsequently, moderate deviations were considered in [5] for martingale difference and ϕ-mixing sequences, in [28] for Markov processes and Markov chains, in [2] and [6] for sharp results of Markov chains, in [15] for fast ergodic diffusion processes, and in [29] for large-time behavior of empirical measures of the solutions to damped Hamiltonian systems.

Moderate deviations bounds are useful for complexity analysis of system identification. In fact, identification complexity is closely associated with probability concentration. To illustrate moderate deviations concepts and their usage, some numerical examples are presented. Due to space limitation, verbatim proofs are omitted and the reader is referred to [10] for detail.

The rest of the paper is arranged as follows. The system setup is given next. Section III presents the main results. Section IV present a couple of examples together with the study of complexity of system identification. Section V provided several final remarks.

II. Formulation: System Identification

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

$$y(t) = \sum_{i=0}^{\infty} a_i u(t-i) + d(t), \quad t = t_0 + 1, \ldots, \tag{2}$$

where $\{y(t)\}$ is the noise corrupted output, $\{d(t)\}$ 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

$$\begin{aligned} \theta &= (a_0, a_1, \ldots, a_{m_0-1})' \in \mathbb{R}^{m_0}, \\ \widetilde{\theta} &= (a_{m_0}, a_{m_0+1}, \ldots)', \end{aligned} \tag{3}$$

where z' denotes the transpose of z, θ is the vector-valued modeled part of the parameters, and $\widetilde{\theta}$ is called the unmodeled dynamics. Separation of the modeled part and unmodeled dynamics is a standard modeling practice for treating model complexity [20], [21], [30], which enables us to treat parameters within a finite dimensional space; see also the related work [17], [18] and the references therein. This model complexity reduction produces modeling errors, due to the "truncation."

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

$$y(t) = \varphi'(t)\theta + \widetilde{\varphi}'(t)\widetilde{\theta} + d(t), \tag{4}$$

where

$$\begin{aligned} &\varphi(t) = (u(t), u(t-1), \ldots, u(t-m_0+1))', \quad \text{and} \\ &\widetilde{\varphi}(t) = (u(t-m_0), u(t-m_0-1), \ldots)', \end{aligned} \tag{5}$$

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

$$\begin{aligned} Y_N(t_0) &= (y(t_0), \ldots, y(t_0+N-1))' \\ D_N(t_0) &= (d(t_0), \ldots, d(t_0+N-1))' \\ \Phi_N(t_0) &= (\varphi(t_0), \ldots, \varphi(t_0+N-1))' \\ \widetilde{\Phi}_N(t_0) &= (\widetilde{\varphi}(t_0), \ldots, \widetilde{\varphi}(t_0+N-1))'. \end{aligned} \tag{6}$$

Estimates will be derived from this relationship, depending on the sensing schemes used for observing y. Moderate deviations of the estimates will be investigated accordingly.

Typical linear finite-dimensional stable systems have rational transfer functions. When representing them by their impulse responses, they are always IIR (infinite impulse response), with a decaying tail. Consequently, it is essential that the model structure (2) starts with an IIR expression.

III. Main Results

We begin by presenting the following general result first. It is concerned with a sequence of random vectors. The first part is a version of the Gärtner-Ellis Theorem (see [12, Lemma 1]). Here and thereafter, we use $\langle a, b \rangle$ to denote the usual inner product in $\mathbb{R}^{m_0}$ for $a, b \in \mathbb{R}^{m_0}$. If (7) holds, we say that $\{X_k\}$ satisfies the *large deviations principle* (LDP) with *speed* $\{\lambda_k\}$ and *rate function* $I(\beta)$.

Let $\{X_k\}$ denote a sequence of $\mathbb{R}^{m_0}$-valued random vectors, for which the following limit exists

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

where $\tau \in \mathbb{R}^{m_0}$, $\lambda_k \to 0$ and $H(\cdot)$ is continuously differentiable. Define the dual function $I(\beta) = \sup_{\tau\in\mathbb{R}^{m_0}}[\langle \tau, \beta \rangle - H(\tau)]$. Then for any open set B in $\mathbb{R}^{m_0}$,

$$\begin{aligned} -\inf_{\beta\in B} I(\beta) &\le \liminf_{k\to\infty} \lambda_k \log P\{X_k \in B\} \\ &\le \limsup_{k\to\infty} \lambda_k \log P\{X_k \in B\} \\ &\le -\inf_{\beta\in\overline{B}} I(\beta), \end{aligned} \tag{7}$$

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

Let $f : \mathbb{R}^r \to \mathbb{R}^m$ be a continuous function. Assume that a family of random variable $\{X_k\}$ on $\mathbb{R}^r$ satisfies the LDP with speed $\{\lambda_k\}$ and rate function $I : \mathbb{R}^r \mapsto [0, +\infty]$. Then the family of random variables $\{Y_k\}$ with $Y_k = f(X_k)$ satisfies the LDP with speed $\{\lambda_k\}$ and rate function $\widetilde{I}(y) = \inf\{I(x) : x \in \mathbb{R}^r, y = f(x)\}$. The above assertion confirms that the LDP is preserved by a continuous mapping, which

is called the contraction principle, see [4, Theorem 4.2.1, p.126]. In fact, the results in [4] are more general and can be applied to mappings between Hausdorff topological spaces. But this assertion is sufficient for this paper.

Using the results above, we get the following MDP theorem.

Proposition 3.1: Let $X_1, \ldots, X_k$ be a sequence of $\mathbb{R}^{m_0}$-valued i.i.d. random vectors such that

$$H(\beta) = \log E[\exp \langle X_1, \beta \rangle] < \infty$$

in some open set containing the origin, $E[X_i] = 0$. Denote the variance matrix of X_i by D and assume it is invertible. Define the empirical mean of the k samples of this sequence by

$$S_k = \frac{X_1 + \cdots + X_k}{k}.$$

Then S_k satisfies MDP with speed $k^{2\alpha-1}$ and rate function $\Lambda(x) = \frac{1}{2}\langle x, D^{-1}x \rangle$, which means the following inequalities hold,

$$\begin{aligned} -\frac{1}{2} \inf_{x \in B^0} \langle x, D^{-1}x \rangle & \\ \le \liminf_{k\to\infty} k^{2\alpha-1} \log P(k^\alpha S_k \in B) & \\ \le \limsup_{k\to\infty} k^{2\alpha-1} \log P(k^\alpha S_k \in B) & \\ \le -\frac{1}{2} \inf_{x \in \overline{B}} \langle x, D^{-1}x \rangle, & \end{aligned}$$

where B^0 and $\overline{B}$ are the interior and closure of B, respectively. Alternatively, we state that S_k satisfies MDP with parameter $\alpha \in (0, \frac{1}{2})$.

A. Regular Sensor

We assume that the following conditions hold.

(A1) (a) $\{d(t)\}$ is a sequence of i.i.d. random variables with zero-mean and variance δ. 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.

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 (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

$$L(t_0) = \frac{1}{k}\Phi_0^{-1}(I_{m_0}, \ldots, I_{m_0}). \tag{8}$$

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

$$\widehat{\theta}_k = \theta + L(t_0)\widetilde{\Phi}_N(t_0)\widetilde{\theta} + L(t_0)D_N(t_0). \tag{9}$$

It follows that the deterministic part of the identification error becomes $\eta_k^d = (I_{m_0}, I_{m_0}, \ldots)\widetilde{\theta}$. Since η_k^d is independent of k, we write it as η^d. It is easily seen that $\|\eta^d\|_1 \le \|\widetilde{\theta}\|_1$. The stochastic part of the identification error is

$$\begin{aligned} \eta_k^s &= \Phi_0^{-1}(U_k^1, \ldots, U_k^{m_0})', \quad \text{where} \\ U_k^i &= \frac{1}{k}\sum_{l=0}^{k-1} d(t_0 + lm_0 + i), \quad \text{for} \quad i = 1, \ldots, m_0. \end{aligned} \tag{10}$$

Since $\{d(t)\}$ is an i.i.d. sequence, η_k^s tends to 0 w.p.1 as $k \to \infty$. As a result, $\lim_{k\to\infty} \widehat{\theta}_k = \theta + \eta^d =: \widehat{\theta}$ w.p.1, where $\|\eta^d\|_1 \le \|\widetilde{\theta}\|_1$. In view of the above together with [9, Theorem 4.3], we obtain the following probabilistic error bounds.

Proposition 3.2: Assume (A1). Then

$$\widetilde{I}(\widetilde{\beta}) = I(G^{-1}(\widetilde{\beta})) = I(\Phi_0(\widetilde{\beta} - \theta - \eta^d))$$

is the rate function for $\{\widehat{\theta}_k\}$, where

$$\begin{aligned} I(\beta) &= \sup_{\tau \in \mathbb{R}^{m_0}} [\langle \beta, \tau \rangle - H(\tau)] \\ &= \sup_{\tau_1, \ldots, \tau_{m_0}} [\sum_{i=1}^{m_0} (\beta_i \tau_i - \log g(\tau_i))]. \end{aligned} \tag{11}$$

That is, for any open set B in $\mathbb{R}^{m_0}$, (7) holds with $I(\beta)$ replaced by $\widetilde{I}(\widetilde{\beta})$, $X_k = \widehat{\theta}_k$, and $\lambda_k = (1/k)$, respectively.

The following result gives MDP error bounds on $P(k^\alpha|\widehat{\theta}_k - \widehat{\theta}| > \varepsilon)$.

Proposition 3.3: $\widehat{\theta}_k - \theta$ satisfies MDP with rate function

$$\widetilde{\Lambda}(\widetilde{x}) = \Lambda(G^{-1}(\widetilde{x})) = \frac{1}{2\delta}\langle \Phi_0\widetilde{x}, \Phi_0\widetilde{x} \rangle = \frac{1}{2\delta}|\Phi_0\widetilde{x}|^2.$$

In particular,

$$\begin{aligned} &\lim_{k\to\infty} k^{2\alpha-1} \log P(k^\alpha|\widehat{\theta}_k - \widehat{\theta}| > \varepsilon) \\ &= -\inf_{|\widetilde{x}|>\varepsilon} \frac{1}{2\delta}|\Phi_0\widetilde{x}|^2. \end{aligned}$$

B. Binary Sensor

In the setup of using a binary sensor, the output y is measured by a binary sensor with a known threshold C. We can observe only

$$s(t) = \chi_{\{y(t)\le C\}} = \begin{cases} 1, & \text{if } y(t) \le C \\ 0, & \text{otherwise,} \end{cases}$$

where χ_A is the indicator of A. In [27], we proposed an algorithm to determine θ, obtained its convergence, and studied the corresponding asymptotic distribution of normalized errors. Using the setup in (4), we recall the algorithm.

Identification Algorithm.

Step 1. Define the empirical mean

$$Z_k = (Z_k^1, \ldots, Z_k^{m_0})'$$

by

$$Z_k^i = \frac{1}{k}\sum_{l=0}^{k-1} s(t_0 + lm_0 + i), \ i = 1, \ldots, m_0. \tag{12}$$

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

$\widetilde{c} = (\widetilde{c}_1, \ldots, \widetilde{c}_{m_0})'$. Then, Z_k^i is the value of the k-sample empirical distribution.
Step 2. Since F is invertible, we can define

$$\begin{array}{l} \gamma_k^i = F^{-1}(Z_k^i),\ i = 1, \ldots, m_0 \ \text{ and} \\ \gamma_k = (\gamma_k^1, \ldots, \gamma_k^{m_0})' = F^{-1}(Z_k), \\ L_k = C\mathbf{1} - \gamma_k, \ \text{where } \mathbf{1} = (1, \ldots, 1)'. \end{array} \tag{13}$$

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

To proceed, we propose the following assumption.

(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.
(b) $\|\widetilde{\theta}\|_1 \le \widetilde{\eta}$.
(c) We choose threshold C such that $\widetilde{c}_i$ is an interior point of the support of the density function $f(x) = F'(x)$ for each $i \le m_0$.

Lemma 3.4: Under Assumption (A2), the following hold.

a) For any compact subset $S \subset \mathbb{R}$

$$\lim_{k\to\infty} \sup_{x\in S} |\widehat{F}_k(x) - F(x)| \to 0 \ w.p.1$$

b) $\widehat{B}_k(\cdot)$ converges weakly to $B(\cdot)$, a stretched Brownian bridge process such that the covariance of $B(\cdot)$ is given by

$$EB(x_1)B(x_2) = \min(F(x_1), F(x_2)) - F(x_1)F(x_2).$$

To proceed, define functions $\mathbf{F}$ and $\mathbf{F}^{-1}$ on $\mathbb{R}^{m_0}$ by

$$\begin{array}{l} \mathbf{F}(v) = (F(v_1), \ldots, F(v_{m_0}))', \\ \mathbf{F}^{-1}(v) = (F^{-1}(v_1), \ldots, F^{-1}(v_{m_0}))' \text{ for } v \in \mathbb{R}^{m_0}. \end{array}$$

We also define the function $\widehat{G} : \mathbb{R}^{m_0} \to \mathbb{R}^{m_0}$ by $\widehat{G}(x) = \Phi_0^{-1}[C\mathbf{1} - \mathbf{F}^{-1}(x)]$ and write $\widehat{\theta}_k = \widehat{G}(Z_k)$. We first recall some results of estimates of $\{\widehat{\theta}_k\}$ in [9].

Proposition 3.5: Define

$$\begin{aligned} I(\beta) &= \sup_{\tau_1,\ldots,\tau_{m_0}} [\sum_{i=1}^{m_0} (\beta_i\tau_i - \log(e^{\tau_i} b_i + 1 - b_i))], \\ &= \begin{cases} \displaystyle\sum_{i=1}^{m_0} \log \frac{\beta_i^{\beta_i}(1-b_i)^{\beta_i - 1}}{b_i^{\beta_i}(1-\beta_i)^{\beta_i-1}}, \\ \infty, \ \text{otherwise}, \end{cases} \end{aligned} \tag{14}$$

where $0 \le \beta_i \le 1$, for $i = 1, \ldots, m_0$. Then $I(\beta)$ is the rate function for the sequence $\{Z_k\}$ defined in (12). For any given open set B in $\mathbb{R}^{m_0}$, (7) holds with $\lambda_k = (1/k)$ and $X_k = Z_k$, respectively.

Note that in the calculation of upper bounds, for any $\varepsilon > 0$,

$$\begin{aligned} &P(k^\alpha|\widehat{\theta}_k - \theta| \ge \varepsilon) \\ &\le P(k^\alpha|\gamma_k - \widetilde{c}| \ge \frac{\varepsilon}{|\Phi_0^{-1}|}) \\ &= P(k^\alpha|\gamma_k - \widetilde{c}| \ge \frac{\varepsilon}{|\Phi_0^{-1}|} \cap |\gamma_k - \widetilde{c}| \le \delta) \\ &+P(k^\alpha|\gamma_k - \widetilde{c}| \ge \frac{\varepsilon}{|\Phi_0^{-1}|} \cap |\gamma_k - \widetilde{c}| > \delta) \\ &\le P\left(k^\alpha|\mathbf{F}(\gamma_k) - \mathbf{F}(\widetilde{c})| \ge \frac{N_1\varepsilon}{|\Phi_0^{-1}|}\right) + P(|\gamma_k - \widetilde{c}| > \delta) \\ &\le P\left(k^\alpha|Z_k - \mathbf{F}(\widetilde{c})| \ge \frac{N_1\varepsilon}{|\Phi_0^{-1}|}\right) \\ &\quad +K\exp\left(-k \inf_{|\widetilde{\beta}-\widetilde{c}|>\delta} \widetilde{I}(\beta)\right) \\ &\le K\exp(-k^{1-2\alpha} \inf_{|t|\ge\frac{N_1\varepsilon}{|\Phi_0^{-1}|}} \frac{1}{2}\langle t, D_b t\rangle), \end{aligned} \tag{15}$$

which means

$$\begin{aligned} &\limsup_{k\to\infty} k^{2\alpha-1} \log P(k^\alpha|\widehat{\theta}_k - \theta| \ge \varepsilon) \\ &\le \limsup_{k\to\infty} k^{2\alpha-1} \log(P\left(k^\alpha|Z_k - \mathbf{F}(\widetilde{c})| \ge \frac{N_1\varepsilon}{|\Phi_0^{-1}|}\right) \\ &\qquad + \exp\left(-kN_3\right)) \\ &= -\inf_{|t|\ge\frac{N_1\varepsilon}{|\Phi_0^{-1}|}} \frac{1}{2}\langle t, D_b t\rangle. \end{aligned}$$

Moreover, consider

$$\widehat{M}_k(x) = k^\alpha(\widehat{F}_k(x) - F(x)) \ \text{ for } \ \alpha \in (0, \frac{1}{2}). \tag{16}$$

To get the rate function of $\widehat{M}_k(x)$, applying Proposition 3.1, we need to calculate the variance of $(\chi_{\{d(t)\le x\}} - F(x))$,

$$\begin{aligned} &\text{Var}((\chi_{\{d(t)\le x\}} - F(x))) \\ &\quad = E((\chi_{\{d(t)\le x\}} - F(x))^2) \\ &\quad = E(\chi_{\{d(t)\le x\}} + F^2(x) - 2F(x)\chi_{\{d(t)\le x\}}) \\ &\quad = F(x) + F^2(x) - 2F^2(x) \\ &\quad = F(x) - F^2(x). \end{aligned}$$

Then the rate function of $\widehat{M}(x)$ defined in (16) is $\Lambda(t) = \frac{1}{2(F(x)-F^2(x))}t^2$. The rate function of $k^\alpha(Z_k - \mathbf{F}(\widetilde{c}))$ is

$$\widehat{\Lambda}(t) = \frac{1}{2}\langle t, D_b t\rangle,$$

where

$$D_b = \text{diag}(\frac{1}{(F(\widetilde{c}_1) - F(\widetilde{c}_1)^2)}, \ldots, \frac{1}{(F(\widetilde{c}_{m_0}) - F(\widetilde{c}_{m_0})^2)}).$$

C. Quantized Sensor

We study system identification under quantized sensors in this section. For simplicity, we consider the case $m_0 = 1$. Consider the gain system given by $y(l) = u(l)\theta + d(l),\ l = 1, 2, \ldots$, 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 $y(l) = \theta + d(l)$. Let $p_i = P(C_{i-1} < y(l) \le C_i) = F(C_i - \theta) - F(C_{i-1} - \theta) := F_i(\theta)$, and $p = (p_1, \ldots, p_m)'$. Consider k measurements on $s(l)$. Then $\xi_k^i = \frac{1}{k}\sum_{l=1}^{k} s^i(l)$, for $i = 1, \ldots, m$ is the sample relative frequency of $y(l)$ taking values in $(C_{i-1}, C_i]$. Throughout this section we impose the following assumption.

(A3) (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. (b) We choose thresholds $-\infty < C_1 < \cdots < C_m < \infty$ such that there exists a compact subset E and constants $0 < N_1 < N_2 < 0$ such that $\theta \in E$ and $N_1|x_1 - x_2| < |F_i(x_1) - F_i(x_2)| < N_2|x_1 - x_2|$ for any $x_1, x_2 \in E, i \le m$.

Under Assumption (A3), $\{y(l)\}$ is an i.i.d. sequence that has the accumulative distribution function $F(x - \theta)$. 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 [22] that $\Theta_k = G(\xi_k)$ is an asymptotically unbiased estimator of $\theta\mathbf{1}$. Define $C = (C_1, \ldots, C_m)'$ and $\widehat{F}(v) = (F_1(v_1), \ldots, F_m(v_m))'$ for $v \in \mathbb{R}^m$. Then we can obtain the following results.

Θ_k satisfies LDP with the rate function

$$\widetilde{I}(\widehat{\beta}) = \inf\{I(\beta) : G(\beta) = \widehat{\beta}\} = I(\widehat{F}(\widehat{\beta})), \tag{17}$$

where

$$I(\beta) = \begin{cases} \sum_{i=1}^{m+1} \beta_i \log \frac{\beta_i}{p_i}, \ \beta \in D \\ \infty, \ \text{otherwise.} \end{cases} \tag{18}$$

where $D = \{\beta : 0 \le \beta_i \le 1, \sum_{i=1}^m \beta_i \le 1\}$, $p_{m+1} = 1 - \mathbf{1}'p$ and $\beta_{m+1} = 1 - \mathbf{1}'\beta$.

Furthermore, we establish the moderate deviations error bounds. Here in the calculation of upper bounds, for any $\varepsilon > 0$, and large enough k,

$$\begin{aligned} &P(k^\alpha|\Theta_k - \theta\mathbf{1}| \ge \varepsilon) \\ &= P(k^\alpha|\Theta_k - \theta\mathbf{1}| \ge \varepsilon \cap |\Theta_k - \theta\mathbf{1}| < \delta) \\ &\quad + P(k^\alpha|\Theta_k - \theta\mathbf{1}| \ge \varepsilon \cap |\Theta_k - \theta\mathbf{1}| \ge \delta) \\ &\le P(k^\alpha|G(\Theta_k) - G(\theta\mathbf{1})| \ge N_1\varepsilon) + P(|\Theta_k - \theta\mathbf{1}| \ge \delta) \\ &= P(k^\alpha|\xi_k - p| \ge N_1\varepsilon) + P(|\Theta_k - \theta\mathbf{1}| \ge \delta) \\ &\le 2K\exp(-k^{1-2\alpha} \inf_{|t| \ge N_1\varepsilon} \frac{1}{2}\langle t, D_q t\rangle). \end{aligned} \tag{19}$$

We can also proceed to getting the lower bounds.

We thus have

$$\begin{aligned} &\liminf_{k\to\infty} k^{2\alpha-1} \log P(k^\alpha|\Theta_k - \theta\mathbf{1}| \ge \varepsilon) \\ &\ge \liminf_{k\to\infty} k^{2\alpha-1} \log P(k^\alpha|\xi_k - p|) \\ &= -\inf_{|t| \ge N_2\varepsilon} \frac{1}{2}\langle t, D_q t\rangle. \end{aligned}$$

IV. Numerical Examples

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

We consider a system identification example, where $\theta' = (1.75, 1.75, 2.75)$ is the true parameter vector and for simplicity assume $\widetilde{\theta} = 0$, i.e., no unmodeled dynamics. Select a 3-periodic input with one-period values $(u(1), u(2), u(3)) = (3, 4, 5)$, which is full rank, and

$$\Phi_0 = \begin{pmatrix} 3 & 4 & 5 \\ 4 & 5 & 3 \\ 5 & 3 & 4 \end{pmatrix}.$$

Regular Sensors. In this simulation, the noise is an i.i.d. sequence of random variables with the standard normal distribution. Let the parameter $\alpha = \frac{1}{4}$ and choose $\varepsilon = 0.5$. In this case we aim to find the convergence rate of

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

where $U_k^j = \frac{1}{k}\sum_{l=0}^{k-1} d(t_0 + lm_0 + j)$, for $j = 1, 2, 3$ and Φ_0 is as above. By virtue of Theorem 3.3, the rate function is $\Lambda(x) = \frac{1}{2}|\Phi_0 x|^2$. We simulate the probability $P(k^\alpha|\widehat{\theta}_k - \theta| > 0.5)$, and then compare with the moderate deviations result, in which this probability is approximated by $K\exp(-k^{1-2\alpha} \inf_{|x|>0.5} \Lambda(x))$ 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 $k^\alpha|\widehat{\theta}_k - \theta| > 0.5$ to approximate $P(k^\alpha|\widehat{\theta}_k - \theta| > 0.5)$.

Step 2. By calculating the value $\inf_{|x|>0.5} \Lambda(x) = \inf_{|x|>0.5} \frac{|\Phi_0 x|^2}{2} = 0.375$, the exponential decay curve is $\exp(-0.375k^{1-2\alpha})$.

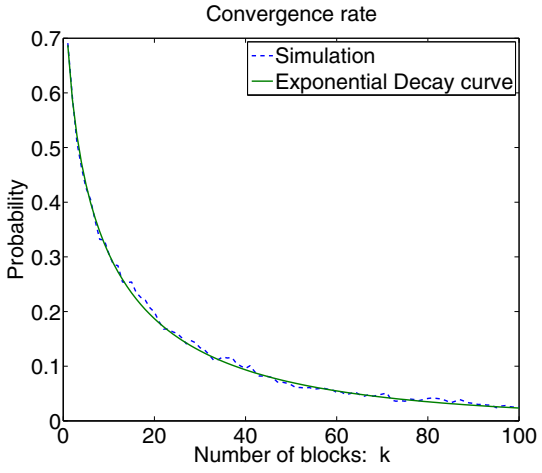

Fig. 1. Comparison of the empirical errors and the MDP bound under a regular sensor.

Binary Sensors. In this case, we consider uniform noise. Assume that $\{d(l)\}$ is a sequence of i.i.d. random variables uniformly distributed on $[-1.5, 1.5]$. Then the cumulative distribution function is $F(x) = \frac{1}{3}x + \frac{1}{2}$, for $x \in [-1.5, 1.5]$. Hence, we have that $N_1 = N_2 = \frac{1}{3}$ in this case. Consider the binary sensor with threshold $C = 25$. For $\varepsilon = 1$ and $\alpha = \frac{1}{4}$, we compare empirical measures of $P(k^\alpha|\widehat{\theta}_k - \theta| > \varepsilon)$, with the calculated moderate deviations upper bound by the result of (15).

Step 1. For each $k = 5, \ldots, 100$, the identification algorithm is repeated 1000 times and the sample frequencies of the event $k^\alpha|\widehat{\theta}_k - \theta| > \varepsilon$ are then calculated as an approximation to $P(k^\alpha|\widehat{\theta}_k - \theta| > \varepsilon)$.

Step 2. From (15), compute

$$\exp -k^{1-2\alpha} \inf_{|t| \geq \frac{N_1 \varepsilon}{|\Phi_0^{-1}|}} \frac{1}{2}\langle t, D_b t\rangle,$$

where $N_1 = 1/3$, and

$$\begin{aligned} &D_b \\ &= \mathrm{diag}\left(\frac{1}{F(-1) - F(-1)^2}, \frac{1}{F(1) - F(1)^2}, \frac{1}{F(0) - F(0)^2}\right) \\ &= \mathrm{diag}(36/5, 36/5, 4). \end{aligned}$$

It is calculated

$$- \inf_{|t| \geq \frac{N_1 \varepsilon}{|\Phi_0^{-1}|}} \frac{1}{2}\langle t, D_b t\rangle = 1.19,$$

and the MDP exponential decaying curve is $\exp(-1.19k^{1/2})$. These are shown in Figure 2.

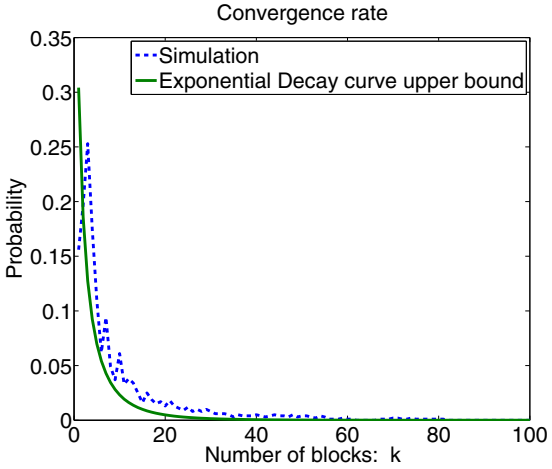

Fig. 2. Comparison of the empirical errors and the MDP bound under a binary sensor.

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

The MDP indicates that the bound of $P(k^\alpha|\widehat{\theta}_k - \theta| \geq \varepsilon) \leq K\exp(-k^{1-2\alpha}\inf_{|\beta| \geq \varepsilon}\Lambda(\beta))$, for some $K > 0$ where Λ 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 a sequence of i.i.d. random variables uniformly distributed on $[-1.5, 1.5]$. We assume that $\theta = 3$, $\alpha = 1/4$, $C_1 = 2$, and $C_2 = 4$. Here, $p_1 = P(d < -1) = 1/6$ and $p_2 = P(-1 < d < 1) = 2/3$. We want to find the probability $P(k^\alpha|\widehat{\theta}_k - \theta| > 1)$.

a. 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 (3.3), the upper bound is $P(k^\alpha|\widehat{\theta}_k^r - \theta| > 1) \leq K\exp(-k^{1-2\alpha}\inf_{|x| \geq \varepsilon}\frac{1}{2\delta}|x|^2)$. Hence the bound of the two dimensional estimator $\Theta_k^r = (\widehat{\theta}_k^r, \widehat{\theta}_k^r)'$ is

$$\begin{aligned} &P(k^\alpha|\Theta_k^r - \Theta| > 1) \\ &= P(k^\alpha|\theta_k^r - \theta| > 1/\sqrt{2}) \\ &\approx K\exp(-k^{1-2\alpha}\inf_{|x| > 1/\sqrt{2}}\frac{|x|^2}{2\delta}) \\ &= K\exp(-0.33k^{1-2\alpha}). \end{aligned}$$

b. 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 $\xi_k^1 = \frac{\sum_{l=1}^k \chi_{\{d(l) \leq C_1 - \theta\}}}{k}$. By (15), $P(k^\alpha|\widehat{\theta}_k^b - \theta|)$ has the upper bound

$$K\exp(-k^{1-2\alpha}\inf_{|t| > N_1/|\Phi^{-1}}\frac{1}{2}\langle t, Dt\rangle),$$

where $N_1 = 1/3$, $D = 1/p_1$. Hence the estimator $\Theta_k^b = (\widehat{\theta}_k^b, \widehat{\theta}_k^b)'$ converge to $\theta\mathbf{1}$. The error bound on $P(k^\alpha|\Theta_k^b - \Theta|)$ is

$$\begin{aligned} &P(k^\alpha|\Theta_k^b - \Theta| > 1) \\ &= P(k^\alpha|\theta_k^b - \theta| > 1/\sqrt{2}) \\ &\approx K\exp(-k^{1-2\alpha}\inf_{|t| > \frac{1}{3}\frac{1}{\sqrt{2}}}\frac{1}{2(p_1 - p_1^2)}t^2) \\ &= K\exp(-0.2k^{1-2\alpha}). \end{aligned}$$

c. 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 $\xi_k^i = \sum_{l=1}^k \chi_{\{d(l) \leq C_i - \theta\}}/k$ for $i = 1, 2$. By (19), $P(k^\alpha|\Theta_k^q - \Theta| > 1)$ has upper bound

$$K\exp(-k^{1-2\alpha}\inf_{|t| \geq 1/3}\tfrac{1}{2}\langle t, Dt\rangle)$$

where $D = \mathrm{diag}(\frac{1}{p_1 - p_1^2}, \frac{1}{p_2 - p_2^2})$. By minimization, we have that $\inf_{|t| \geq 1/3}\frac{1}{2}\langle t, Dt\rangle = -0.25$. Thus the upper bound is

$$K\exp(-0.25k^{1-2\alpha}).$$

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

$$P(k^\alpha|\theta_k - \theta| > 1) \approx \begin{cases} K\exp(-0.20k^{1-2\alpha}) & \text{binary sensor,} \\ K\exp(-0.25k^{1-2\alpha}) & \text{quantized sensor,} \\ K\exp(-0.33k^{1-2\alpha}) & \text{regular sensor,} \end{cases} \quad (20)$$

where the quantized sensor has two threshold values. Figure 3 displays the comparison results of the exponential curves.

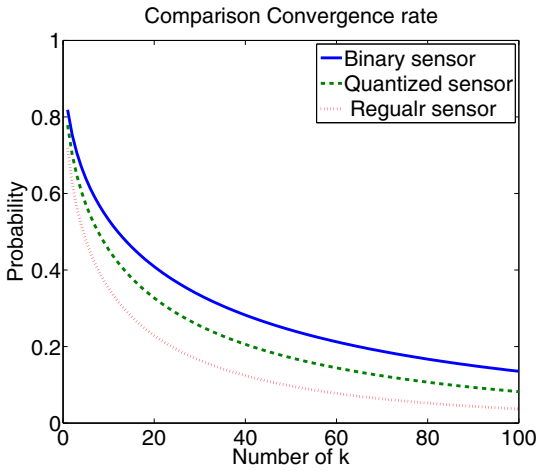

Fig. 3. Comparison of convergence rates among different sensors.

In view of the study in [25], the two-threshold quantized sensor design is a refinement of the binary sensor case, and the regular sensor is an infinite refinement of quantized sensors. The moderate deviations error bounds give precise descriptions on convergence rates, hence can be used in selecting sensor complexity levels. For further discussion on complexity issues, see [25].

V. Further Remarks

This paper has been devoted to moderate deviations in system identification. The effort can be considered as a way of studying concentration probability. One of the uses is the analysis for complexity in system identification. As for specific applications, as in [9], we can carry out case studies for battery diagnosis, medical signal processing, and electric machine.

References

[1] K. Aström and B. Wittenmark, *Adaptive Control*, Addison-Wesley, 1989.
[2] A. de Acosta, Moderate deviations for empirical measure of Markov chains: Lower bound. *Ann. Probab.* **25** (1998), 259–284.
[3] H.-F. Chen and L. Guo, *Identification and Stochastic Adaptive Control*, Birkhäuser, Boston, 1991.
[4] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd Edition, Springer-Verlag, New York, 1998.
[5] H. Djellout, Moderate deviations for martingale differences and applications to ϕ-mixing sequences, *Stoch. Stoch. Rep.* **73** (2002), 37–63.
[6] H. Djellout and A. Guillin, Moderate deviations for Markov chains with atom. *Stochastic Process. Appl.* **95** (2001), no. 2, 203–217.
[7] M.I. Friedlin and A.D. Wentzel, *Random Perturbations of Dynamical Systems*, Springer-Verlag, New York, 1984.
[8] A.A. Giannopoulos and V. Milman, Concentration property on probability spaces, *Advances in Mathematics* **156** (2000), 77–106.
[9] Q. He, L. Wang, and G. Yin, System Identication Using Regular and Quantized Observations: Applications of Large Deviations Principles, *Springer Briefs in Mathematics*. 2013.
[10] Q. He, G. Yin, and L.Y. Wang, System identification under regular, binary, and quantized observations: Moderate deviations error bounds, under revision for *IEEE Trans. Automat. Control.*
[11] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification and Adaptive Control*, Prentice-Hall, Englewood Cliffs, NJ, 1986.
[12] H.J. Kushner, Asymptotic behavior of stochastic approximation and large deviation, *IEEE Trans. Automat. Control*, **29** (1984), 984–990.
[13] H.J. Kushner and G. Yin, *Stochastic Approximation Algorithms and Applications*, Springer-Verlag, New York, 2nd Ed., 2003.
[14] M. Ledoux, *The Concentration of Measure Phenomenon*, 2001, Amer. Math. Soc.
[15] R.S. Liptser and V. Spokoiny, Modrate deviations type evaluation for integral functional of diffusion processes. *Eletron. J. Probab*, **4** (1999), 1–25.
[16] L. Ljung and T. Söderström, *Theory and Practice of Recursive Identification*, MIT Press, Cambridge, MA, 1983.
[17] M. Milanese and G. Belforte, Estimation theory and uncertainty intervals evaluation in the presence of unknown but bounded errors: Linear families of models and estimators, *IEEE Trans. Automat. Control*, **27** (1982), 408–414.
[18] M. Milanese and A. Vicino, Optimal estimation theory for dynamic systems with set membership uncertainty: An overview, *Automatica*, **27** (1991), 997–1009.
[19] M. Talagrand, A new look at independence, *Ann. Probab.* **24** (1996), 1-34.
[20] 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.
[21] L.Y. Wang and G. Yin, Persistent Identification of Systems with Unmodeled Dynamics and Exogenous Disturbances, *IEEE Trans. Automat. Control*, **45** (2000), 1246–1256.
[22] L.Y. Wang and G. Yin, Asymptotically efficient parameter estimation using quantized output observations, *Automatica*, **43** (2007), 1178–1191.
[23] 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.
[24] 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.
[25] 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.
[26] L.Y. Wang, G. Yin, J.-F. Zhang, and Y.L. Zhao, *System Identification with Quantized Observations: Theory and Applications*, Birkhäuser, Boston, 2010
[27] L.Y. Wang, J.-F. Zhang, and G. Yin, System Identification Using Binary Sensors, *IEEE Trans. Automat. Control*, **48** (2003), 1892–1907.
[28] L. Wu, Moderate deviations of dependent random variables related to CLT. *Ann. Probab.* **23**, (1995) 420–445.
[29] L. Wu, Large and moderate deviations and exponential convergence for stochastic damping hamiltonian systems. *Stochas. Process. Appl.* **91**, (2001) 205–238.
[30] 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.