State dimension reduction and analysis of quantized estimation systems

Hui Zhang a,⁎, Ying Shen b, Li Chai c, Li-Xin Gao d

a State Key Laboratory of Industrial Control Technology, Institute of Industrial Process Control, Department of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China
b School of Engineering, Huzhou Teachers College, Huzhou 31300, China
c Engineering Research Center of Metallurgical Automation and Measurement Technology, Wuhan University of Science and Technology, Wuhan 430081, China
d Institute of Intelligent Systems and Decision, Wenzhou University, Zhejiang 325027, China

ARTICLE INFO

Article history:
Received 28 August 2013
Received in revised form 14 April 2014
Accepted 10 June 2014
Available online 23 June 2014

Keywords:
Communication constraint
Quantized linear system
Quantizer design
State dimension reduction
Estimability
Mutual information

ABSTRACT

The problem of state dimension reduction and quantizer design under communication constraints is discussed for state estimation in quantized linear systems. Subject to the limited signal power, number and bandwidth of the parallel channels, a differential pulse code modulation (DPCM)-like structure is adopted to generate the quantized innovations as the transmitted signals, and the multi-level quantized Kalman filter (MLQ-KF) is used to serve as the pre- and post-filters. The dimension reduction matrix and quantizer are designed jointly under the MMSE criterion of estimation at the channel receiver. To demonstrate the validity of state estimation under the adopted framework, the state estimability based on quantized innovations is analyzed by using information theoretic method. This leads to a sufficient and necessary condition of a certain estimability Gramian matrix having full rank. The quantized Gramian is proved to converge to that of the original unquantized system when the quantization intervals turn to zero. Our work also provides an auxiliary analytic support for the estimation under 1-bit quantization. Simulations show that under communication constraints, the estimation performance is satisfactory when the designed dimension reduction method and quantizer are applied. The analytic conclusion of estimability is also verified by illustrative simulations.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Recently there have been plenty of works concerning the quantized estimation problem, e.g. [1–7]. A modified Kalman filter called SOI-KF was proposed by using the sign of innovations or the 1-bit quantization of innovations [1]. The filtering problem based on multi-level quantizers was studied in [2–4], where the optimal filters and the multi-level quantization schemes are designed simultaneously. Lu et al. [5] and Liu et al. [6] discussed the problem of state estimation under quantization, while Farias and Brossier [7] considered the quantizer design for estimation. Another important methodology [40–42] dealing with quantized estimation adopts the so-called expectation maximization (EM) algorithm involving the Monte Carlo (MC) techniques. In this paper, we address the problem of joint state dimension reduction and quantizer design for state estimation in linear dynamical systems with a differential pulse code modulation.
In fact, the dimension reduction, pre-filter and quantizer designing the predictive quantizer under the Gaussian Ramabadran and Sinha [12] and Arildsen et al. [15] without considering state dimension reduction, while the classic book [13] Curry provided the methods of additive quantization noise model (LAQNM). Specially, in Msechu et al. [2], You et al. [3], Fu and de Souza [4] and Ramabadran and Sinha [12] and Arildsen et al. [15].

In the pre-processing part, the generalized quantization the recently proposed approaches involving predictor MLQ-KF [3] under GA is adopted in our system as both [18] and Khaled et al. [19] the dimension reduction matrix approximate the input property of the original system as precise as possible, whereas the later is looking for a compression matrix which reduces the quantity of transmission data while making the state estimation at the receiver as precise as possible. Contributions of Zhu et al. [10] and Goela and Gastpar [11] consider the communication in network setting by using reduced-dimension transformation to deal with the bandwidth limitation, while our work focuses on the state estimation under reduced-dimension transmission where the bandwidth limitation is further dealt with by quantization.

DPCM is an effective way of reducing the redundancy of transmission data [12–15]. Our work bases on a DPCM-like scheme where the quantized innovations are transmitted. There are a pre-filter which generates innovations to be transmitted and a post-filter which estimates the states at the receiver of noiseless channel. The present structure is different from the traditional DPCM in that, in our system the estimation performance is pursued partly by designing the dimension reduction matrix (which can also be considered as a designable system output matrix). Our design work is also enlightened by but different from Msechu et al. [2], You et al. [3], Fu and de Souza [4], Curry [13], Ramabadran and Sinha [12] and Arildsen et al. [15]. Msechu et al. [2], You et al. [3], Fu and de Souza [4] and Curry [13] design the quantized estimator and quantizer without considering state dimension reduction, while Ramabadran and Sinha [12] and Arildsen et al. [15] consider the coder-reduction (measurements selection) and decoder (filter) design with a representation of linear additive quantization noise model (LAQNM). Specially, in the classic book [13] Curry provided the methods of designing the predictive quantizer under the Gaussian assumption (GA) [13, Section 3.3], and the joint optimization of quantizer and filter under LAQNM [13, Section 4.5]. In fact, the dimension reduction, pre-filter and quantizer play jointly the role of preprocessing or coding, and the post-filter is the decoder. In this sense, the design problem in the present paper is similar to the preceding design for fixed receiver [16], joint source-channel coding [17], and joint design of precoder and decoder [18,19]. Different from Wiesel et al. [16], Persson et al. [17], Sampath et al. [18] and Khaled et al. [19] the dimension reduction matrix and quantizer are involved here for design, and the MLQ-KF [3] under GA is adopted in our system as both of the pre- and post-filters. Our scheme is also similar to the recently proposed approaches involving predictor in the pre-processing part, the generalized quantization scheme (GQS) [42] and its special form of the noise shaping quantizer (NSQ) [43], but different from them in two points. One point is that in our scheme the adopted MLQ-KF [3] (see Section 2.3) provides analytic formula of the estimator, while in [42] the EM algorithm is adopted based on the MC technique. The other point is in our scheme the dimension reduction matrix is designed jointly with the quantizer levels. The reduction matrix can also be considered as a static filter under the framework of the GQS [42].

Although many works concerning the quantized estimation algorithm and performance have been published, a problem is slightly ignored: the validity or effectiveness of the estimation may be destroyed by improper quantization scheme. Few contributions concerning the property of state reconstruction under quantization were found. In [45] the conditions of input and quantization design were proposed for guaranteeing the active observability under binary valued observations. Here, we focus on the estimability property of quantized systems which is important from both the system theory point of view and the engineering point of view. This is another topic of this paper. The concept of estimability was proposed by Baram and Kailath to evaluate the validity of state estimation under the minimum mean square error (MMSE) criterion in [20], where it is stated that the system is estimable if and only if the posterior estimation error is strictly less than the prior estimation error. The estimability was also discussed from information theoretic viewpoint [21–23] based on the minimum maximum error entropy (MMEE) estimation [24–26]. Referring to this methodology, the present paper discusses the estimability given quantized innovations, which is different from that in [23] concerning quantized outputs. Our discussion involves the ‘true’ quantizer. This is different from that based on the assumption of additive uniform noise for approximating quantization [43,46].

In short, this paper consists of two main works: the optimal design of state dimension reduction and quantizer for state estimation in linear systems under limited channel (or subchannel) number, bandwidth and power of transmitted signal; the estimability analysis given quantized innovations. Under the MMSE criterion, the reduction matrix and quantizer are designed jointly based on a DPCM-like structure and the estimator of MLQ-KF. Simulation results show that the method presented in this paper has satisfactory estimation performance. The analysis work leads to an algebraic condition of estimability for quantized linear Gaussian systems, i.e. certain Gramian matrix having full rank, and shows the quantized Gramian converges to that of the original unquantized system when the quantization intervals turn to zero. The relation between the present result and that in [23] is discussed. Analytical results are also illustrated by simulations. The design work was presented primarily in the conference paper [27], and the estimability analysis is an extension of [28,23].

The rest of the paper is organized as follows: Section 2 introduces the system and the problem of state preprocessing, reviews the quantizer and the MLQ-KF; The state dimension reduction and quantizer are designed jointly under MMSE criterion in Section 3; Section 4 introduces
the concept of estimability, gives analytic results based on quantized innovations, and proves the convergence of quantized estimability Gramian; Sections 5 and 6 are illustrative simulations and conclusion, respectively.

Throughout this paper, we use $Y^k$ to denote a vector $[y_1^k(k_0) \ldots y_{[n]}^k(k_0 + 1)]^T$ which consists of a sequence of $y(k), k \geq k_0$ of vectors (or variables), where $k$ denotes the time instants. The initial instant can be set as $k_0 = 0$ for time-invariant systems without losing generality. We use $\hat{x}(j|k)$ to denote the MMSE or MMEE estimation of system state $x(j)$ at instant $j$ given data $Y^k$. Other notations are as follows. $\mathbb{R}$, $\mathbb{R}^n$ denote the set of all real numbers and the $n$-dimensional real space, respectively. $A^T$, $\text{tr}(A)$ and $\text{rank}(A)$ denote respectively the transposition, trace and rank of a matrix (or vector) $A$. For random variables or vectors $x$ and $y$, $p(x)$ and $p(x|y)$ denote the probability density of $x$ and that of $x$ given $y$, respectively; $E[x]$ and $E[x|y]$ denote respectively the mathematical expectation of $x$ and the conditional expectation of $x$ given $y$; $H(x)$, $H(x|y)$ and $I(x; y)$ denote the entropy of $x$, the conditional entropy of $x$ given $y$ and the mutual information between $x$, $y$, respectively.

2. System design framework

2.1. System state

The system state is modeled as

$$x(k + 1) = Ax(k) + Bw(k)$$

(1)

where $x(k) \in \mathbb{R}^n$, $w(k) \in \mathbb{R}^m$; $A, B$ are constant matrices with appropriate dimensions; $w(k)$ is a zero-mean white Gaussian random vector with constant covariance $Q$, and is uncorrelated with the initial state $x(0)$ which is Gaussian with zero mean and bounded covariance. The state $x(k)$ is measured and preprocessed before transmission, and reconstructed at the receiver.

2.2. State preprocessing

The channel under discussion is assumed to be the parallel channel [18,19,29], which consists of $p$ independent noiseless subchannels with limited bandwidth and a power constraint on the transmitted signal, where $p$ is generally assumed to be smaller than $n$ (the dimension of system state). The number of parallel data streams is then required to be not greater than $p$. This will be met by state dimension reduction. We reduce the state dimension linearly, namely generate a linear combination of state measurements by aggregation. In general, the sensing is affected by thermal and other unobservable tiny noises, and the reduction error is inevitable. In areas of communication and data compression, independent white Gaussian noise is commonly used to approximate kinds of noise, uncertainty and disturbance [30,31]. Then in our problem, various uncertainties caused by sensing and dimension reduction are also approximated by an additive white Gaussian process, according to the central limit theorem [32]. In this way, the state dimension reduction can be described as

$$y(k) = Cx(k) + w(k)$$

(2)

where $y(k) \in \mathbb{R}^p$ is the output of state dimension reduction; $C$ is the reduction matrix to be designed; $w(k) \in \mathbb{R}^p$ is a Gaussian white noise with zero-mean and constant covariance $R$, and is uncorrelated with $w(k)$ and $x(0)$. In the design problem, we consider the case of $p = 1$, i.e. $y(k),v(k) \in \mathbb{R}$.

Due to the bandwidth constraint, quantization is involved. We adopt a DPCM-like framework, where the innovations are generated and quantized for transmission:

$$\hat{y}(k) = y(k) - \hat{y}(k - 1)$$

(3)

where $\hat{y}(k - 1) = E[y(k)|\hat{Y}^k_q]$ is the prediction of $y(k)$ given $\hat{Y}^k_q = [\hat{y}(0) \ldots \hat{y}(k - 1)]^T$ which contains the quantized innovations from instant 0 to $k - 1$. The designable quantizer is a symmetrical one, by which the innovation is partitioned into $2L + 1$ contiguous and non-overlapping intervals:

$$\hat{y}(k) = Q_{a_L}(\hat{y}(k)) = \begin{cases} a_L, & a_L < \hat{y}(k) \\ \vdots & \vdots \\ a_1, & a_1 < \hat{y}(k) \leq a_2 \\ \vdots & \vdots \\ a_0, & 0 < \hat{y}(k) \leq a_1 \\ -Q_{a_L}(-\hat{y}(k)), & \hat{y}(k) < 0 \end{cases}$$

(4)

where $Q_{a_L}(-\hat{y}(k))$ denotes the quantized $\hat{y}(k)$, $Q_{a_L}(\cdot)$ is the symmetrical quantizer, $\{a_l, a_{l + 1}, \ldots, a_{l - 1}, a_0, a_1, \ldots, a_{l + 1}\}$ are the thresholds, $0 = a_0 < a_1 < \cdots < a_L < +\infty = a_{L + 1}$. We use $a^L$ to denote the set of thresholds $\{a_1, a_2, \ldots, a_L\}$.

In contrast to the case of transmitting states directly, data redundancy is reduced when innovations are transmitted: $\hat{Y}^k_q$ is the output of state dimension reduction; $C$ is the reduction matrix to be designed; $w(k) \in \mathbb{R}^p$ is a Gaussian white noise with zero-mean and constant covariance $R$, and is uncorrelated with $w(k)$ and $x(0).$ In the design problem, we consider the case of $p = 1$, i.e. $y(k),v(k) \in \mathbb{R}$.

Due to the bandwidth constraint, quantization is involved. We adopt a DPCM-like framework, where the innovations are generated and quantized for transmission:

$$\hat{y}(k) = y(k) - \hat{y}(k - 1)$$

(3)

where $\hat{y}(k - 1) = E[y(k)|\hat{Y}^k_q]$ is the prediction of $y(k)$ given $\hat{Y}^k_q = [\hat{y}(0) \ldots \hat{y}(k - 1)]^T$ which contains the quantized innovations from instant 0 to $k - 1$. The designable quantizer is a symmetrical one, by which the innovation is partitioned into $2L + 1$ contiguous and non-overlapping intervals:

$$\hat{y}(k) = Q_{a_L}(\hat{y}(k)) = \begin{cases} a_L, & a_L < \hat{y}(k) \\ \vdots & \vdots \\ a_1, & a_1 < \hat{y}(k) \leq a_2 \\ \vdots & \vdots \\ a_0, & 0 < \hat{y}(k) \leq a_1 \\ -Q_{a_L}(-\hat{y}(k)), & \hat{y}(k) < 0 \end{cases}$$

(4)

where $Q_{a_L}(-\hat{y}(k))$ denotes the quantized $\hat{y}(k)$, $Q_{a_L}(\cdot)$ is the symmetrical quantizer, $\{a_l, a_{l + 1}, \ldots, a_{l - 1}, a_0, a_1, \ldots, a_{l + 1}\}$ are the thresholds, $0 = a_0 < a_1 < \cdots < a_L < +\infty = a_{L + 1}$. We use $a^L$ to denote the set of thresholds $\{a_1, a_2, \ldots, a_L\}$.

In contrast to the case of transmitting states directly, data redundancy is reduced when innovations are transmitted, and the signal power is cut down. The power of innovation is

$$\sigma^2(k) = E[\hat{y}^2(k)].$$

(5)

The channel is assumed to be lossless. The receiver estimates $x(k)$ by using quantized innovations $\hat{Y}^k_q$. The predictor and the filter will be introduced in the following context. The system with a DPCM-like structure is shown in Fig. 1.

2.3. MLQ-KF

For system (1)–(4), the MLQ-KF [3] is adopted as both the pre- and post-filters (i.e., the predictor and the filter in Fig. 1). The MLQ-KF is an MMSE estimator which seeks the pre- and post-filters (i.e., the predictor and the filter in Fig. 1) of the innovation.

The channel is assumed to be lossless. The receiver estimates $x(k)$ by using quantized innovations $\hat{Y}^k_q$. The predictor and the filter will be introduced in the following context. The system with a DPCM-like structure is shown in Fig. 1.

Fig. 1. System structure.
levels turns to infinite. Equations of the MLQ-KF algorithm used here are as follows:

One-step-ahead predictions of the state and output at time instant $k$:

$$\hat{x}(k|k-1) = A\hat{x}(k-1|k-1),$$

$$y(k|k-1) = Cx(k|k-1).$$

Estimation of the state at time instant $k$:

$$\hat{x}(k|k) = \hat{x}(k|k-1) + \frac{f(\mathbf{a}^l, \tilde{y}_q(k))P(k|k-1)\mathbf{C}^T}{\sqrt{CP(k|k-1)\mathbf{C}^T + R}}.$$  

Covariance of the state prediction error at time instant $k$:

$$P(k|k-1) = AP(k|k-1|k-1)A^T + Q.$$  

Covariance of the state estimation error at time instant $k$:

$$P(k|k) = P(k|k-1) - F(\mathbf{a}^l)P(k|k-1)\mathbf{C}^T CP(k|k-1)\mathbf{C}^T + R.$$  

In these equations: $\tilde{y}_q(k)$ is produced by (3) and (4):

$$f(\mathbf{a}^l, \tilde{y}_q(k)) = \sum_{i=1}^{I} I_{a(l)}(\tilde{y}_q(k)) \frac{\phi(a_i/\sigma_q(k)) - \phi(a_{i+1}/\sigma_q(k))}{T(a_i/\sigma_q(k)) - T(a_{i+1}/\sigma_q(k))},$$

$$F(\mathbf{a}^l) = 2 \sum_{i=0}^{I} \frac{(\phi(a_i/\sigma_q(k)) - \phi(a_{i+1}/\sigma_q(k)))^2}{T(a_i/\sigma_q(k)) - T(a_{i+1}/\sigma_q(k))},$$

where $I_{a(l)}(\cdot)$ is a standard sign function:

$$I_{a(l)}(x) = \begin{cases} 1, & x = a_l, \\ 0, & \text{otherwise}. \end{cases}$$

$\phi(\cdot)$ denotes the probability density function of a standard Gaussian variable with zero mean and unit covariance $\mathbf{I}$, and $T(\cdot)$ denotes its probability distribution function; $\sigma_q^2(k) = \mathbf{CP}(k|k-1)\mathbf{C}^T + R$ is the innovation covariance, the innovation $\tilde{y}(k)$ is defined as (3).

3. Joint design of reduction matrix and quantizer

There are three factors affecting the state reconstruction performance: the dimension reduction, the quantizer and the quantized filter. When the filter is defined, the design problem is then to look for the dimension matrix $C$ and the quantization algorithm under preferred optimal principle of state estimation at the channel receiver. There is no reason to think uncritically that independently designed reduction and quantizer will be optimal for state estimation. This leads to a joint optimization problem stated as follows.

3.1. Design method

It is expected to recover the states from the quantized innovations. Under MMSE principle, we look for a linear dimension reduction matrix and a quantizer which make $tr(P(k|k))$ be minimal. The effect of noise can be attenuated by increasing the power of the useful signal. However the signal power is always limited by measurement equipment and other practical factors [8]. Many works characterized this limitation in terms of signal-to-noise ratio (SNR) constraints [16,33,34]. Instead of this approach, here we assume a power constraint $W$ on the transmitted innovations as in [18]:

$$\sigma_q^2(k) \leq W.$$  

Thus our problem boils down to the problem as follows:

$$\begin{array}{ll}
\min & \text{tr}(P(k|k)) \\
\text{s.t.} & \sigma_q^2(k) \leq W.
\end{array}$$  

Constraint (15) is equivalent to

$$C(k)P(k|k-1)C(k)^T + R \leq W.$$  

The optimal estimation performance is expected when (18) is satisfied. This will be implied in Section 3.2.

We employ the iterative method to solve this problem: compute the dimension reduction matrix $C(k)$ and quantizer $a^l(k) = [a_1(k), a_2(k), \ldots, a_{l}(k)]$ at every time instant $k$, which are expected to approach to a constant matrix when (18) is satisfied. This will be implied in Section 3.2.

3. Joint design of reduction matrix and quantizer

There are three factors affecting the state reconstruction performance: the dimension reduction, the quantizer and the quantized filter. When the filter is defined, the design problem is then to look for the dimension matrix $C$ and the quantization algorithm under preferred optimal principle of state estimation at the channel receiver. There is no reason to think uncritically that independently designed reduction and quantizer will be optimal for state estimation. This leads to a joint optimization problem stated as follows.

3.1. Design method

It is expected to recover the states from the quantized innovations. Under MMSE principle, we look for a linear dimension reduction matrix and a quantizer which make $tr(P(k|k))$ be minimal. The effect of noise can be attenuated by increasing the power of the useful signal. However the signal power is always limited by measurement equipment and other practical factors [8]. Many works characterized this limitation in terms of signal-to-noise ratio (SNR) constraints [16,33,34]. Instead of this approach, here we assume a power constraint $W$ on the transmitted innovations as in [18]:

$$\sigma_q^2(k) \leq W.$$  

Thus our problem boils down to the problem as follows:

$$\begin{array}{ll}
\min & \text{tr}(P(k|k)) \\
\text{s.t.} & \sigma_q^2(k) \leq W.
\end{array}$$  

Constraint (15) is equivalent to

$$C(k)P(k|k-1)C(k)^T + R \leq W.$$  

The optimal estimation performance is expected when (18) is satisfied. This will be implied in Section 3.2.

We employ the iterative method to solve this problem: compute the dimension reduction matrix $C(k)$ and quantizer $a^l(k) = [a_1(k), a_2(k), \ldots, a_{l}(k)]$ at every time instant $k$, which are expected to approach to a constant matrix $C$ and a threshold set $a^l$ of a time-invariant symmetrical quantizer, respectively.

From (9),

$$tr(P(k|k)) = tr(P(k|k-1)) - F(\mathbf{a}^l(k))tr\left(\frac{P(k|k-1)C(k)^T C(k)P(k|k-1)}{C(k)P(k|k-1)C(k)^T + R}\right)$$

$$= tr(P(k|k-1)) - F(\mathbf{a}^l(k))\frac{C(k)P(k|k-1)C(k)^T}{C(k)P(k|k-1)C(k)^T + R}.$$  

At time instant $k$, $P(k|k-1)$ is known, then

$$\min tr(P(k|k)) \Leftrightarrow \max F(\mathbf{a}^l(k))\frac{C(k)P^2(k|k-1)C(k)^T}{C(k)P(k|k-1)C(k)^T + R}.$$  

where “ $\Leftrightarrow$ ” denotes “be equivalent to”. Let $\tilde{y}(k) = \tilde{y}(k)/\sigma_q(k)$. For $y(k)$ is a standard Gaussian variable with constant covariance 1, the quantizer for $\tilde{y}(k)$ can be designed as $\mathbf{a}^l = [\tilde{\tau}_1, \tilde{\tau}_2, \ldots, \tilde{\tau}_l]$, then the thresholds $\mathbf{a}^l$ at time instant $k$ can be obtained by $a_i(k) = \tilde{\tau}_i\sigma_q(k)$, $l = 1, 2, \ldots, L$. Thus the value of $F(\mathbf{a}^l(k))$ formulated as (12) is not affected by the covariance of innovation $\tilde{y}(k)$. As a result,

$$\min tr(P(k|k)) \Leftrightarrow \max F(\mathbf{a}^l)\frac{C(k)P^2(k|k-1)C(k)^T}{C(k)P(k|k-1)C(k)^T + R}$$  

where

$$F(\mathbf{a}^l) = 2 \sum_{i=0}^{I} \frac{(\phi(a_i) - \phi(a_{i+1}))^2}{T(a_{i+1}) - T(a_i)}.$$
is related with the quantizer only and irrelevant to $C(k)$. While,
\[
\frac{C(k)P^T(k|k-1)C(k)}{C(k)P(k|k-1)C(k)^T + R}
\]
is related with $C(k)$ only and irrelevant to the quantizer $\tilde{d}^k$. Thus the quantizer and the dimension reduction matrix can be designed separately at each time instant. Consequently, problem (16) is decomposed into the following two sub-problems:

Sub-problem I: Design of the optimal quantizer, i.e.
\[
\max_{\tilde{d}^k} \frac{1}{2} \sum_{l=0}^{L} \left( \phi(\tilde{a}_l) - \phi(\tilde{a}_{l+1}) \right)^2
\]
s.t. $0 < \tilde{a}_1 < \cdots < \tilde{a}_L < +\infty. \tag{22}
\]

Sub-problem II: Design of the optimal dimension reduction matrix, i.e.
\[
\max_{C(k)} \frac{C(k)P^2(k|k-1)C(k)^T}{C(k)P(k|k-1)C(k)^T + R}
\]
s.t. $C(k)P(k|k-1)C(k)^T + R \leq W. \tag{23}
\]

3.2. Solutions

In the inequality constrained optimization problem of Sub-problem I, the objective function $F(\tilde{d}^k)$ is bounded and infinitely differentiable, while the constraints define a convex set of the standard Gaussian variables $\tilde{a}_l$, $l = 1, 2, \ldots, L$. Due to these properties and the special form of the constraints matrices, the gradient and Hessian conditions of the optimality [47] can be checked analytically for simple cases, e.g. $L = 1$. Numerical methods are practical consideration for solving the constraints of high levels. From which we prefer the sequential quadratic programming (SQP) method [48] based on expanding quadratically the objective function and checking the Karush–Kuhn–Tucker conditions. By checking the feasible and bounded conditions [48], it is seen that the SQP is applicable to solve the Sub-problem I, and the local optimal solution is expected to be obtained by evoking the following Matlab command:

\[
[x, y] = \text{fmincon}(fun, x_0, A, b, Aeq, beq, lb, ub).
\]

The “fmincon” attempts to find a constrained minimum of a scalar function of several variables starting at an initial estimate, and it is a gradient-based approach to problems where the objective and constraint functions are both continuous and have continuous first derivatives, so its use in this sub-problem is reasonable. This command starts at an initial value $x_0$ and attempts to find a minimizer $x$ of the function described by $fun$ subject to the linear inequalities $A \cdot x \leq b$ and linear equalities $Aeq \cdot x = beq$. The solution $x$ is set in $lb \leq x \leq ub$, and the minimal value of $fun$ is returned by $y$. In our problem, only the inequalities $A \cdot x \leq b$ are required and formulated by using constraints on $\tilde{a}^k$ in (22); $fun = -F(\tilde{d}^k)$. The “fmincon” uses one of three algorithms: active-set, interior-point, or trust-region-reflective, and the optimization result is local minimum possible. In our procedure, the active-set algorithm which is implemented by SQP method will be selected and illustrated by Monte Carlo (MC) simulations. The solution $\tilde{d}^k$ is then used to get $\tilde{d}^k = \tilde{a} - \sigma_s(k) = \tilde{a} \cdot \sqrt{W}$. This means $\tilde{a}^k = \tilde{d}^k$ is constant and can be computed independently of the solution to Sub-problem II.

The Lagrange multiplier method is employed to solve Sub-problem II. The objective function is then written as
\[
f(C(k)) = \frac{C(k)P^2(k|k-1)C(k)^T}{C(k)P(k|k-1)C(k)^T + R} + \rho(C(k)P(k|k-1)C(k)^T + R - W)
\]
where $\rho$ is the Lagrange multiplier. The optimal $C(k)$ can be obtained by solving
\[
df(C(k))/dC(k) = 0.
\]

Let $\lambda_1(k) \geq \lambda_2(k) \geq \cdots \geq \lambda_d(k)$ be the eigenvalues of $P(k|k-1)$ and $e_1(k), e_2(k), \ldots, e_d(k)$ be the corresponding unit orthogonal eigenvectors, then the optimal reduction matrix at time $k$ is
\[
C(k) = \sqrt{(W-R)/\lambda_1(k)} e_1^T(k).
\]
Condition (18) is satisfied by using (26).

Let the difference value between the traces of estimation error covariance matrices of the adjacent iteration times be $d(i) = tr(P(i|i-1)) - tr(P(i|i))$, the expected precision of $d(i)$ be delta. The sub-problem II is then solved by the following algorithm:

Step 0: Set quantization level $L$, power constraint $W$, and compute the optimal quantizer (the solution to sub-problem I);

Step 1 (initiation): Let iteration times $i=0$, set prediction error covariance $P(0|0) = delta$, and $d(0) = \infty$, compute $C(0)$;

Step 2: Let $i = i + 1$, compute $P(i|i-1)$, then $C(i)$ by (26) and $d(i)$ by using (19);

Step 3: If $d(i) \leq delta$, then stop; otherwise, go to step 2.

When the algorithm stops, the obtained matrix $C(i) = C$ is the constant dimension reduction matrix we are looking for. The parameter $\delta$ represents the convergence degree of the algorithm. The smaller the $\delta$ is, the closer to the optimal point the solution is. The precision is at the expense of computation time. For time invariant systems, we can choose to get the dimension reduction matrix offline. In this case, we pay more attention to precision than computation, and hence tend to set $\delta$ as a very small value.

4. Estimability analysis

In order to evaluate the validity of state estimation in systems involving the DPCM-like structure as Fig. 1, here we discuss the concept of estimability proposed for the time-varying linear system (27) which takes system (1), (2) as a special case:

\[
\begin{align*}
\dot{x}(k+1) &= A(k)x(k) + B(k)w(k) \\
y(k) &= C(k)x(k) + w(k)
\end{align*}
\]
where \( \mathbf{x}(k) \in \mathbb{R}^n \), \( \mathbf{w}(k) \in \mathbb{R}^m \), \( \mathbf{y}(k), \mathbf{v}(k) \in \mathbb{R}^p \); \( \mathbf{A}(k), \mathbf{B}(k), \mathbf{C}(k) \) are time-varying matrices with appropriate dimensions; the initial state possesses zero mean and bounded covariance; \( \mathbf{E}[\mathbf{x}(0)\mathbf{w}^T(k)] = 0 \), \( \mathbf{E}[\mathbf{x}(0)\mathbf{v}^T(k)] = 0 \), \( \mathbf{E}[\mathbf{w}(k)\mathbf{w}^T(j)] = 0 \), \( \mathbf{E}[\mathbf{w}(k)\mathbf{v}^T(j)] = 0 \), \( \mathbf{E}[\mathbf{v}(k)\mathbf{v}^T(j)] = \mathbf{R}(k) \delta_{kj} \), where \( \delta_{kj} = 1 \) when \( k = j \), otherwise \( \delta_{kj} = 0 \).

It was proposed in [20] that the system (27) is estimable if the optimal posterior mean-square error of state estimation is strictly smaller than the prior estimation error. This leads to a necessary and sufficient condition that the following estimability Gramian has full-rank:

\[
W_e[k_0, k] = \sum_{j = k_0}^{k} \mathbf{P}(k, j) \mathbf{H}(j) \mathbf{C}^T(j) \mathbf{H}(j) \mathbf{P}(k, j),
\]

(28)

where \( \mathbf{P}(k, j) = \mathbf{A}(k - 1) \mathbf{A}(k - 2) \cdots \mathbf{A}(j), \quad j < k \), \( \mathbf{H}(k, k) = \mathbf{I} \), and \( \mathbf{H}(k) = \mathbf{E}[\mathbf{x}(k)\mathbf{v}^T(k)] \). Eq. (28) can be interpreted by using the mutual information between state and the data received by the estimator [21–23]. In (27), estimation given outputs is equivalent to that given innovations [35]. However, the relation between estimability based on quantized innovations and that based on quantized outputs [23] needs to be discussed.

### 4.1. Estimability based on quantized innovations

We discuss the estimability of system (27) (with \( p = 1 \)) under the framework of Fig. 1 with a slight difference that the measurement output matrix is not designable. In order to get general result, here we do not focus on a certain quantizer, but on the general time-varying \( N \)-level quantization [13,14] described by (29) which takes (4) as a special case.

\[
\hat{y}_q(k) = Q_d(\hat{y}(k)) = \mathbf{z}_k(k) \quad \text{for} \quad \hat{y}(k) \in \kappa(k), \quad l = 1, 2, \ldots, N
\]

(29)

where \( \mathbf{z}_k(k) = \mathbf{y}(k) - \hat{y}(k-1) \) are innovations, \( \hat{y}(k-1) = \mathbf{E}[\mathbf{y}(k)\hat{y}_q(k-1)] \) is the prediction of \( \mathbf{y}(k) \) in (27) given quantized innovations \( \hat{y}_q(k) = \{y_q(k_0)\hat{y}_q(k_0 + 1) \cdots \hat{y}_q(k-1)\}^T \); \( \hat{y}_q(k) \in \{z_1(k), z_2(k), \ldots, z_N(k)\} \) are quantizer outputs; \( \kappa(k) = (a_0(k), a_1(k), 1) \) are quantization intervals, \( \{a_i(k)\}_{i=1}^{N+1}, -\infty = a_0(k) < a_1(k) < \cdots < a_{N+1}(k) = +\infty \) are thresholds. Also for the consideration of generality, we assume that the quantized innovations \( \hat{y}_q(k) \) are decoded by

\[
\hat{y}_d(k) = D_d(\hat{y}_q(k))
\]

(30)

at the channel receiver before used to estimate, where \( D_d(\cdot) \) is a one-to-one decoding mapping. A common decoding model in accord with (30) is [13]

\[
D_d(\hat{y}_q(k)) = \mathbf{E}[\mathbf{y}(k)\hat{y}_q(k) = \mathbf{z}_k(k)], \quad l = 1, 2, \ldots, N.
\]

In our system stated in Section 2, the one-to-one mapping is implied in the MLQKF [3];

\[
D_d(\hat{y}_q(k)) = f(a^*, \hat{y}_q(k)) \sqrt{\mathbf{CP}(k(k-1))} + \mathbf{R}.
\]

The one-to-one property of decoding implies that state estimation given quantized innovations is equivalent to that given the decoded signals.

The quantized estimability is then discussed under the following assumption.

**Assumption 1.** In system (27) the initial state \( \mathbf{x}(k_0) \), noises \( \mathbf{w}(k) \) and \( \mathbf{v}(k) \) are mutually independent Gaussian variables; The outputs are scalar, i.e. \( \mathbf{y}(k), \mathbf{v}(k) \in \mathbb{R} \); The outputs are preprocessed to generate quantized innovations by (29) before transmission, \( \mathbf{y}_q(k) \) are decoded by (30) at the lossless channel receiver and then used to estimate the states.

It was shown in [23] that the quantized estimability means the mutual information between state and quantized outputs is larger than zero. While, it can be seen [24–26] that when the optimal posterior errors are achieved, \( H(\mathbf{g}^T \hat{x}(k)) = H(\mathbf{g}^T \mathbf{x}(k)|\mathbf{y}_d(k)) = H(\mathbf{g}^T \mathbf{x}(k)) - I(\mathbf{g}^T \mathbf{x}(k), \hat{y}_d(k)) \), \( \forall \mathbf{g} \in \mathbb{R}^n, \mathbf{g} \neq 0 \), where \( \hat{x}^*(k) = \mathbf{x}(k) - \hat{x}^*(k), \quad \hat{x}^*(k) = \mathbf{E}[\mathbf{x}(k)|\hat{y}_d(k)] \), \( \hat{y}_d(k) = [\mathbf{y}_d(k_0)|\mathbf{y}_d(k_0 + 1) | \cdots | \mathbf{y}_d(k)]^T \). Then referring to [22–26, 29], we prefer the following condition for the analysis of estimation given quantized innovations: The system (27), (29) and (30) is estimable if \( \forall k \geq k_0 + n - 1 \), \( \forall \mathbf{g} \in \mathbb{R}^n, \mathbf{g} \neq 0 \),

\[
I(\mathbf{g}^T \mathbf{x}(k), \hat{y}_d(k)) > 0,
\]

(31)

or equivalently,

\[
H(\mathbf{g}^T \mathbf{x}(k)) > H(\mathbf{g}^T \mathbf{x}(k)|\hat{y}_d(k))
\]

(32)

As information theoretic quantities [29], entropy describes the information amount of a variable, while mutual information measures the information amount commonly contained in and the statistic dependence between two variables. So, (31) and (32) imply that the quantized system is estimable if the received data contain the information of any element of the state, which can be extracted by proper estimator. It is known that \( I(\cdot ; \cdot) \geq 0 \) with the equality holds if and only if these two variables are mutual independent. So (31) also means that any direction of state space is not orthogonal to all the past decoded data, i.e. \( \forall \mathbf{g} \in \mathbb{R}^n, \mathbf{g} \neq 0 \),

\[
\mathbf{g}^T \mathbf{E}[\mathbf{x}(k)\hat{y}_q(j)] \neq 0, \quad \exists j \leq k, \quad \forall k \geq k_0 + n - 1
\]

(33)

From (33) the following result is gotten.

**Theorem 1.** Under Assumption 1, the quantized Gaussian system (27), (29), (30) is estimable if and only if

\[
\text{rank} \{W_q[k_0, k]\} = n, \quad \forall k \geq k_0 + n - 1,
\]

(34)

where the quantized estimability Gramian is

\[
W_q[k_0, k] = \sum_{j = k_0}^{k} \psi^2(j) \mathbf{F}(k, j) \mathbf{H}(j) \mathbf{C}(j) \mathbf{H}(j) \mathbf{F}(k, j)
\]

(35)

with

\[
\psi(j) = \sum_{l = 1}^{N} \frac{D_d(\mathbf{z}_l(j))}{\sqrt{\mathbf{C}(j) \mathbf{H}(j) \mathbf{C}(j) + \mathbf{R}(j)}} \left( \phi \left( \frac{a_0(j) + \mathbf{C}(j) \hat{x}(j)}{\sqrt{\mathbf{C}(j) \mathbf{H}(j) \mathbf{C}(j) + \mathbf{R}(j)}} \right) - \phi \left( a_{N+1}(j) + \mathbf{C}(j) \hat{x}(j) \right) \right),
\]

(36)

\( \phi(\cdot) \) is the probability density function of a standard Gaussian variable as in Section 2.3.
**Proof.** From Eqs. (27), (29) and (30) we have
\[
E(x(k)y_0(j)) = E(\Phi(k,j)x) + \sum_{\ell=0}^{k-j-1} \Phi(k,j+\ell+1)B(\ell)w(j+\ell)i_0y_0(j))
= \Phi(k,j)E(x(j)y_0(j)). \tag{37}
\]
Let \(x = x(j), \hat{x} = \hat{x}(j-1), y_0(j) = \hat{y}(j), \hat{y} = \hat{y}(j-1), R = R(j)\) for notation simplicity, then
\[
E(x(j)y_0(j)) = \sum_{l=1}^{N} \int_{x \in R^n} x \cdot D_j(z_l) \cdot p(x, y_0 = D_j(z_l))dx
= \sum_{l=1}^{N} D_j(z_l) \int_{x \in R^n} x \cdot p(x, y_0 = D_j(z_l))dx. \tag{38}
\]
Based on the Bayesian law
\[
p(x, y_0 = D_j(z_l)) = p(y_0 = D_j(z_l)|x)p(x)
= p(a_l < \hat{y} \leq a_{l+1}|x)p(x)
= p(a_l + Cx + v - Cx \leq a_{l+1}|x)p(x)
= p(a_l + Cx - Cx < \hat{y} \leq a_{l+1} + Cx - Cx)p(x)
= p(a_l + Cx - Cx < \hat{y} \leq a_{l+1} + Cx - Cx)p(x)
\]
where the last equality holds because \(x(j)\) and \(v(j)\) are mutually independent. Then
\[
\int_{x \in R^n} x \cdot p(x, y_0 = D_j(z_l))dx = \int_{x \in R^n} x \cdot p(a_l + Cx - Cx < \hat{y} \leq a_{l+1} + Cx - Cx)p(x) dx
\]
\[
= \int_{x \in R^n} (T^{-1}q_{l+1} - \sqrt{C} - Cx) - (a_{l+1} + Cx - Cx) \Gamma(0, II) dx \tag{39}
\]
where \(T(\cdot)\) is the probability distribution function value of a standard Gaussian variable with zero mean and covariance \(1\) as in Section 2.3, \(\Gamma(0, II)\) denotes the probability density function which means that the stochastic vector \(x\) is Gaussian with zero-mean and covariance matrix \(II\) (\(II\) here denotes \(II(j)\) for short). Compute the integral in (39) by variable replacements, from (38) we have
\[
E(x(j)y_0(j)) = \sum_{l=1}^{N} \frac{D_j(z_l)}{C_j(j)C_j^{-1}j(j) + R(j)} \left( \begin{array}{c} a_l + Cj\hat{x}(j-1) \\ \sqrt{C_j(j)C_j^{-1}j(j) + R(j)} \end{array} \right)
- \frac{1}{C_j(j)C_j^{-1}j(j) + R(j)} C_j(j)C_j^{-1}j(j)
= \left( \begin{array}{c} a_l + Cj\hat{x}(j-1) \\ \sqrt{C_j(j)C_j^{-1}j(j) + R(j)} \end{array} \right)
- \frac{1}{C_j(j)C_j^{-1}j(j) + R(j)} C_j(j)C_j^{-1}j(j)
\]
Then (37) shows that (33) is equivalent to
\[
g^T \sum_{l=1}^{N} \frac{D_j(z_l)}{C_j(j)C_j^{-1}j(j) + R(j)} \left( \begin{array}{c} a_l(j) + Cj\hat{x}(j-1) \\ \sqrt{C_j(j)C_j^{-1}j(j) + R(j)} \end{array} \right)
- \frac{1}{C_j(j)C_j^{-1}j(j) + R(j)} C_j(j)C_j^{-1}j(j)
\]
\[
\exists j \leq k, \forall k \geq k_0 + n - 1, \forall g \in R^n, g \neq 0.
\]
Thus (34) is obtained. \(\Box\)

**Remark 1.** Comparing (35) with (28) shows that the quantized estimability is defined by the quantizer and the intrinsic property of the original system jointly. When \(\psi(j) \neq 0, \forall j \geq k_0\), the estimability of the original system can be preserved under quantization, even if the quantizer is 1-bit \([1, 44, 45]\). To see this, suppose a time-varying quantizer is designed as
\[
D_j(z_l(j)) = c_l \sqrt{C_j(j)\Pi(j)(j) + R(j)}, \quad l = 1, ..., N
\]
where \(c_l\) and \(d_l\) are chosen to make \(\psi(j)\) nonzero constant, i.e., \(\psi(j) = \psi = \Sigma_{l=1}^{N} c_l(\phi(d_l) - \phi(d_{l+1})) \neq 0, j = k_0, k_0 + 1, \ldots\).
Then comparing (35) with (28) we know such a quantizer does not change the estimability of the original system. Especially, when \(N = 2\) in the above equations, let \(c_1 = -1, c_2 = 1, d_1 = -\infty, d_2 = 0, d_3 = +\infty\), then \(\psi(j) = \sqrt{2}/\pi\).
Namely a 1-bit coarse quantizer can preserve the estimability.

**Remark 2.** If the estimation \(\hat{x}(k)\) given \(\hat{Y}_k^g\) is optimal, \(H(g^T\hat{x}(k)|\hat{Y}_k^g) = H(g^T\hat{x}(k)|\hat{Y}_k^g)\), then the quantized system is estimable if the posterior error entropy is strictly less than the prior entropy:
\[
H(g^T\hat{x}(k)) < H(g^T\hat{x}(k)), \quad \forall k \geq k_0 + n - 1, g \in R^n, g \neq 0. \tag{40}
\]
This condition is an information theoretic counterpart of that in [20], i.e. the optimal posterior mean-square estimation error is strictly smaller than the prior error. If \(\hat{x}(k)\) is not optimal, \(H(g^T\hat{x}(k)) > H(g^T\hat{x}(k)|\hat{Y}_k^g)\), then (40) is a sufficient condition for practical consideration.

**Remark 3.** There are three supplementary explanations to the applicability of Theorem 1 to the system in Sections 2 and 3 where the design method is proposed: (1) Because of the one-to-one property of the decoder (30), \(l(g^T\hat{x}(k); \hat{Y}_k^g) = l(g^T\hat{x}(k); \hat{Y}_k^g)\), so (31) is equivalent to \(l(g^T\hat{x}(k); \hat{Y}_k^g) > 0\) which can be applied to the analysis for the case that the state is estimated by using \(\hat{Y}_k^g\) (as in Sections 2 and 3) without decoding (30); (2) The designed procedure in Section 3 provides innovations under a power level (15), while Theorem 1 is derived for any innovation power; (3) The reduction matrix (2) of the system in Sections 2 and 3 is designable, while the output matrix in (27) of the framework in this section is not, this does not affect the estimability analysis. In fact the Gramian (35) involves two key items playing roles in quantized estimability, the C concerning the dimension reduction, and the \(\psi\) concerning the quantization. The estimability yields constraints on the design of C and the quantizer. If no quantization is involved, the constraints will be reduced to constraint on the dimension of C, because when C is designed under the index of (16) or (23), the information loss is minimal in the procedure of dimension reduction.

**Remark 4.** Gramian (35) based on quantized innovations is slightly different from that based on quantized outputs [23]:
\[
W_k^g[k, k] = \sum_{j = k_0}^{k} \psi_j^2(j)\Phi(k,j)\Pi(j)\Pi(j)\Phi(k,j) \tag{41}
\]
with
\[ \psi(j) = \sum_{i=1}^{N} \psi_{i}(j) = \frac{D_{i}(z_{i}(j))}{\sqrt{C(j)l_{i}l_{i}'(j)C^{t}(j)+R_{i}(j)}} \left( \phi \left( \frac{a_{i}(j)}{\sqrt{C(j)l_{i}l_{i}'(j)C^{t}(j)+R_{i}(j)}} \right) - \phi \left( \frac{a_{i+1}(j)}{\sqrt{C(j)l_{i}l_{i}'(j)C^{t}(j)+R_{i}(j)}} \right) \right). \]

where \( D_{i}, z_{i}(j), a_{i}(j) \) are respectively the decoding, quantizer outputs and levels corresponding to system outputs. However, if the quantizer (29) and decoder (30) are applied for both cases, \( a_{i}(j) \) can be replaced by \( a_{i}(j)+C(j)\hat{x}(j)\Delta(j), \) so can \( D_{i}, z_{i}(j). \) It can be seen from the following equivalent form of \( \psi(j) \) that the Gramian (35) is equivalent to (41). So under the same quantization and decoding, the estimability based on quantized innovations is equivalent to that based on quantized outputs.

4.2. Convergence

Let \( \Delta = \sup_{1 \leq l \leq N, k \geq k_{0}} \Delta_{l}(k), \) where \( \Delta_{l}(k) = a_{l+1}(k) - a_{l}(k). \) Consider the Gramian in Theorem 1 when \( \Delta \to 0. \)

**Theorem 2.** Under Assumption 1, the estimability Gramian (35) of quantized system (27), (29) and (30) converges to the estimability Gramian (28) of unquantized system (27) when \( \Delta \to 0. \)

**Proof.** We have
\[ \lim_{\Delta \to 0} \psi(j) = \lim_{\Delta \to 0} \sum_{i=1}^{N} \psi_{i}(j) = \sum_{i=1}^{N} \frac{D_{i}(z_{i}(j))}{\sqrt{C(j)l_{i}l_{i}'(j)C^{t}(j)+R_{i}(j)}} \left( \phi \left( \frac{a_{i}(j)+C(j)\hat{x}(j)\Delta(j)}{\sqrt{C(j)l_{i}l_{i}'(j)C^{t}(j)+R_{i}(j)}} \right) - \phi \left( \frac{a_{i+1}(j)+C(j)\hat{x}(j)\Delta(j)}{\sqrt{C(j)l_{i}l_{i}'(j)C^{t}(j)+R_{i}(j)}} \right) \right). \]

Then noting \( N \to \infty, \Delta_{l} \to 0 \) when \( \Delta \to 0, \) from (42) we have
\[ \lim_{\Delta \to 0} \psi(j) = \lim_{\Delta \to 0} \sum_{i=1}^{N} \frac{D_{i}(z_{i}(j))}{\sqrt{C(j)l_{i}l_{i}'(j)C^{t}(j)+R_{i}(j)}} \left( \phi \left( \frac{a_{i}(j)+C(j)\hat{x}(j)\Delta(j)}{\sqrt{C(j)l_{i}l_{i}'(j)C^{t}(j)+R_{i}(j)}} \right) - \phi \left( \frac{a_{i+1}(j)+C(j)\hat{x}(j)\Delta(j)}{\sqrt{C(j)l_{i}l_{i}'(j)C^{t}(j)+R_{i}(j)}} \right) \right). \]

**5. Simulation**

Simulations of different models show that the estimation performance of the proposed method is satisfactory. Analytic results on estimability given quantized innovations are also verified by simulations. In this section, only the simulation results of 3 models are stated. The design method of dimension reduction and quantizer is illustrated mainly by Model 1 and Model 2. The analysis of estimability is illustrated further by Model 3. These models are simulated with different power constraints \( W \) and different quantization levels \( L \) for the symmetrical quantizer (4) or \( N \) for the general quantizer (29).

5.1. Model 1 and Model 2

Model 1 described by (1) is from [37]. The parameters are:
\[
A = \begin{bmatrix} -1.1 & 0 & 0 & 0 \\ 0.01 & 0 & 1 & 0 \\ 0.275 & 0 & 0 & 1 \\ 0.06 & 0 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad Q = 1, \quad R = 1.
\]

Model 2 from [38] is as
\[
x(k+1) = Ax(k) + Bu(k) + \Gamma w(k),
\]
where \( u(k) \) is the control input which does not affect the state estimation. We then let \( u(k) = 0 \) and \( \Gamma = B \) in accord with (1). The parameters of Model 2 are:
\[
A = \begin{bmatrix} 0.3718 & -0.3111 & 0.1357 & -0.0557 \\ 0.9575 & 0.1524 & -0.0685 & 0.1043 \\ 0.3408 & -0.3287 & 0.3923 & -0.2060 \\ 0.0529 & -0.2922 & 0.9320 & 0.4835 \end{bmatrix},
\]
These two models are mainly used to illustrate the algorithms in Section 3.2 which takes the symmetrical model (4) as the quantizer and the MLQ-KF as the predictor and filter.

MC simulation method is employed to illustrate the design method which involves numerical technique, the SQP. The active-set algorithm is selected here when using the Matlab function “fmincon”. Set delta = 0.00001 in the algorithm. By using “fmincon”, the optimization depends on the initial setting of the normalized quantizer thresholds \( \overline{a}_l \), \( l = 1, 2, \ldots, L \) in (22). We set the initial values of these standard Gaussian variables randomly under the constraints in (22). Simulation results illustrate our design method. For the case of \( L = 3, 200 \) optimization results of quantizer thresholds are shown in Fig. 2(a), where \( q^s_l \) denote the optimization results of \( \overline{a}_l \), the star points are defined by values of triple \( \{\pi_1, \pi_2, \pi_3\} \). We evaluate the corresponding estimation performance by using the trace of the steady estimation error covariance \( P \). For Model 1, 200 simulation results of \( \text{tr}(P) \) in the case of \( W = 3 \) and \( L = 3 \), corresponding to the quantizer thresholds in Fig. 2(a), are shown in Fig. 2(b). Results of one simulation realization for different power constraints and quantization levels are also shown respectively in Tables 1 and 2, and Figs. 3 and 4. Along with the \( \text{tr}(P) \), time average square errors defined as follows are also given:

\[
(\Delta x_i)^2 = \frac{1}{T} \sum_{k=1}^{T} (x_i(k) - \hat{x}_i(k))^2
\]

where \( x_i(k) \) and \( \hat{x}_i(k) \) are respectively the \( i \)-th elements of the actual states and estimates, \( T \) denotes simulation time. Data in Tables 1 and 2 are \( \Delta x^2_i \) and \( \text{tr}(P) \) in the cases of different \( W \) and \( L \) (\( T = 100 \)). It is seen that the square error decreases when signal power or quantization level increases. This is consistent with the intuition. The data show that the performance of \( L = 3 \) is close to that of \( L = \infty \) (i.e. the signals are not quantized) when the transmitted power remains invariant. This illustrates the advantage of the method. Performances of cases of \( W = 3, L = \infty \) and \( W = 6, L = 1 \) are very close. This suggests that the performance decrease caused by lower quantization level can be compensated by higher power. Only the 1st state elements of these models in the case of \( W = 3 \) and \( L = 1 \) are shown by figures for saving text space.

Table 1

<table>
<thead>
<tr>
<th>( W = 3, L = 1 )</th>
<th>( \Delta x^2_1 )</th>
<th>( \Delta x^2_2 )</th>
<th>( \Delta x^2_3 )</th>
<th>( \text{tr}(P) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8487</td>
<td>0.0255</td>
<td>0.0324</td>
<td>0.0017</td>
<td>1.0566</td>
</tr>
<tr>
<td>0.5876</td>
<td>0.0235</td>
<td>0.0263</td>
<td>0.0015</td>
<td>0.6976</td>
</tr>
<tr>
<td>0.4147</td>
<td>0.0242</td>
<td>0.0255</td>
<td>0.0013</td>
<td>0.6099</td>
</tr>
<tr>
<td>0.4631</td>
<td>0.0260</td>
<td>0.0257</td>
<td>0.0014</td>
<td>0.5869</td>
</tr>
<tr>
<td>0.2212</td>
<td>0.0139</td>
<td>0.0142</td>
<td>0.0007</td>
<td>0.3040</td>
</tr>
<tr>
<td>0.1734</td>
<td>0.0112</td>
<td>0.0113</td>
<td>0.0005</td>
<td>0.2368</td>
</tr>
</tbody>
</table>

Table 2

<table>
<thead>
<tr>
<th>( W = 3, L = 1 )</th>
<th>( \Delta x^2_1 )</th>
<th>( \Delta x^2_2 )</th>
<th>( \Delta x^2_3 )</th>
<th>( \text{tr}(P) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5055</td>
<td>2.9696</td>
<td>3.7679</td>
<td>6.4590</td>
<td>17.8801</td>
</tr>
<tr>
<td>1.4906</td>
<td>2.5241</td>
<td>3.1872</td>
<td>5.2531</td>
<td>13.9071</td>
</tr>
<tr>
<td>1.0252</td>
<td>2.3152</td>
<td>2.1876</td>
<td>3.0971</td>
<td>12.4909</td>
</tr>
<tr>
<td>1.1475</td>
<td>2.4792</td>
<td>2.5080</td>
<td>4.1069</td>
<td>12.4848</td>
</tr>
<tr>
<td>0.9749</td>
<td>2.3826</td>
<td>2.4776</td>
<td>3.3500</td>
<td>8.2446</td>
</tr>
<tr>
<td>0.9009</td>
<td>1.8176</td>
<td>1.8104</td>
<td>2.4057</td>
<td>7.0180</td>
</tr>
</tbody>
</table>
In quantized systems, the probability distribution of $g^T \hat{x}(k)$ is undefined but can be supposed to make its entropy maximal according to the well known Jaynes' "maximum entropy principle" [39,26]. The maximum entropy distribution is Gaussian when the covariance is given, the posterior error entropy $H(g^T \hat{x}(k))$ in simulation can also be calculated by using that as the above formula. When $g^T = [1 0 0 0]$, condition (40) means $H(x_1) > H(\hat{x}_1)$ which is shown in Fig. 3(c), where the "entropy difference of $x_1$" is $H(x_1) - H(\hat{x}_1)$. This implies that the estimability of the original system (Model 1) is preserved under the designed framework proposed in Sections 2 and 3, even if a coarse quantizer ($L=1$) is used. For Model 2, we have also verified estimability results for different cases. For example, Fig. 4(c) shows that $H(x_1) > H(\hat{x}_1)$ in the case of $W=3$ and $L=1$.

### 5.2. Model 3

Model 3 from [4] is simulated to verify the analysis of quantized estimability. Its parameters (as in (1) and (2)) are:

\[
A = \begin{bmatrix}
2.4744 & -2.811 & 1.7038 & -0.5444 & 0.0723 \\
1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 & 0
\end{bmatrix},
\]

\[
B = \begin{bmatrix}
1 \\
0 \\
0 \\
0 \\
0
\end{bmatrix},
\]

\[
C = \begin{bmatrix}
0.245 & 0.236 & 0.384 & 0.146 & 0.035
\end{bmatrix},
\]

\[
Q = 1, \quad R = \frac{1}{16}
\]

Different from that in the designed algorithm, here we choose the filter of Gaussian fit algorithm [13] and the Max–Lloyd quantizer [8] to illustrate the general conclusion.

Cases of quantization level $N=3$ and 2 in (28) are shown. The quantizer outputs are respectively $\sigma_{y_1}(k) = (-1.2240, 0, 1.2240)$ and $\sigma_{y_1}(k) = (-0.7979, 0.7979)$ for $N=3$ and 2, where $\sigma_{y_1}(k)$ is the standard deviation of $\hat{y}(k)$.

By [20] we know Model 3 is estimable without quantization. Computation shows $\psi(j) \neq 0, j = 0, 1, 2, \ldots$, for both cases of $N=3$ and 2 (Figs. 5(a) and 6(a)). So, the estimability is not changed by quantization according to Remark 1. Figs. 5(b) and 6(b) show respectively the actual values of $x_1(k)$ by solid lines and its estimates $\hat{x}_1(k)$ by dotted lines. Figs. 5(c) and 6(c) are differences between the prior and posterior entropies. The "entropy difference of $x_1$" $= H(x_1(k)) - H(\hat{x}_1(k))$, where $H(x_1(k))$ of estimation errors $\hat{x}_1(k)$ are also computed under the "maximum entropy principle" as above. We observe that all of the entropy differences are greater than zero, i.e. the posterior error entropies are strictly smaller than prior entropies. This indicates the computation conclusion that the quantized systems with the 3-level quantizer and the 1-bit quantizer are both

![Fig. 3.](image-url)
estimable, according to (40), and also illustrated the Remark 4 concerning the estimability equivalence between quantizing innovation and quantizing outputs.

6. Conclusion

The present paper discussed the problem of joint design of state dimension reduction and quantizer for
state estimation in quantized linear dynamic systems with a DPCM-like structure. For systems under constraints of limited signal power, number and bandwidth of the communication (sub) channels, the dimension reduction matrix and quantizer thresholds were obtained based on MMSE criterion and the MLQ-KF. To illustrate the validity of state estimation in the designed system, the state estimability given quantized innovations was analyzed based on the quantity of mutual information, along with the design. The quantized estimability Gramian was obtained, and its convergence property was also analyzed. It was shown that the estimability of the original system can be preserved even if a 1-bit quantizer is involved. This provides an auxiliary support for the results on 1-bit quantized estimation [1,44,45]. The estimability equivalence between cases of quantized innovations and quantized outputs [23] was also discussed. Monte Carlo simulation results show that the performance of state estimation under communication constraints is satisfactory when the derived dimension reduction and quantizers are applied. The analytic conclusion on estimability was also verified by illustrative simulations.

In the present paper, the design method was given and verified for the case of one subchannel. For the case of more than one parallel subchannels, it is suspected by the authors that the performance of estimation can be improved partly by considering the signal power allocation.

On the other hand, as stated in Section 1, the scheme of the pre-processing part is similar to the GQS [42] and its special form of the NSQ [43]. A very interesting result is that the NSQ improves the accuracy of the estimates. Also, while the estimability equivalence between quantizing innovations and quantizing output observations was gotten in this paper, an important fact was exploited in [4,40] that quantizing innovations requires fewer bits than quantizing observations directly. These contributions provide valuable hints for further consideration of estimation performance of predictive quantizer or the DPCM-like scheme.

Acknowledgment

The authors thank the anonymous reviewers whose comments and suggestions are very valuable and helpful for improving the quality of our work.

References

