Preprints of the 19th World Congress
The International Federation of Automatic Control
Cape Town, South Africa. August 24-29, 2014

Iterative Parameter Estimate with Batched Binary-Valued Observations: Convergence with an Exponential Rate ⋆

Wenjian Bi * **Yanlong Zhao** *

* *Academy of Mathematics and Systems Science, CAS, Beijing, 100190 China, (e-mail: biwenjian10@mails.ucas.ac.cn, ylzhao@amss.ac.cn).*

Abstract: This paper considers the linear system identification with batched binary-valued observations. An iterative parameter estimate algorithm is constructed to achieve the Maximum Likelihood (ML) estimate. The first interesting result is that there exists at most one finite ML solution for this specific maximum likelihood problem, which is induced by the fact that the Hessian matrix of the log-likelihood function is negative definite under binary data and Gaussian system noises. The global concave property and local strongly concave property of the log-likelihood function are obtained. Under mild conditions on the system input, the ML function can be proved to have unique maximum point. The second main result is that the proposed iterative estimate algorithm converges to a fixed vector with an exponential rate which are proved by constructing a Lyapunov function. The more interesting result is that the limit of the iterative algorithm achieves the maximization of the ML function. Numerical simulations are illustrated to support the theoretical results obtained in this paper well.

Keywords: Binary-valued observation, maximum likelihood estimate, strongly convex, system identification, exponential rate.

1. INTRODUCTION

System identification with set-valued measurements has been showing their wide applications in different fields, such as networked control systems, biological networks, and communication systems, see, e.g., Wang et al. (2010). The set-valued sensor introduces substantial difficulties since only very limited information is available to the system identification. Fortunately, the latest research results that it can be converted to a corresponding Maximum Likelihood (ML) problem, which shed light on the study of the set-valued system identification, see, e.g. Godoya et al. (2011), Chen et al. (2012). However, the solution of the ML problem usually cannot be obtained explicitly even limited to the linear system with binary observations. Instead, people usually search for iterative estimates that can converge to the ML estimate.

In the ML field, the Expectation Maximization (EM) algorithm proposed by Dempster et al. (1977) has obtained great success. It produces a sequence of the iterative estimates $\{\hat{\theta}_t, t = 1, 2, ...\}$ of the parameter θ. As the iteration step goes on, it is guaranteed that the log-likelihood function $\{l(\hat{\theta}_t)\}$ is non-decreasing. Coupled with the upper bound of the log-likelihood function, there exists an l^* which is the limit of $\{l(\hat{\theta}_t)\}$. But the convergence of the parameter estimates $\{\hat{\theta}_t\}$ cannot be concluded by the convergence of $\{l(\hat{\theta}_t)\}$, see, e.g. Wu (1983). Hence, its theoretical feasibility based on the specific log-likelihood function with binary data and Gaussian system noises is worthy of being derived.

For binary-valued systems, the EM algorithm is introduced by Godoya et al. (2011) to estimate the model parameter and simulation results show the convergence property of the iterative procedures. However, there are still some fundamental questions to be answered such as how to construct a convergent iterative estimate algorithm? What is the convergence rate? What kind of properties does the limit of the iteration have?

This paper constructs an iterative algorithm to estimate the system parameter based on batched binary data to achieve the ML estimate. Under mild conditions on the system input, the ML function is proved to have unique maximum point, the necessary and sufficient condition for which is given. The algorithm is proved to be convergent with an exponential rate to a fixed vector, which is exactly the ML estimate under batched binary-valued observations.

The rest of the paper is organized as following: Section 2 introduces the identification problem and its corresponding ML criterion; and an iteration estimate algorithm is constructed. Section 3 analyzes the likelihood function and obtains a sufficient and necessary condition for the existence and uniqueness of the maximum point of the likelihood function. Section 4 derives the convergence of the algorithm and obtains an exponential convergence rate. Some results are illustrated through extensive numerical simulations in Section 5. Section 6 concludes the paper and discusses related future works.

⋆ This work was supported by National Natural Science Foundation of China under grants 61174042 and 11171333, and Youth Innovation Promotion Association Foundation of CAS under grant 4106960. Corresponding author Yanlong Zhao. Tel. +86-10-62651446.

2. PROBLEM FORMULATION

In this section, the system with batched binary-valued observations is introduced and the corresponding parameter identification problem is formulated. The ML estimate problem is given and an iterative algorithm is constructed.

2.1 Binary-valued system

We consider the scheme where noise enters between the n-dimensional linear system and a binary sensor. The system can be described by

$$\begin{cases} y_k = \phi_k^T\theta + e_k, \\ s_k = I_{[y_k \le C]}, \quad 1 \le k \le N, \end{cases} \tag{1}$$

where $\phi_k \in \mathbb{R}^n$ is the system input, $y_k \in \mathbb{R}$ is the system output and $\theta \in \mathbb{R}^n$ is a constant but unknown parameter vector; $s_k \in \{0,1\}$ is the binary-valued observation generated by the comparison between the system output and a given sensor threshold $C \in \mathbb{R}$, I is the indicator function. The data length is N. For all $k \le N$, system noise $\mathscr{E}_N = \{e_1, e_2, \dots, e_N\}$ is assumed to be independent with a zero-mean and variance 1 Gaussian distribution.

Assumption 1. Matrix $A = \sum_{k=1}^N \phi_k\phi_k^T$ is positive definite.

Remark 1. Assumption 1 is the mathematical description of persistent excitation condition, which is a common assumption in the research of system identification, see, e.g. Ljung (1999).

The problem of interest is to estimate the parameter θ using the binary-valued observations $\mathscr{O}_N = \{s_1, s_2, \dots, s_N\}$ and input data $\mathscr{I}_N = \{\phi_1, \phi_2, \dots, \phi_N\}$.

2.2 Maximum likelihood criterion

Consider system (1) with the input data $\mathscr{I}_N$ and output data $\mathscr{O}_N$, the log-likelihood function $l(\theta)$ is given by

$$\begin{aligned} l(\theta) = &\textstyle\sum_{\{k\le N, s_k=1\}} \log[F(C-\phi_k^T\theta)] \\ &+ \textstyle\sum_{\{k\le N, s_k=0\}} \log[1-F(C-\phi_k^T\theta)], \end{aligned} \tag{2}$$

where $F(x)$ and $f(x)$ are the Cumulative Distribution Function(CDF) and Probability Density Function(PDF) of the standard normal distribution.

The corresponding ML estimate is the parameter that maximize the log-likelihood function:

$$\hat\theta = \arg\max_\theta l(\theta). \tag{3}$$

Remark 2. Since log-likelihood function and related ML estimate are relevant to N observations, $l_N(\theta)$ and $\hat\theta_N$ are more accurate representation. In this paper, for the convenience of description, symbols $l(\theta)$ and $\hat\theta$ are employed.

2.3 Iterative estimate algorithm

At first, we introduce the basic idea and some common properties of the EM algorithm. For the details of the EM algorithm, see, e.g. Dempster et al. (1977). Given $\hat\theta_t$ which is the estimate at iteration t, the core idea of the EM algorithm is to construct a function $l(\theta|\hat\theta_t)$ that satisfies the following 2 properties:

(i) $l(\theta|\hat\theta_t) \le l(\theta)$ holds for all θ,

(ii) $l(\hat\theta_t|\hat\theta_t) = l(\hat\theta_t)$,

and then calculate the argument $\hat\theta_{t+1}$ of the maximum value of function $l(\theta|\hat\theta_t)$. Hence,

$$l(\hat\theta_{t+1}) \ge \max l(\theta|\hat\theta_t) \ge l(\hat\theta_t|\hat\theta_t) = l(\hat\theta_t).$$

This guarantees the non-decrease of log-likelihood function. The construction process of function $l(\theta|\hat\theta_t)$ is the E-step and maximization process is the M-step.

Back to the binary-valued model, the E-step provides the following $l(\theta|\hat\theta_t)$ with quadratic form:

$$\begin{aligned} l(\theta|\hat\theta_t) = &-\tfrac{1}{2}\theta^T\Big(\textstyle\sum_{k=1}^N \phi_k\phi_k^T\Big)\theta + \Big[\Big(\textstyle\sum_{k=1}^N \phi_k\phi_k^T\Big)\hat\theta_t \\ &-\Big(\textstyle\sum_{k=1}^N \phi_k \cdot f(C-\phi_k^T\hat\theta_t)\big[\frac{I_{[s_k=1]}}{F(C-\phi_k^T\hat\theta_t)} - \frac{I_{[s_k=0]}}{F(\phi_k^T\hat\theta_t - C)}\big]\Big)\Big]^T\theta \\ &+l_1(\hat\theta_t), \end{aligned}$$

where $l_1(\hat\theta_t)$ is the part which is independent from θ.

Under Assumption 1, the iterative algorithm is obtained as following:

$$\begin{aligned} \hat\theta_{t+1} &= \arg\max_\theta l(\theta|\hat\theta_t) \\ &= \hat\theta_t - \Big(\textstyle\sum_{k=1}^N \phi_k\phi_k^T\Big)^{-1}\Big(\textstyle\sum_{k=1}^N \phi_k f(C-\phi_k^T\hat\theta_t) \\ &\quad \cdot\big[\frac{I_{[s_k=1]}}{F(C-\phi_k^T\hat\theta_t)} - \frac{I_{[s_k=0]}}{1-F(C-\phi_k^T\hat\theta_t)}\big]\Big). \end{aligned} \tag{4}$$

The above algorithm will be showed to converge to the ML estimate (3) with an exponential rate in the rest of this paper. As the initial issue, the uniqueness of the solution of (3) is need to be assured, which makes the estimate is exactly the system parameter.

3. EXISTENCE AND UNIQUENESS OF THE ML ESTIMATE

In this section, we explore the properties of ML estimate by analyzing the log-likelihood function (2) and prove the existence and uniqueness of ML estimate. The main reason is that the special likelihood function with binary-valued observations is concave.

Lemma 1. Function $p(x) = \frac{f(x)xF(x)+f^2(x)}{F^2(x)}$ is a strictly decreasing function and $0 < p(x) < 1$ for $x \in (-\infty, \infty)$.

Proof. The result can be obtained by repeated use of L'Hôpital's rule, see, e.g., Apostol (1974). □

The non-negative property of $p(x)$ induces the concaveness of the log-likelihood function, which can be described in the following lemma.

Lemma 2. Under Assumption 1, log-likelihood function $l(\theta)$ given in (2) is a concave function on $\mathbb{R}^n$. Given any $r > 0$, $l(\theta)$ is a strongly concave function on set $\mathscr{S} = \{\theta, \|\theta\| \le r\}$.

Proof. Based on Assumption 1, $A > 0$. Hence, there exists a minimal eigenvalue $\lambda_1 = \lambda_{\min}(A)$ such that $A \ge \lambda_1 I$.

Calculate the gradient vector and Hessian matrix of log-likelihood function:

$$\nabla l(\theta) = \Big[\sum_{\{s_k=1\}} \frac{-f(C-\phi_k^T\theta)}{F(C-\phi_k^T\theta)} + \sum_{\{s_k=0\}} \frac{f(\phi_k^T\theta - C)}{F(\phi_k^T\theta - C)} \Big] \phi_k,$$

$$\nabla^2 l(\theta) = -\Big[\sum_{\{s_k=1\}} \frac{f(x)xF(x)+f^2(x)}{F^2(x)}\Big|_{x=C-\phi_k^T\theta} + \sum_{\{s_k=0\}} \frac{f(x)xF(x)+f^2(x)}{F^2(x)}\Big|_{x=\phi_k^T\theta-C} \Big] \phi_k\phi_k^T$$

for $k \le N$.

Define $p(x) = \frac{f(x)xF(x)+f^2(x)}{F^2(x)}$ and rewrite the Hessian matrix as following:

$$\nabla^2 l(\theta) = -\sum_{k=1}^{N} \big[p(C-\phi_k^T\theta) I_{[s_k=1]} + p(\phi_k^T\theta - C) I_{[s_k=0]} \big] \phi_k\phi_k^T.$$

Lemma 1 tells us the monotonicity and boundedness of function $p(x)$. Hence, $\nabla^2 l(\theta) \le 0$ can be directly concluded through $p(x) > 0$, which infers the concave property of $l(\theta)$. If fixed on the set $\mathscr{S} = \{\theta, \|\theta\| \le r\}$, the boundedness of θ and finiteness of ϕ_k guarantee that there exists $\mu > 0$ such that

$$\min_k \big(p(C-\phi_k^T\theta) I_{[s_k=1]} + p(\phi_k^T\theta - C) I_{[s_k=0]} \big) \ge \mu.$$

Hence, $\nabla^2 l(\theta) \le -\mu \sum_{k=1}^N \phi_k\phi_k^T = -\mu A$. Coupled with $A \ge \lambda_1 I$, we can see that $\nabla^2 l(\theta) \le -\lambda_1\mu I$, which infers the strongly concave property of function $l(\theta)$ on the set $\mathscr{S}$. □

Theorem 3. Under Assumption 1, log-likelihood function $l(\theta)$ given in (2) has at most one maximum point.

Proof. Assume there exist two maximum points θ_1 and θ_2. Let $r_1 = \max(\|\theta_1\|, \|\theta_2\|)$, Lemma 2 provides the strongly concave property of $l(\theta)$ on the set $\mathscr{S} = \{\theta, \|\theta\| \le r_1\}$, which makes it impossible to come up two different maximum point on the set $\mathscr{S}$. The contradiction shows that there is at most one maximum point for $l(\theta)$. □

To reveal the condition that ML estimate exists. Some novel conditions are given.

Definition 1. Denote

$$\Psi = (\phi_1(I_{[s_1=0]} - I_{[s_1=1]}), \cdots, \phi_N(I_{[s_N=0]} - I_{[s_N=1]}))$$

as the integrate matrix which combines the information of both input data $\mathscr{I}_N$ and binary-valued observation $\mathscr{O}_N$.

Definition 2. Given input $\mathscr{I}_N$ and binary-valued observations $\mathscr{O}_N$, If there does not exist non-zero vector $\gamma \in \mathbb{R}^n$ such that $\Psi^T\gamma \ge 0$, then data $(\mathscr{I}_N, \mathscr{O}_N)$ is called effective, otherwise it is called ineffective.

Lemma 4. Under Assumption 1, if data $(\mathscr{I}_N, \mathscr{O}_N)$ is effective, then $\forall\ b \in \mathbb{R}$, the set $\{\theta, l(\theta) \ge b\}$ is bounded.

Proof. The detailed proof is in Appendix A.1.

Based on the previous discussion, we can give an explicit description for the existence and uniqueness of ML estimate.

Theorem 5. Under Assumption 1, the log-likelihood function $l(\theta)$ given in (2) has and only has one maximum point iff that data $(\mathscr{I}_N, \mathscr{O}_N)$ is effective.

Proof. We prove the theorem from two directions.

Sufficiency. Given any θ_1, the global maximum point of $l(\theta)$ is on set $\mathscr{S}_{\theta_1} = \{\theta, l(\theta) \ge l(\theta_1)\}$. From Lemma 4, $\mathscr{S}_{\theta_1}$ is bounded set, which infers the existence of maximum point of $l(\theta)$. Coupled with the result of Theorem 3, there is exactly one maximum point for $l(\theta)$.

Necessity. If $(\mathscr{I}_N, \mathscr{O}_N)$ is ineffective, there is a non-zero vector $\gamma \in \mathbb{R}^n$ such that $\Psi^T\gamma \ge 0$. Because $\Psi\Psi^T = \sum_{k=1}^N \phi_k\phi_k^T$, Assumption 1 rejects that $\Psi^T\gamma = 0$. Hence, there is at least one positive component for vector $\Psi^T\gamma$. The form of the ith-component of $\Psi^T\gamma$ is as following:

$$(\Psi^T\gamma)_i = -\phi_i^T\gamma I_{[s_i=1]} + \phi_i^T\gamma I_{[s_i=0]}.$$

Given any parameter θ, we define a scalar function $h_\theta(r)$:

$$\begin{aligned} h_\theta(r) &= l(\theta + r\gamma) \\ &= \textstyle\sum_{\{k,s_k=1\}} \log[F(-\phi_k^T\gamma r + C - \phi_k^T\theta)] \\ &\quad + \textstyle\sum_{\{k,s_k=0\}} \log[F(\phi_k^T\gamma r + \phi_k^T\theta - C)]. \end{aligned}$$

Obviously, $h_\theta(r)$ is a strictly increasing function.

Suppose θ^* is a maximum value point, $h_{\theta^*}(r) = l(\theta^* + r\gamma)$ should be non-increasing at $r = 0$, which is contradictory with the strictly increasing property of $h_{\theta^*}(r)$. So, there does not exist any finite maximum point for $l(\theta)$. □

Remark 3. Since the effective property of data is not easy to verify through Definition 2, we construct a criterion on which only needs the existence of one point at N-dimensional space.

Criterion 1. If there exists $\rho \in \mathbb{R}^N > 0$, s.t. $\Psi\rho = 0$, then data $(\mathscr{I}_N, \mathscr{O}_N)$ is effective.

Proof. Suppose the data $(\mathscr{I}_N, \mathscr{O}_N)$ is ineffective, there is a non-zero vector $\gamma \in \mathbb{R}^n$ such that $\Psi^T\gamma \ge 0$. If there exists $\rho > 0$, then $\rho^T(\Psi^T\gamma) > 0$. While because $\Psi\rho = 0$, $\rho^T\Psi^T = 0$ infers $\rho^T(\Psi^T\gamma) = 0$, which concludes the contradiction. Hence, data $(\mathscr{I}_N, \mathscr{O}_N)$ is effective. □

4. CONVERGENCE OF THE ITERATIVE ESTIMATE

In this section, we will prove the proposed iterative algorithm converging to the ML estimate with an exponential rate by using a Lyapunov method.

Assumption 2. Data $(\mathscr{I}_N, \mathscr{O}_N)$ is effective.

Remark 4. As Theorem 5 shows, if data $(\mathscr{I}_N, \mathscr{O}_N)$ is ineffective, the log-likelihood function does not have any finite maximum point. Hence, Assumption 2 is necessary on the convergence to the ML estimate. Additionally, if N is much larger than n, then the kernel of integrate matrix Ψ is an $(N-n)$ dimensional sub-space, which is very dense in N-dimensional space. This means it is probably true that Criterion 1 is satisfied.

Denote $A = \sum_{k=1}^N \phi_k\phi_k^T$ and

$$Q_1 = (1-\epsilon)^{-1}(\hat\theta_2 - \hat\theta_1)^T A(\hat\theta_2 - \hat\theta_1).$$

Then, we have the following main result, which infers that the iterative estimate $\hat\theta_t$ constructed in (4) converges to the ML estimate with exponential convergence rate.

Theorem 6. Under Assumptions 1 and 2 , the iteration $\{\hat{\theta}_t\}$ based on (4) satisfies

$$\|\hat{\theta}_t - \hat{\theta}\| \leq \sqrt{\frac{Q_1}{\lambda_{\min}(A)}} \cdot \frac{\sqrt{(1-\epsilon)}^t}{1-\sqrt{(1-\epsilon)}},$$

where $\hat{\theta}$ is the ML estimate (3), $0 < \epsilon < 1$ is a constant value dependent on the input data $\mathscr{I}_N$ and binary-valued observation $\mathscr{O}_N$, $\lambda_{\min}(A)$ is the minimal eigenvalue of A and $\|\cdot\|$ is the Euclidean norm.

Proof. Since EM algorithm increases the log-likelihood function value as the iteration goes on, $\{\hat{\theta}_t, t \geq 1\} \subset \{\theta, l(\theta) \geq l(\hat{\theta}_1)\}$ is bounded according to Lemma 4. For the simplicity, for all $1 \leq k \leq N$, define

$$g_k(x) = -f(x)\left[\frac{I_{[s_k=1]}}{F(x)} - \frac{I_{[s_k=0]}}{1-F(x)}\right].$$

Then,

$$\begin{aligned} g_k'(x) = & I_{[s_k=1]} \frac{f(x)xF(x) + f^2(x)}{F^2(x)} \\ & + I_{[s_k=0]} \frac{-f(x)x(1-F(x)) + f^2(x)}{(1-F(x))^2}. \end{aligned}$$

Hence, (4) can be transformed to

$$\begin{aligned} \hat{\theta}_{t+1} &= \hat{\theta}_t - \Big(\sum_{k=1}^N \phi_k \phi_k^T\Big)^{-1}\Big(-\sum_{k=1}^N \phi_k g_k(C - \phi_k^T \hat{\theta}_t)\Big) \\ &= \hat{\theta}_t - A^{-1}\Big(-\sum_{k=1}^N \phi_k g_k(C - \phi_k^T \hat{\theta}_t)\Big). \end{aligned}$$

Furthermore, we can see that

$$\begin{aligned} \hat{\theta}_{t+1} - \hat{\theta}_t &= \hat{\theta}_t - \hat{\theta}_{t-1} - A^{-1}\Big(-\sum_{k=1}^N \phi_k \big(g_k(C - \phi_k^T \hat{\theta}_t) \\ &\quad - g_k(C - \phi_k^T \hat{\theta}_{t-1})\big)\Big) \\ &= \hat{\theta}_t - \hat{\theta}_{t-1} - A^{-1}\Big(\sum_{k=1}^N \phi_k g_k'(C - \phi_k^T \tilde{\theta}_{kt}) \\ &\quad \cdot \big(\phi_k^T(\hat{\theta}_t - \hat{\theta}_{t-1})\big)\Big) \\ &= \Big[I_n - A^{-1}\Big(\sum_{k=1}^N \phi_k \phi_k^T g_k'(C - \phi_k^T \tilde{\theta}_{kt})\Big)\Big] \\ &\quad \cdot (\hat{\theta}_t - \hat{\theta}_{t-1}), \end{aligned} \tag{5}$$

where $\tilde{\theta}_{kt} = \lambda_{kt}\hat{\theta}_t + (1-\lambda_{kt})\hat{\theta}_{t-1}, 0 < \lambda_{kt} < 1$ for all $1 \leq k \leq N$, $t \geq 1$. From Lemma 1, $p(x) = \frac{f(x)xF(x)+f^2(x)}{F^2(x)}$ is a strictly decreasing function and $0 < p(x) < 1$. Coupled with the boundedness of $\hat{\theta}_t, t = 1, 2, \ldots$ provided by Lemma 4 and the finiteness of $\phi_k, k = 1, 2, \ldots, N$, we can obtain a lower bound ϵ satisfying $1 > \epsilon > 0$ such that for all (k, t)

$$1 > p(C - \phi_k^T \tilde{\theta}_{kt}) > \epsilon$$

and so does $p(-(C - \phi_k^T \tilde{\theta}_{kt}))$.

Let $p_{k,t}$ denote $g_k'(C - \phi_k^T \tilde{\theta}_{kt})$. Then for all (k, t),

$$\begin{aligned} p_{k,t} &= g_k'(C - \phi_k^T \tilde{\theta}_{kt}) \\ &= I_{[s_k=1]} p(C - \phi_k^T \tilde{\theta}_{kt}) + I_{[s_k=0]} p(-(C - \phi_k^T \tilde{\theta}_{kt})) \\ &\in (\epsilon, 1). \end{aligned}$$

For the simplicity, for all $t \geq 1$, we define

$$\begin{aligned} B_t &= \textstyle\sum_{k=1}^N \phi_k \phi_k^T g_k'(C - \phi_k^T \tilde{\theta}_{kt}) = \sum_{k=1}^N \phi_k \phi_k^T p_{k,t}, \\ x_t &= \hat{\theta}_{t+1} - \hat{\theta}_t. \end{aligned}$$

According to the boundedness of $p_{k,t}$ and Assumption 1, $\forall t \geq 1, A > B_t > \epsilon A > 0$. Equation (5) can be translated to the following form:

$$x_t = (I_n - A^{-1}B_t)x_{t-1}. \tag{6}$$

We prove the convergence of $\{\hat{\theta}_t\}$ by analyzing the property of a Lyapunov function $Q_t = (1-\epsilon)^{-t} x_t^T A x_t$.

$$\begin{aligned} Q_t &= (1-\epsilon)^{-t} x_t^T A x_t \\ &= (1-\epsilon)^{-t} x_{t-1}^T (I_n - A^{-1}B_t)^T A (I_n - A^{-1}B_t) x_{t-1} \\ &= Q_{t-1} + (1-\epsilon)^{-t} x_{t-1}^T \big[(I_n - A^{-1}B_t)^T A (I_n - A^{-1}B_t) \\ &\qquad -(1-\epsilon)A\big] x_{t-1} \\ &= Q_{t-1} + (1-\epsilon)^{-t} x_{t-1}^T \big[A - 2B_t + B_t A^{-1} B_t \\ &\qquad -(1-\epsilon)A\big] x_{t-1} \\ &= Q_{t-1} + (1-\epsilon)^{-t} x_{t-1}^T \big[B_t A^{-1} B_t - B_t \\ &\qquad +\epsilon A - B_t\big] x_{t-1} \end{aligned}$$

For all $t \geq 1, A > B_t > 0$, so that $A^{-1} < B_t^{-1}$, and furthermore, $B_t A^{-1} B_t < B_t B_t^{-1} B_t = B_t$. Coupled with $\epsilon A < B_t$, we can see that

$$B_t A^{-1} B_t - B_t + \epsilon A - B_t < 0 \Rightarrow Q_t \leq Q_{t-1}.$$

For all $t > 1$,

$$\begin{aligned} Q_t \leq Q_1 &\Rightarrow (1-\epsilon)^{-t} x_t^T A x_t \leq Q_1 \\ &\Rightarrow \|x_t\|^2 \leq \tfrac{Q_1}{\lambda_{\min}(A)}(1-\epsilon)^t \end{aligned}$$

where $\lambda_{\min}(A)$ is the minimal eigenvalue of matrix A. Hence,

$$\begin{aligned} \|\hat{\theta}_{t+r} - \hat{\theta}_t\| &= \|x_t + x_{t+1} + \cdots + x_{t+r-1}\| \\ &\leq \|x_t\| + \|x_{t+1}\| + \cdots + \|x_{t+r-1}\| \\ &\leq \sqrt{\tfrac{Q_1}{\lambda_{\min}(A)}}\Big[\sqrt{(1-\epsilon)}^t + \sqrt{(1-\epsilon)}^{t+1} \\ &\qquad + \cdots + \sqrt{(1-\epsilon)}^{t+r-1}\Big] \\ &= \sqrt{\tfrac{Q_1}{\lambda_{\min}(A)}} \cdot \tfrac{\sqrt{(1-\epsilon)}^t (1 - \sqrt{(1-\epsilon)}^r)}{1-\sqrt{(1-\epsilon)}} \\ &\to 0 \quad \text{as} \quad t \to \infty, r \to \infty \end{aligned}$$

Based on Cauchy criterion, $\{\hat{\theta}_t\}$ convergence to some point $\hat{\theta}$. Additionally, fix t and let $r \to \infty$

$$\|\hat{\theta} - \hat{\theta}_t\| \leq \sqrt{\frac{Q_1}{\lambda_{\min}(A)}} \cdot \frac{\sqrt{(1-\epsilon)}^t}{1-\sqrt{(1-\epsilon)}}.$$

We can see that convergence rate is of the same order with the exponential rate $\sqrt{(1-\epsilon)}^t$.

Recall that $\hat{\theta}_{t+1} = \hat{\theta}_t + A^{-1}\nabla l(\hat{\theta}_t)$. As $t \to \infty$, $\hat{\theta} = \hat{\theta} + A^{-1}\nabla l(\hat{\theta})$ infers that $\nabla l(\hat{\theta}) = 0$. This concludes that $\hat{\theta}$ is the ML estimate. In addition, by Theorem 5, $\hat{\theta}$ is the unique maximum value of log-likelihood function. □

Remark 5. Assumption 2 provides restriction on the input information and output information. Hence, the exponential convergence is not completely in the sense of probability. The more strict case such as periodic input has been

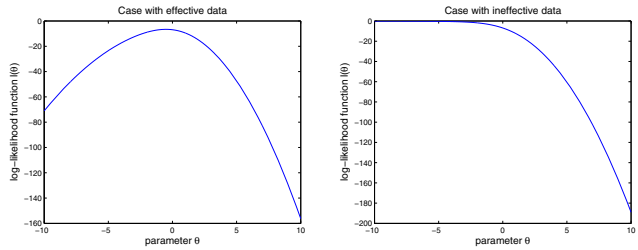


Fig. 1. Curve of the log-likelihood function $l(\theta)$: the left is the effective data case, the right is the ineffective data case.

discussed thoroughly by Wang et al. (2003), where the ML estimate can be obtained in a closed form.

5. NUMERICAL SIMULATIONS

In this section, we illustrate the main results by extensive simulations.

5.1 Log-likelihood function curve

To illustrate the log-likelihood function intuitively, we limit the model dimension to $n = 1$. In this case, the Assumption 1 degenerates to that $\sum_{k=1}^{N} \phi_k^2 > 0$. If the assumption is not satisfied, then $\forall k, \phi_k = 0$, means we cannot obtain any useful input information.

That data $\{\mathscr{I}_N, \mathscr{O}_N\}$ is ineffective is equivalent that one of $(A1)$ and $(A2)$ is true.

$$\begin{aligned}
(A1)\ &\text{For all } k \text{ that } \phi_k \geq 0,\ s_k = 1;\\
&\text{For all } k \text{ that } \phi_k \leq 0,\ s_k = 0.\\
(A2)\ &\text{For all } k \text{ that } \phi_k \leq 0,\ s_k = 1;\\
&\text{For all } k \text{ that } \phi_k \geq 0,\ s_k = 0.
\end{aligned}$$

If we generate the data based on the model, these cases hardly emerge. To illustrate the necessity of "effective property", we adopt a kind of data generate process which is nothing to do with the model.

Data Generate Process. Fix sample size $N = 10$, divide observations $\mathscr{O}_N$ equally into two parts:

(B): *s=[ones(N/2,1);zeros(N/2,1)];*

As for input data $\mathscr{I}_N$, two cases are considered:

(C1): *phi=[rand(N/2,1); -rand(N/2,1)];*

(C2): *phi=[randn(N,1)];*

(B)+(C1) corresponds to the ineffective data $(\mathscr{I}_N, \mathscr{O}_N)$, and (B)+(C2) corresponds to the effective data $(\mathscr{I}_N, \mathscr{O}_N)$. In both cases, log-likelihood function $l(\theta)$ where $\theta \in (-10, 10)$ is shown in Figure 1. The "effective property" indeed provides the existence and uniqueness of finite ML estimate.

5.2 Convergence of the proposed iterative algorithm

In this section, the convergence of the constructed algorithm (4) is illustrated by numerical simulations. The brief simulation process is as following:

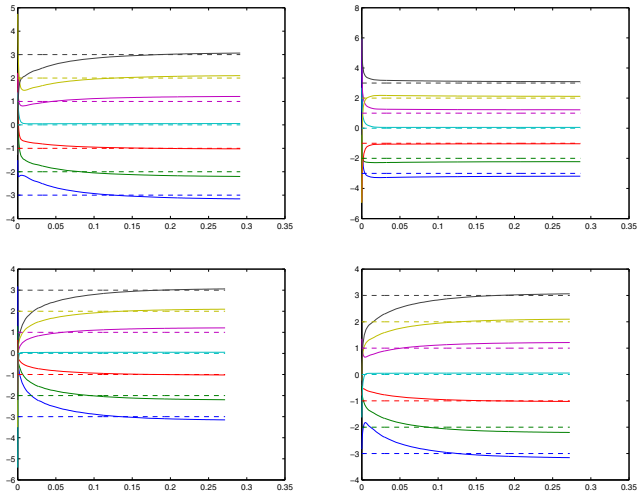


Fig. 2. The solid lines are the curves of all 7 components of estimated parameter $\hat{\theta}_t$. The dotted lines correspond to the true parameter.

Step 1: Data Generate. Fix the sample size $N = 500$, model dimension $n = 7$, sensor threshold $C = 0$, and model parameter $\theta = (-3, -2, -1, 0, 1, 2, 3)^T$. Suppose error $\mathscr{E}_N$ and input data $\mathscr{I}_N$ obey standard normal distribution. The binary-valued observations $\mathscr{O}_N$ is generated by (1).

Step 2: Initial Vector Select. To prove the EM algorithm can converge to the uniform ML estimate, under the same effective data $\{\mathscr{I}_N, \mathscr{O}_N\}$, we adopt random vector as the iterative initial vector $\hat{\theta}_1$. All components of $\hat{\theta}_1$ are generated by normal distribution with mean 0 and covariance 3.

Step 3: Parameter Estimate. Based on the initial value $\hat{\theta}_1$ and iteration process (4), we can generate the iteration estimates $\{\hat{\theta}_t, t \geq 1\}$.

The simulation results of iteration estimates are shown in Figure 2. Under various initial vectors, all components of estimates $\{\hat{\theta}_t\}$ converge to the unique ML estimate which is quite close to the true parameter. In addition, the curves of Figure 2 indicate the exponential convergence rate.

5.3 Consistence of the ML estimate

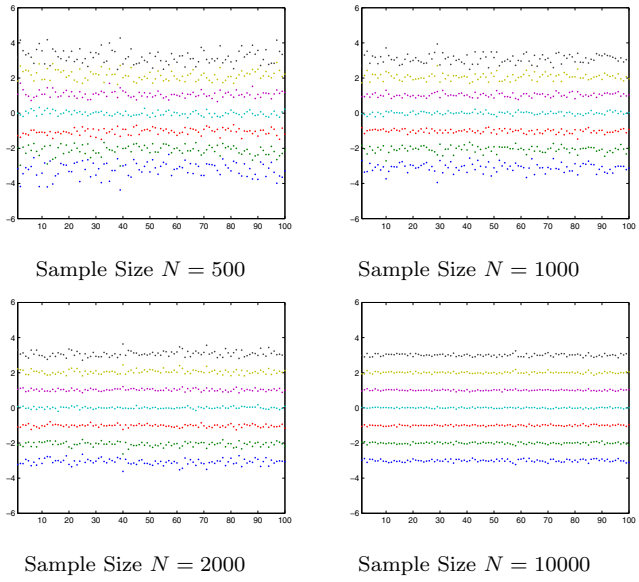
In this subsection, we illustrate the consistence of the ML estimate by numerical simulations. In other words, whether the ML estimate converges to the true parameter with the increase of sample size.

Select sample size $N = 500, 100, 1500, 2000$, respectively. In each case, we repeat the data generation and parameter estimate for 100 iterative steps.

Figure 3 shows the distribution of each component of the ML estimates. We can see that with the increase of sample size, the ML estimate converges to the true parameter in some probability meaning.

6. SUMMARY

In this paper, we have considered system identification with batched binary data through the ML criterion. The



Sample Size $N = 500$ Sample Size $N = 1000$

Sample Size $N = 2000$ Sample Size $N = 10000$

Fig. 3. The corresponding component of the ML estimates for 100 times sampling

local and global concave properties of log-likelihood function has been obtained by the negativeness of the Hessian matrix, which infers that there is at most one finite ML estimate. Furthermore, a necessary and sufficient condition for that there is unique finite solution for the ML problem has been given. An iterative algorithm is constructed to estimate the parameter and the convergence of the algorithm have been obtained, the limit of which is exactly the ML estimate. Surprisingly, the convergence has an exponential rate.

In this paper, the threshold is assumed to be known, which is not common in many cases. How to construct a algorithm to estimate the parameter and threshold simultaneously is also an attractive question. The development from finite impulse response models to general linear and even nonlinear models are also promising.

REFERENCES

Apostol, T. (1974). *Mathematical Analysis (2nd ed.)*. Addison Wesley Publishing Company.

Chen, T., Zhao, Y., and Ljung, L. (2012). Impulse response estimation with binary measurements: a regularized fir model approach. *16th IFAC Symposium on System Identification*, 113–118.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), 1–38.

Godoya, B., Goodwin, G., Agueroa, J., Marelli, D., and Wigrenb, T. (2011). On identification of fir systems having quantized output data. *Automatica*, 47, 1905–1915.

Ljung, L. (1999). *System identification (2nd ed.)*. Prentice-Hall, Englewood Cliffs NJ.

Wang, L., Zhang, J., Yin, G., and Zhao, Y.L. (2010). *System identification with quantized observations*. BirkhAauser, Boston.

Wang, L., Zhang, J., and Yin, G. (2003). System identification using binary sensors. *IEEE Transactions on Automatic Control*, 48(11), 1892–1907.

Wu, C. (1983). On the convergence properties of the em algorithm. *The Annals of Statistics*, 11(1), 95–103.

Appendix A. PROOF OF LEMMAS 1 AND 4

A.1 Proof of Lemma 4

Proof. The lemma is equivalent that for all $b < 0$, there exists an upper bound $r(b) > 0$, if $\|\theta\| \geq r(b), l(\theta) < b$. The below proof is to construct the upper bound $r(b)$.

For the simplicity, some definitions are given in advance,

$$\begin{aligned} f_{\alpha,k}(x) &= \log[F(C - \phi_k^T \alpha x)] I_{[s_k=1]} \\ &\quad + \log[F(\phi_k^T \alpha x - C)] I_{[s_k=0]}, \\ x_{b,k}(\alpha) &= \frac{C - F^{-1}(e^b)}{\phi_k^T \alpha} I_{[s_k=1]} + \frac{C + F^{-1}(e^b)}{\phi_k^T \alpha} I_{[s_k=0]}. \end{aligned}$$

where $\alpha \in \mathbb{R}^n, b, x \in \mathbb{R}, k = 1, 2, \cdots, N$. Additionally, $f_{\alpha,k}(x) < 0$ because the CDF has a upper bound 1.

Arbitrarily select an unit vector α, Definition 2 tells that there exists a k_1 such that $\phi_{k_1}^T \alpha (I_{[s_{k_1}=0]} - I_{[s_{k_1}=1]}) < 0$. Suppose $s_{k_1} = 1 (s_{k_1} = 0$ is a similar case), then $\phi_{k_1}^T \alpha > 0$, furthermore, function $f_{\alpha,k_1}(x) = \log[F(C - \phi_k^T \alpha x)]$ is a strictly decreasing function and tends to $-\infty$ as $x \to \infty$. Hence, for any given $b < 0$,

$$\begin{aligned} x \geq x_{b,k_1}(\alpha) &\Rightarrow x \geq \frac{C - F^{-1}(e^b)}{\phi_k^T \alpha} \\ &\Rightarrow C - \phi_k^T \alpha x \leq F^{-1}(e^b) \\ &\Rightarrow log(F(C - \phi_k^T \alpha x)) \leq b \\ &\Rightarrow f_{\alpha,k_1}(x) \leq b. \end{aligned}$$

For all $x \geq x_{b,k_1}(\alpha), l(x\alpha) = \sum_{k=1}^N f_{\alpha,k}(x) < f_{\alpha,k_1}(x) \leq b$. This is equivalent that for any vector θ whose corresponding unit vector is α, if $\|\theta\| \geq x_{b,k_1}(\alpha), l(\theta) < b$. That is, the set $\{\theta, l(\theta) \geq b, \frac{\theta}{\|\theta\|} = \alpha\}$ is bounded by $x_{b,k_1}(\alpha)$.

Because $\phi_{k_1}^T \alpha (I_{[s_{k_1}=0]} - I_{[s_{k_1}=1]}) < 0$, there exists an $\epsilon > 0$ and a set $A_\alpha(\epsilon) = \{\alpha_0 : \|\alpha_0\| = 1, \|\alpha_0 - \alpha\| \leq \epsilon\}$ such that for any $\beta \in A_\alpha(\epsilon), \phi_{k_1}^T \beta (I_{[s_{k_1}=0]} - I_{[s_{k_1}=1]}) < 0$. This means for any $\beta \in A_\alpha(\epsilon)$, there exists $x_{b,k_1}(\beta)$ as the bound of $\{\theta, l(\theta) \geq b, \frac{\theta}{\|\theta\|} = \beta\}$. The continuity of function $x_{b,k_1}(\alpha_0)$ and compactness of $A_\alpha(\epsilon)$ infers that $\max_{\alpha_0 \in A_\alpha(\epsilon)} x_{b,k_1}(\alpha_0) < \infty$. Define an open subset $B_\alpha = \{\alpha_0 : \|\alpha_0\| = 1, \|\alpha_0 - \alpha\| < \epsilon/2\}$ and $r(\alpha) = \max_{\alpha_0 \in B_\alpha(\epsilon)} x_{b,k_1}(\alpha_0) < \infty$. We can see that for all θ whose corresponding unit vector is within open set B_α, if $\|\theta\| \geq r(\alpha), l(\theta) < r$.

Let $S = \{\alpha : \|\alpha\| = 1\}$ denote N-dimensional sphere whose radius is 1. $\{B_\alpha, \alpha \in S\}$ is an open cover of S. Coupled with the compactness of S, there is a finite subcover $\{B_{\alpha_i}, i \leq m\}$ of S. Let r denote $\max_{1 \leq i \leq m} r(\alpha_i)$, we can see that for all θ, if $\|\theta\| \geq r, l(\theta) < b$. □