UKACC International Conference on Control 2012
Cardiff, UK, 3-5 September 2012

Counter-examples Design to Global Convergence of Maximum Likelihood Estimators

Yiqun Zou, Xiafei Tang and Zhengtao Ding

***Abstract*— MLE(Maximum Likelihood Estimation) is widely applied in system identification because of its consistency, asymptotic efficiency and sufficiency. However gradient-based optimization of the likelihood function might end up in local convergence. To overcome this difficulty, the non-local-minimum conditions are very useful. Here we suggest a heuristic method of constructing local minimum examples for ARMAX, ARARMAX and BJ models. Based on them the derivation of non-local-minimum conditions can be inspired by analyzing these examples.**

I. INTRODUCTION

Many methods have been presented in the area of system identification, such as **MLE**[1][9], frequency domain analysis[11] and subspace method[12]. Amongst them, maximum likelihood estimation is one of the most popular approaches.

The idea of **MLE** introduced by [7] and further proven by [14] is to obtain the maximum likelihood estimate $\hat{\theta}_{ML}$ through maximizing the likelihood function or minimizing its corresponding natural negative logarithmic form. An efficient method is gradient descent search [9] which is applied extensively in optimization. However the if the landscape of the objective function has at least one local minimum, the gradient search may get stuck in local convergence when badly initialized. In this case, the **MLE** will produce wrong system information. Hence so-called non-local-minimum conditions[15] have been developed to judge whether there exists any local minimum.

N.B. To clarify the difference between the global and local minimum, here we refer the local minimum to the "false" non-global minimum as in [9].

An innovative method to derive non-local-minimum conditions can be described as follows. First of all, we design so-called "local minimum examples", i.e. the particular model structures with local minima. Secondly, via tuning the model dynamics or the input signals etc of such examples in simulation, the condition which affects the local minimum existence can be tested. At last we analyze such condition theoretically in order to derive the corresponding non-local-minimum condition. In this paper we only look at the first step and make some suggestions on the design of local minimum examples.

The structure of this paper is organized as follows. Next section explains the background of **MLE**. In section 3 the general methodology of the construction of local minimum examples is provided. Section 4 shows details of local minimum examples construction for open loop **OE**, **ARMAX**, **ARARMAX** and **BJ** models respectively. In section 5, simulation examples for each model above are given. Section 6 summarizes our contributions and points out the future works.

II. Background of Maximum Likelihood Estimation

To illustrate the concept of **MLE**, Ljung [9] lets the observations represented by the random variable $y^N = (y(1), y(2), \ldots, y(t), \ldots, y(N))$ which takes values in R^N and the **PDF** (**P**robability **D**ensity **F**unction) of y^N by $f_y(\hat{\theta}, y^N)$. If the observed value of y^N is y_*^N, the probability that the observation should take the value y_*^N is proportional to

$$f_y(\hat{\theta}, y_*^N) \tag{1}$$

This is a deterministic function of $\hat{\theta}$ which is known as *the likelihood function*. A reasonable estimator $\hat{\theta}$ or explicitly $\hat{\theta}_{ML}$ can be chosen so that the observed event becomes "as likely as possible". That is

$$\hat{\theta}_{ML}(y_*^N) = \arg\max_{\hat{\theta}} f_y(\hat{\theta}, y_*^N) \tag{2}$$

where the maximization is performed for fixed y_*^N. This function is known as *maximum likelihood estimator*.

Such an estimator is reasonable because of its three advantages: firstly it provides a consistent estimate asymptotically, i.e.

$$\hat{\theta}_{ML} \to \theta \quad \text{with probability } 1 \tag{3}$$

for different model structures, e.g. [2], [4]. This property is known as consistency [14]. Secondly the covariance of $\hat{\theta}_{ML}$ is lower bounded by the inverse of Fisher information matrix, i.e.

$$E[\hat{\theta}_{ML} - \theta][\hat{\theta}_{ML} - \theta]^T \geq (-E\left[\frac{d^2}{d\hat{\theta}^2} \log f_y(\hat{\theta}, y^N)|_{\hat{\theta}=\theta}\right])^{-1} \tag{4}$$

Y. Zou is with Department of Intelligence Science and Technology, School of Information Science and Engineering, Central South University, 410083, China Email: yiqunzou@gmail.com

X. Tang is with Control Systems Centre, School of Electrical and Electronic Engineering, University of Manchester, M13 9PL, UK Email: xiafei.tang@postgrad.manchester.ac.uk

Z. Ding is with Control Systems Centre, School of Electrical and Electronic Engineering, University of Manchester, M13 9PL, UK Email: zhengtao.ding@postgrad.manchester.ac.uk

978-1-4673-1558-6/12/$31.00 ©2012 IEEE

where the equality holds asymptotically. This property is known as asymptotic efficiency [5]. Here E represents the mathematical expectation. Thirdly assume $S(y^N)$ is a sufficient statistic. According to the **Factorisation Theorem**, (1) can be rewritten to

$$f_y(\hat{\theta}, y_*^N) = \Psi(S(y_*^N), \hat{\theta})h(y_*^N) \tag{5}$$

and further transformed into its logarithmic form

$$\log f_y(\hat{\theta}, y_*^N) = \log \Psi(S(y_*^N), \hat{\theta}) + \log h(y_*^N) \tag{6}$$

Maximizing $f_y(\hat{\theta}, y_*^N)$ with respect to $\hat{\theta}$ is equivalent to maximising $\log \Psi(S(y_*^N), \hat{\theta})$ with respect to $\hat{\theta}$. This implies $f_y(\hat{\theta}, y^N)$ depends on y^N through every sufficient statistic $S(y^N)$. This property of **MLE** is sufficiency [7].

In the scope of this paper only **MLE** in open loop is considered. In Fig.1, a general open loop process is shown. We assume the system dynamics can be described by the common family of model structures[9]

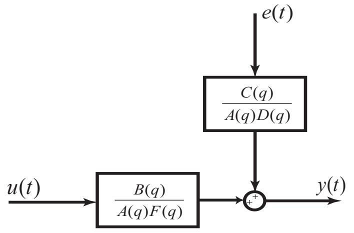

Fig. 1. General Linear Model Structure of **SISO** Open Loop Systems

$$A(q)y(t) = \frac{B(q)}{F(q)}u(t) + \frac{C(q)}{D(q)}e(t) \tag{7}$$

We also assume that the estimate model is governed by

$$\hat{A}(q)y(t) = \frac{\hat{B}(q)}{\hat{F}(q)}u(t) + \frac{\hat{C}(q)}{\hat{D}(q)}\varepsilon(t, \hat{\theta}) \tag{8}$$

where the polynomials $\hat{A}(q)$, $\hat{B}(q)$, $\hat{C}(q)$, $\hat{D}(q)$ and $\hat{F}(q)$ are rational functions characterized by

$$\hat{A}(q) = 1 + \hat{a}_1 q^{-1} + \ldots + \hat{a}_{n_a} q^{-n_a} \tag{9}$$

$$\hat{B}(q) = \hat{b}_1 q^{-1} + \ldots + \hat{b}_{n_b} q^{-n_b} \tag{10}$$

$$\hat{C}(q) = 1 + \hat{c}_1 q^{-1} + \ldots + \hat{c}_{n_c} q^{-n_c} \tag{11}$$

$$\hat{D}(q) = 1 + \hat{d}_1 q^{-1} + \ldots + \hat{d}_{n_d} q^{-n_d} \tag{12}$$

$$\hat{F}(q) = 1 + \hat{f}_1 q^{-1} + \ldots + \hat{f}_{n_f} q^{-n_f} \tag{13}$$

In the notations above, $u(t)$, $y(t)$ and $e(t)$ are the input, output and noise signal respectively. The super-index "ˆ" represents the estimates. Combining (7) and (8), the one-step-ahead prediction error $\varepsilon(t, \hat{\theta})$ can be expressed as

$$\varepsilon(t, \hat{\theta}) = \frac{\hat{D}(q)}{\hat{C}(q)}\left(\frac{\hat{A}(q)B(q)}{A(q)F(q)} - \frac{\hat{B}(q)}{\hat{F}(q)}\right)u(t) + \frac{\hat{A}(q)C(q)\hat{D}(q)}{A(q)\hat{C}(q)D(q)}e(t) \tag{14}$$

Its generation is shown in Fig. 2. The estimate coefficient vector $\hat{\theta} = [\hat{a}_1 \ldots \hat{a}_{n_a}\, \hat{b}_1 \ldots \hat{b}_{n_b}\, \hat{c}_1 \ldots \hat{c}_{n_c}\, \hat{d}_1 \ldots \hat{d}_{n_d}\, \hat{f}_1 \ldots \hat{f}_{n_f}]^T$ and the true parameter θ in a similar manner are defined. It is worthwhile to point out that if necessary we will also use the notations $[A(q)\, B(q)\, C(q)\, D(q)\, F(q)]$ and $[\hat{A}(q)\, \hat{B}(q)\, \hat{C}(q)\, \hat{D}(q)\, \hat{F}(q)]$ to represent θ and $\hat{\theta}$ respectively.

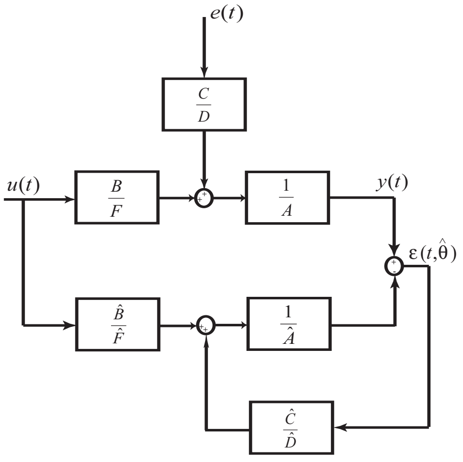
$e(t)$
$\frac{C}{D}$
$u(t)$
$\frac{B}{F}$
$\frac{1}{A}$
$y(t)$
$\varepsilon(t,\hat{\theta})$
$\frac{\hat{B}}{\hat{F}}$
$\frac{1}{\hat{A}}$
$\frac{\hat{C}}{\hat{D}}$

Fig. 2. Generation of Model Based One-step-ahead Prediction Error $\varepsilon(t, \hat{\theta})$

To clarify different model structures used in this paper, their definitions are provided in TABLE I. According to it,

Model Structures	Characteristics
ARMAX[2]	$D(q) = F(q) = 1$
ARARMAX[9]	$F(q) = 1$
OE[6]	$A(q) = C(q) = D(q) = 1$
BJ[4]	$A(q) = 1$

TABLE I

DEFINITIONS OF MODEL STRUCTURES

ARMAX and **OE** models can be produced by **ARARMAX** and **BJ** models via choosing $D(q) = 1$ and $C(q) = D(q) = 1$ respectively. For reference convenience, we omit the forward shift operator (q) and its frequency domain interpretation $(e^{j\omega})$ when there is no misunderstanding. In addition, we define an auxiliary notation

$$d(\hat{X}, \hat{Y}) \quad = \quad X(q)\hat{Y}(q) - \hat{X}(q)Y(q) \tag{15}$$

where $\hat{X}(q)$ and $\hat{Y}(q)$ are the corresponding estimate polynomials of $X(q)$ and $Y(q)$.

The following assumptions are postulated through out this paper:
(AP1). The input $u(t)$ persistently exciting of relevant system orders is deterministic and periodic or filtered white noise [13].
(AP2). The noise signal $e(t)$ and true prediction error $\varepsilon(t, \theta)$ are identical independent distributed subject to $N(0, \sigma^2)$ where the variance is known.
(AP3). The true polynomials A, C, D, F and estimated

polynomials $\hat{A}$, $\hat{C}$, $\hat{D}$, $\hat{F}$ all have the roots inside the unit circle.
(AP4). The orders of estimate polynomials are equal to the true ones, i.e. $n_a = \hat{n}_a$, $n_b = \hat{n}_b$, $n_c = \hat{n}_c$, $n_d = \hat{n}_d$ and $n_f = \hat{n}_f$.
(AP5). The number of data N goes to infinity.

Next let us derive the likelihood function for y^N which is described by the estimate model (8). Assume $\varepsilon(t,\hat{\theta})$ has the **PDF** $f_e(\varepsilon(t,\hat{\theta}),t,\hat{\theta})$. According to the joint **PDF** for the observations y^N provided in **Lemma 5.1** in [9], the likelihood function turns to be

$$f_y(\hat{\theta},y^N) = \prod_{t=1}^{N} f_e(\varepsilon(t,\hat{\theta}),t,\hat{\theta}) \tag{16}$$

Maximizing (16) is equivalent to minimizing

$$-\frac{1}{N}\log(f_y(\hat{\theta},y^N)) = -\frac{1}{N}\sum_{t=1}^{N}\log f_e(\varepsilon(t,\hat{\theta}),t,\hat{\theta}) \tag{17}$$

When $\hat{\theta} = \theta$ holds, i.e. $\varepsilon(t,\hat{\theta}) = e(t)$. Then the random function $f_e(\varepsilon(t,\hat{\theta}),t,\hat{\theta})$ turns to be a Gaussian density function. Thus (17) can be simplified into the loss function

$$V_N(t,\hat{\theta}) = \frac{1}{2N}\sum_{t=1}^{N}\varepsilon^2(t,\hat{\theta}) \tag{18}$$

The value of $V_N(t,\hat{\theta})$ is stochastic with fixed $\hat{\theta}$ and a small N. It is hard to analyze the property of $V_N(t,\hat{\theta})$ in this case. In order to avoid this difficulty, combining (AP5) we introduce the asymptotic loss function

$$V(\hat{\theta}) = \lim_{N\to\infty} V_N(t,\hat{\theta}) \quad \text{with probability 1} \tag{19}$$

in the following instead of $V_N(t,\hat{\theta})$. Note that (AP1), (AP2), (AP3) and (AP4) ensure the existence of $V(\hat{\theta})$ [13]. For convenience we replace $\lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}$ with the symbol $\bar{E}$. Then

$$V(\hat{\theta}) = \frac{1}{2}\bar{E}\varepsilon^2(t,\hat{\theta}) \tag{20}$$

holds. Since

$$\bar{E}\varepsilon^2(t,\hat{\theta}) \geq \bar{E}\varepsilon^2(t,\theta) \tag{21}$$

stands under (AP2), maximum likelihood estimator can be obtained by the minimization of (20).

III. Methodology

The methodology of local minimum design originates from [8]. Goodwin *et al* define the **DEPEN**(**D**ecreasing **E**uclidean **P**arameter **E**rror **N**orm) region as Γ in parameter space. Its elements $\hat{\theta}$ are those which will get closer to the true parameter θ in the Euclidean sense when an infinitesimal step is taken along the negative gradient direction of $V(\hat{\theta})$. For such region, it holds that
Lemma 1[8]. *$\hat{\theta} \in \Gamma$ if and only if the inner product between $\tilde{\theta} = \hat{\theta} - \theta$ and $V'(\hat{\theta})$ is positive, i.e.*

$$\tilde{\theta}^T V'(\hat{\theta}) > 0 \tag{22}$$

Proof: Let $\hat{\theta}_i$ denote the current estimate. The **SDS**(**S**teepest **D**escent **S**earch) can be expressed as

$$\hat{\theta}_{i+1} = \hat{\theta}_i - \zeta_i V(\hat{\theta}_i) \tag{23}$$

[9] where ζ_i is the step size. Subtracting θ from both sides gives

$$\tilde{\theta}_{i+1} = \tilde{\theta}_i - \zeta_i V(\hat{\theta}_i) \tag{24}$$

Squaring both sides provides

$$\tilde{\theta}_{i+1}^T\tilde{\theta}_{i+1} = \tilde{\theta}_i^T\tilde{\theta}_i - 2\zeta_i\tilde{\theta}_i^T V'(\hat{\theta}_i) + \zeta_i^2(V'(\hat{\theta}_i))^T V'(\hat{\theta}_i) \tag{25}$$

When ζ_i is sufficiently small, we could neglect the last term on the right side of (25) and prove the lemma. □

According to (22), those $\hat{\theta}$ circled by the dash-line belong to the **DEPEN** region in Fig. 3 since $\tilde{\theta}$ and $V'(\hat{\theta})$ point at the same direction. Conversely, **SDS** starting off at points like $\hat{\theta}_1$ which satisfies

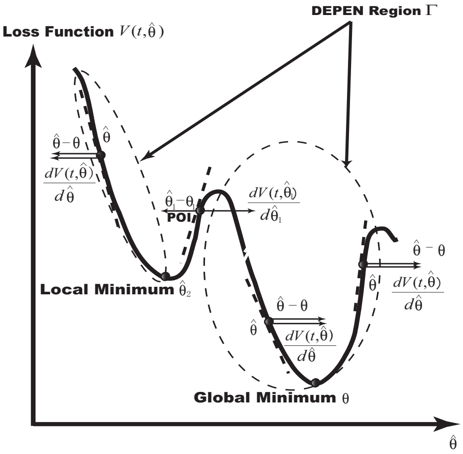
DEPEN Region Γ
Loss Function $V(t,\hat{\theta})$
$\hat{\theta}-\theta$
$\hat{\theta}$
$\frac{dV(t,\hat{\theta})}{d\hat{\theta}}$
$\hat{\theta}_1-\theta_1$
POI
$\frac{dV(t,\hat{\theta}_1)}{d\hat{\theta}_1}$
$\hat{\theta}-\theta$
$\hat{\theta}$
$\frac{dV(t,\hat{\theta})}{d\hat{\theta}}$
Local Minimum $\hat{\theta}_2$
$\hat{\theta}-\theta$
$\hat{\theta}$
$\frac{dV(t,\hat{\theta})}{d\hat{\theta}}$
Global Minimum θ
$\hat{\theta}$

Fig. 3. Illustration of **DEPEN** Region and **POI**

$$\tilde{\theta}_1^T V'(\hat{\theta}_1) \leq 0 \tag{26}$$

could *possibly* converge to the nearest local minimum $\hat{\theta}_2$. We call such $\hat{\theta}_{ini}$ *point of interest* or **POI** and the dash-line circled landscape constituted by **POI** the *region of interest*. To sum up, the construction of local minimum examples can be transformed into the design of **POI** satisfying (26) followed used in **SDS** as the initial value.

Remarks:
(1). In Fig. 3, the directions of the two arrows starting at $\hat{\theta}$ represent the sign of $\tilde{\theta}$ and $V'(\hat{\theta})$ respectively which are both scalars on the two-dimensional plot.

(2). The search may also possible converge to other kinds of stationary point, such as the saddle point.

IV. Local Minimum Examples Design

In this section we review how to construct local minimum examples for **OE** models[13] first and then implement the method described in section 3 to design local minimum examples for **ARMAX**, **ARARMAX** and **BJ** models.

For open loop **OE** models

$$y(t) = \frac{B}{F}u(t) + e(t) \tag{27}$$

according to the form of the prediction error $\varepsilon(t,\hat{\theta})$ in (14), its asymptotic loss function can be derived as

$$V(\hat{\theta}) = \frac{1}{2}\bar{E}\left[\left(\frac{B}{F} - \frac{\hat{B}}{\hat{F}}\right)u(t)\right]^2 + \frac{1}{2}\bar{E}[e(t)]^2 \tag{28}$$

Söderström suggested that optimizing (28) with respect to $\hat{b}$ firstly, $V(t,\hat{\theta})$ can be transformed into the pseudo-asymptotic function[**?**]

$$\tilde{V}(\hat{f}) = \frac{1}{2}(V_0 - V_1(\hat{f})^T V_2(\hat{f})^{-1} V_1(\hat{f})) \tag{29}$$

where

$$V_0 = \bar{E}\left[\frac{B}{F}u(t)\right]^2 + \bar{E}[e(t)]^2 \tag{30}$$

$$V_1(\hat{f}) = \bar{E}[\frac{B}{F}u(t)]\left[\frac{1}{\hat{F}}u(t-i)\right] \qquad 1 \le i \le n_b \tag{31}$$

$$V_2(\hat{f}) = \bar{E}\left[\frac{1}{\hat{F}}u(t-i)\right]\left[\frac{1}{\hat{F}}u(t-j)\right]^T \qquad 1 \le i,j \le n_b \tag{32}$$

Here $\hat{f} = [\hat{f}_1 \ \dots \ \hat{f}_{n_f}]^T$. $\left[\frac{1}{\hat{F}}u(t-i)\right]$ and $\left[\frac{1}{\hat{F}}u(t-j)\right]$ are n_f-column vectors. Since $V_2(\hat{f})$ is positive definite, the following inequality

$$\tilde{V}(\hat{f}) \le \frac{1}{2}V_0 \tag{33}$$

always holds. The equality only holds when $V_1(\hat{f}) = 0$. If the curve of $\tilde{V}(\hat{f}) = \frac{1}{2}V_0$ bisects the stability region of $\hat{f}$, normally there is at least one minimum at each side of the curve.

The **POI** of **ARMAX** and **BJ** models are more or less related to the local minimum of the **OE** models. This can be seen as follows.

Design of ***POI*** *for* ***ARMAX*** *models*: Assume the following **OE** model

$$y(t) = \frac{B(q)}{F(q)}u_1(t) + e(t) \tag{34}$$

has a local minimum at $[\hat{B}_1 \ \hat{F}_1]$. Then we design the provisional **ARMAX** model

$$A_p(q)y(t) = B_p(q)u_p(t) + C_p(q)e(t) \tag{35}$$

where

$$\begin{cases} A_p(q) = F(q) + \delta(q), \\ B_p(q) = B(q), \\ C_p(q) = F(q), \\ u_p(t) = \alpha u_1(t). \end{cases} \tag{36}$$

Here α is a positive coefficient attached before the input signal to adjust the **SNR**(**S**ignal-to-**N**oise-**R**atio)[**?**][**?**] and $\delta(q)$ is a deviation polynomial

$$\delta_1 q^{-1} + \delta_2 q^{-2} \dots + \delta_{n_a} q^{-n_a} \tag{37}$$

introduced to avoid the overlap between **OE** and **ARMAX** models. The selection of it yields to

$$\forall\sqrt{\delta_1^2 + \dots + \delta_{n_a}^2} < \xi \quad \text{s.t. } \tilde{\theta}^T V'(t,\hat{\theta}) \le 0 \,\text{at}\, [\hat{F}_1 \ \hat{B}_1 \ \hat{F}_1] \tag{38}$$

where ξ is a small positive scalar. **SDS** initialized at $\hat{\theta} = [\hat{F}_1 \ \hat{B}_1 \ \hat{F}_1]$ probably converges to the nearby local minimum $[\hat{A}_2 \ \hat{B}_2 \ \hat{C}_2]$.

Design of ***POI*** *for* ***ARARMAX*** *models:* Suppose the **ARMAX** model (35) has a local minimum $[\hat{A}_2 \ \hat{B}_2 \ \hat{C}_2]$. For the following **ARARMAX** model

$$A_p y(t) = B_p u_3(t) + \frac{C_p}{D_p}e(t), \tag{39}$$

where the input signal is

$$u_3(t) = \frac{1}{D_p}u(t), \tag{40}$$

the equalities

$$\begin{aligned} \tilde{\theta}^T V'(\hat{\theta}) &= \frac{1}{\pi}\int_0^\pi \Re(\frac{C}{\hat{C}})(|G_1|^2\Phi_{uu}(\omega) + |G_2|^2\sigma^2)d\omega \\ &= 0 \end{aligned} \tag{41}$$

[8][**?**] hold at $\hat{\theta} = [\hat{A}_2 \ \hat{B}_2 \ \hat{C}_2 \ D_p]$. Here

$$G_1 = \frac{d(\hat{B},\hat{A})}{A\hat{C}} \quad G_2 = \frac{d(\hat{C},\hat{A})}{A\hat{C}} \tag{42}$$

Therefore **SDS** initialized at such point converges to the nearby stationary point $[\hat{A}_3 \ \hat{B}_3 \ \hat{C}_3 \ \hat{D}_3]$. To ensure it is a local minimum, we only apply those D_p which let the Hessian matrix of the stationary point positive definite.

Design of ***POI*** *for* ***BJ*** *models:* It also starts from (34) which has the local minimum $[\hat{B}_1 \ \hat{F}_1]$. For the following **BJ** model

$$y(t) = \frac{B}{F}u_2(t) + \frac{C}{D}e(t), \tag{43}$$

where B and F are the same polynomials as in (34) and the input signal

$$u_2(t) = \frac{C}{D}u_1(t). \tag{44}$$

The inner product of $\tilde{\theta}$ and $V'(\hat{\theta})$ is equal to zero at $[\hat{B}_1 \ C \ D \ \hat{F}_1]$. Hence **SDS** initialized at this point $\hat{\theta} = [\hat{B}_1 \ C \ D \ \hat{F}_1]$ converges to the nearby stationary point $[\hat{B}_2 \ \hat{C}_2 \ \hat{D}_2 \ \hat{F}_2]$ differing from the global minimum. To ensure it is a local minimum point, again we only adopt those polynomials C and D which make the Hessian matrix of $[\hat{B}_2 \ \hat{C}_2 \ \hat{D}_2 \ \hat{F}_2]$ positive definite.

V. Simulation Examples

In this section we give a local minimum example for each model above. In all examples, the length of all data points N is 10000. Noise signal $e(t)$ has unit variance. For **ARMAX**, **ARARMAX** and **BJ** example, **SDS** based on (23) initialized at the **POI** is applied iteratively until the condition

$$\frac{|V(\hat{\theta}_i)-V(\hat{\theta}_{i-1})|}{V(\hat{\theta}_i)} < 0.0005 \tag{45}$$

is met or a maximum iteration number 20 is reached.

Example 1: The dynamics of the **OE** model is given as

$$\begin{cases} B = q^{-1} \\ F = 1-1.2q^{-1}+0.36q^{-2} \end{cases} \tag{46}$$

The input signal is

$$u_1(t) = (1-0.72q^{-2}+0.1296q^{-4})v_1(t) \tag{47}$$

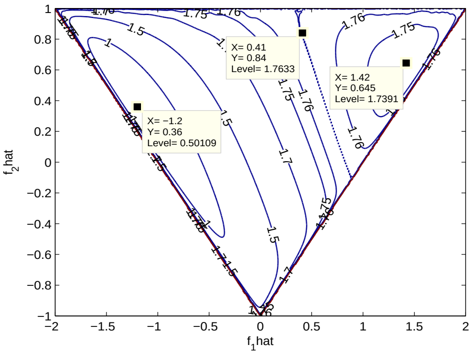

Fig. 4. The Contour of $\tilde{V}(\hat{f})$ for Example 1

Here $v_1(t)$ is i.i.d Gaussian signal with unit variance. The contour of $\tilde{V}(\hat{f})$ is shown in Fig. 4. The curve $\tilde{V}(\hat{f}) = 1.76619$ bisects the stability triangle where the coefficients $\hat{f}_1$ and $\hat{f}_2$ satisfy [10]

$$-1 < \hat{f}_2 < 1 \tag{48a}$$

$$\hat{f}_1 - 1 < \hat{f}_2 \tag{48b}$$

$$-\hat{f}_1 - 1 < \hat{f}_2 \tag{48c}$$

into two subsets. Each subset has one minimum. They are

$$\theta = [1\ -1.200\ 0.360]^T \quad \text{(global minimum)} \tag{49a}$$

$$\hat{\theta}_1 = [-0.121\ 1.419\ 0.670]^T \quad \text{(local minimum)} \tag{49b}$$

Example 2: Given an **ARMAX** model where

$$\begin{cases} A_p(q) = (1-1.2q^{-1}+0.36q^{-2})+\delta(q) \\ B_p(q) = q^{-1} \\ C_p(q) = 1-1.2q^{-1}+0.36q^{-2} \end{cases} \tag{50}$$

The input signal of the system is

$$u_p(t) = 0.3u_1(t) \tag{51}$$

where $u_1(t)$ is the input signal used in example 1. At

$$\hat{\theta} = [1.419\ 0.670\ -0.121\ 1.419\ 0.670]^T \tag{52}$$

selecting

$$\delta(q) = 0.2q^{-1}+0.09q^{-2} \tag{53}$$

makes

$$\tilde{\theta}^T V'(\hat{\theta}) = -0.0305 \tag{54}$$

which implies (38) satisfied. Starting from (52), **SDS** converges to the following stationary point

$$\begin{cases} \hat{A}_2 = 1+1.304q^{-1}+0.5435q^{-2} \\ \hat{B}_2 = -0.1035q^{-1} \\ \hat{C}_2 = 1+1.386q^{-1}+0.6022q^{-2} \end{cases} \tag{55}$$

after four iterations. Its Hessian matrix

$$\begin{bmatrix} 7.07 & -5.99 & 0.14 & -7.94 & 6.50 \\ -5.99 & 7.07 & -0.05 & 7.16 & -7.94 \\ 0.14 & -0.05 & 0.33 & -0.16 & 0.02 \\ -7.94 & 7.16 & -0.16 & 9.11 & -7.92 \\ 6.50 & -7.94 & 0.02 & -7.92 & 9.11 \end{bmatrix}$$

is positive definite. The trace of loss function in **SDS** is shown in Fig. 5.

Example 3: For such **ARARMAX** model which has the same A_p, B_p and C_p with in example 2, we assign its D_p and input signal as

$$\begin{cases} D_p = 1+0.8257q^{-1} \\ u_3(t) = \dfrac{1}{D_p}u_p(t) \end{cases} \tag{56}$$

SDS starting at $\hat{\theta} = [\hat{A}_2\ \hat{B}_2\ \hat{C}_2\ D_p]$ eventually ends at the stationary point

$$\begin{cases} \hat{A}_3 = 1+1.312q^{-1}+0.5464q^{-2} \\ \hat{B}_3 = -0.1035q^{-1} \\ \hat{C}_3 = 1+1.377q^{-1}+0.6009q^{-2} \\ \hat{D}_3 = 1+0.815q^{-1} \end{cases} \tag{57}$$

after five iterations. The Hessian matrix of this point

$$\begin{bmatrix} 7.06 & -5.98 & 0.14 & -7.67 & 6.37 & 4.72 \\ -5.98 & 7.06 & -0.05 & 6.73 & -7.67 & -3.85 \\ 0.14 & -0.05 & 0.32 & -0.16 & 0.03 & 0.12 \\ -7.67 & 6.73 & -0.16 & 8.43 & -7.25 & -5.03 \\ 6.37 & -7.67 & 0.03 & -7.25 & 8.43 & 4.11 \\ 4.72 & -3.85 & 0.12 & -5.03 & 4.11 & 4.13 \end{bmatrix}$$

is positive definite. The trace of loss function in **SDS** is shown in Fig. 6.

Example 4: The true polynomials of this **BJ** model are

$$\begin{cases} B = q^{-1} \\ C = 1+0.6428q^{-1} \\ D = 1-0.6616q^{-1}+0.1792q^{-2} \\ F = 1-1.2q^{-1}+0.36q^{-2} \end{cases} \tag{58}$$

Its input signal is

$$u_2(t) = \frac{C}{D}u_1(t) \tag{59}$$

Let **SDS** begins at $\hat{\theta} = [\hat{B}_1 \; C \; D \; \hat{F}_1]$. After eight steps of iteration, **SDS** ends up at the following local minimum point

$$\begin{cases} \hat{B}_2 = -0.0804q^{-1} \\ \hat{C}_2 = 1 + 0.6927q^{-1} \\ \hat{D}_2 = 1 - 1.062q^{-1} + 0.4294q^{-2} \\ \hat{F}_2 = 1 + 1.398q^{-1} + 0.6252q^{-2} \end{cases} \tag{60}$$

Its Hessian matrix

$$\begin{bmatrix} 7.13 & -1.95 & -0.37 & -0.94 & -0.41 & -0.19 \\ -1.95 & 6.09 & -1.39 & 0.99 & 0.32 & -0.17 \\ -0.37 & -1.39 & 7.57 & 5.61 & 0.29 & 0.11 \\ -0.94 & 0.99 & 5.61 & 7.57 & 0.09 & -0.44 \\ -0.41 & 0.32 & 0.29 & 0.09 & 0.24 & -0.23 \\ -0.19 & -0.17 & 0.11 & -0.44 & -0.23 & 0.66 \end{bmatrix}$$

is positive definite. The trace of loss function in **SDS** is shown in Fig. 7.

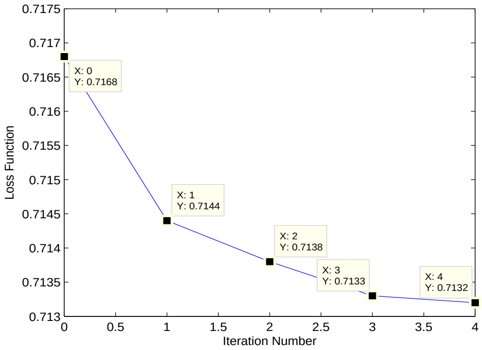

Fig. 5. Evaluation of Asymptotic Loss Function $V(\hat{\theta})$ in **SDS** for Example 2

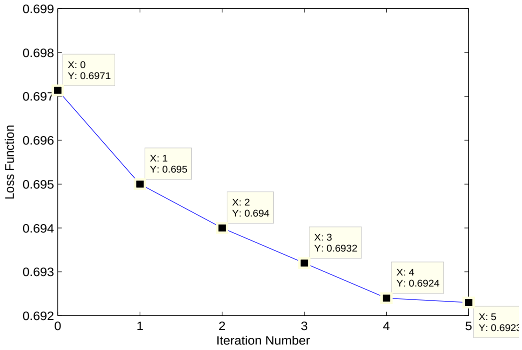

Fig. 6. Evaluation of $V(\hat{\theta})$ in **SDS** for Example 3

VI. Conclusions and Future Works

In this paper, we design the **POI** followed by **SDS** to construct local minimum examples for open loop **ARMAX**, **ARARMAX** and **BJ** models. In particular, the **POI**s for **ARMAX** and **BJ** models have strong links to the local minimum in the corresponding **OE** models. Furthermore, simulation examples are also provided for each model structure above. These examples play as a key to the development of the non-local-minimum conditions in **MLE**. In the future, we will investigate the non-local-minimum conditions for these models starting by changing the examples dynamics etc.

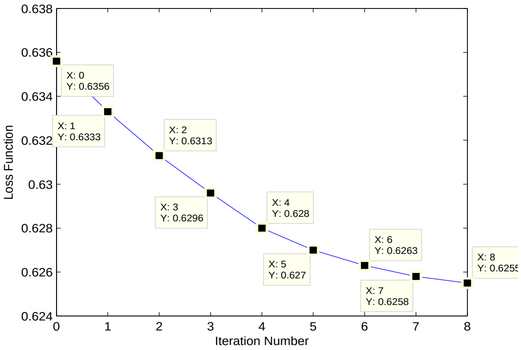

Fig. 7. Evaluation of $V(\hat{\theta})$ in **SDS** for Example 4

VII. Acknowledgments

The authors would like to thank Dr. Daniel Coca and Dr. Tim Breikin for their constructive comments on the way of discussing the simulation examples.

References

[1] Åström, K.J. *Maximum Likelihood and Prediction Error Methods*. Automatica, 16:551–574, 1980.

[2] Åström, K.J. and Bohlin, T. *Numerical Identification of Linear Dynamic Systems from Normal Operating Records*. In IFAC Symposium on Self-Adaptive Systems, Teddington, UK, 1965.

[3] Åström, K.J. and Söderström, T. *Uniqueness of Maximum Likelihood Estimates of the Parameters of an ARMA model*. IEEE Transactions on Automatic Control, 19(6):769-773, 1974

[4] Box, G.E.P. and Jenkins, G.M. *Time Series Analysis, Forecasting and Control(3rd ed. 1994)*. Holden-Day, 1970.

[5] Cramér, H. *Mathematical Methods of Statistics*. Princeton University Press, Princeton, 1946.

[6] Dugard, L. and Landau, I.D. *Recursive Output Error Identification Algorithms*. Automatica, 16:443-462, 1980. Messenger of Mathmatics, 41:155–160, 1912.

[7] Fisher, R.A. *On the Mathematical Foundations of Theoretical Statistics*. Philosophical Transactions of the Royal Society of London. Series A, 222:309–368, 1921.

[8] Goodwin, G.C., Carlos, J.A. and Skelton, R.E. *Conditions for Local Convergence of Maximum Likelihood Estimation for ARMAX Models*. In 13th IFAC Symposium on System Identification, Rotterdam, The Netherland, 2003.

[9] Ljung, L. *System Idenfication:Theory for the User, 2nd Edition*. Prentice Hall, 1999.

[10] Ogata, K. *Modern Control Engineering, 3rd Edition*. Prentice Hall, 1996.

[11] Pintelon, R. and Schoukens, J. *System Identification: A Frequency Domain Approach*. IEEE Press, 2001.

[12] van Overschee, P. and DeMoor, B. *Subspace Identification for Linear Systems: Theory, Implementation, Applications*. Kluwer Academic Publishers, 1996.

[13] Söderström, T. *On the Uniqueness of Maximum Likelihood Identification*. Automatica, 11:193–197, 1975.

[14] Wald, A. *Note on the Consistency of the Maximum Likelihood Estimate*. The Annals of Mathematical Statistics, 20(4):595–601, 1949.

[15] Zou, Y. and Heath, W.P. *Global Convergence Conditions in Maximum Likelihood Estimation*. International Journal of Control, DOI: 10.1080/00207179.2012.658085, 2012.