Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Quantized innovations Kalman filter: stability and modification with scaling quantization*

Jian XU[1,2], Jian-xun LI[1,2], Sheng XU[3]

([1]*Science and Technology on Avionics Integration Laboratory, Shanghai Jiao Tong University, Shanghai 200240, China*)

([2]*MOE Key Laboratory of System Control and Information Processing, Shanghai Jiao Tong University, Shanghai 200240, China*)

([3]*School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China*)

E-mail: xujian2001-1@163.com; lijx@sjtu.edu.cn; xusheng2007-1@163.com

Received June 11, 2011; Revision accepted Oct. 8, 2011; Crosschecked Jan. 6, 2012

Abstract: The stability of quantized innovations Kalman filtering (QIKF) is analyzed. In the analysis, the correlation between quantization errors and measurement noises is considered. By taking the quantization errors as a random perturbation in the observation system, the QIKF for the original system is equivalent to a Kalman-like filtering for the equivalent state-observation system. Thus, the estimate error covariance matrix of QIKF can be more exactly analyzed. The boundedness of the estimate error covariance matrix of QIKF is obtained under some weak conditions. The design of the number of quantized levels is discussed to guarantee the stability of QIKF. To overcome the instability and divergence of QIKF when the number of quantization levels is small, we propose a Kalman filter using scaling quantized innovations. Numerical simulations show the validity of the theorems and algorithms.

Key words: Kalman filtering, Quantized innovation, Stability, Scaling quantization, Wireless sensor network
doi:10.1631/jzus.C1100161 **Document code:** A **CLC number:** TP13

1 Introduction

Because of the finite computation precision in micro-processors (Karlsson and Gustafsson, 2005a; 2005b), estimation with quantized data has long been a well studied topic in digital signal processing (DSP) (Curry *et al.*, 1970; Clements and Haddad, 1972).

Recently, wireless sensor networks (WSNs) have attracted much attention due to their applications in environmental monitoring, intelligent transportation, space exploration, military surveillance, etc. (Karlsson and Gustafsson, 2005a; 2005b; Xiao *et al.*, 2006; Sun *et al.*, 2007; Duan *et al.*, 2008; Msechu *et al.*, 2008; Yu *et al.*, 2009). Because of the bandwidth constraint, each sensor is only able to transmit a finite number of bits. Observations have to be quantized before transmission. Thus, a revisit of the estimation with quantized data in WSNs is warranted.

Much of the early work (Curry, 1970; Curry *et al.*, 1970; Clements and Haddad, 1972) devised approximate point estimators of the Kalman-type filter based on the optimal conditional mean estimator, which in general requires numerical integration for implementation. Sviestins and Wigren (2000) derived an exact density filter based on solving the Fokker-Planck equation and Bayes' rule for a special case of the problem, under somewhat

* Project supported by the National Natural Science Foundation of China (Nos. 61175008, 60935001, and 60874104), the National Basic Research Program (973) of China (Nos. 2009CB824900 and 2010CB734103), the Space Foundation of Supporting-Technology (No. 2011-HT-SHJD002), and the Aeronautical Science Foundation of China (No. 20105557007)

restrictive assumptions. Recently, Karlsson and Gustafsson (2005a; 2005b) and Sukhavasi and Hassibi (2009) applied particle filtering to solve the estimation problems, which practically amounts to repeatedly using Bayes' rule.

Currently, quantized Kalman filtering is applied to solve estimation problems in WSNs (Ribeiro, 2005; Ribeiro *et al.*, 2006; You *et al.*, 2008; 2011; Xu and Li, 2011). The sensor measurements are quantized into several bits to save energy. It is obvious that quantizing sensor measurements can lead to large quantized noise when the observed value is large, which then leads to poor estimation accuracy. In Ribeiro (2005) and Ribeiro *et al.* (2006), this limitation was overcome by developing an elegant distributed estimation approach based on quantizing the innovation to one bit (the so-called sign of innovation or SOI) (Fig. 1). There the quantization Kalman filter is obtained under the Gaussian assumption of the predicted distribution. In You *et al.* (2008; 2011) and Msechu *et al.* (2008), quantized innovations Kalman filtering (QIKF) was generalized to handle multiple quantization levels. All of these discussions are based on the Gaussian assumption of the predicted distribution and the independence assumption between quantization errors and measurement noises. This approximation is quite prevalent in low-complexity algorithms for nonlinear filtering. However, despite its superior practical usefulness, the stability of QIKF has not been analyzed in a rigorous mathematical way for a long time.

In the context of analyzing the stability of the closed loop system under state feedback control, it is important to investigate the stability of the observer. Now the question that needs to be answered is whether QIKF can still maintain stability in the face of quantization errors. On the other hand, does the true estimation error covariance matrix of QIKF remain following the modified Riccati recursion? If the answer is negative, then under what condition is the QIKF stable?

The main objective of the current study is to more exactly analyze the stability of QIKF from the view of control theory. The other objective of this work is to give a modified filtering algorithm to improve the stability of QIKF when the number of quantization levels is small.

Notations: The superscript 'T' stands for matrix transpose, superscript '−1' stands for matrix inverse, and '−T' stands for the transposition of matrix inverse. E(·) stands for expectation. We use $\|\cdot\|$ to denote the matrix norm. Given an $n \times m$ matrix $\boldsymbol{A}$, the norm of $\boldsymbol{A}$ is defined as $\|\boldsymbol{A}\| = (\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}))^{1/2}$, where tr(·) is the trace of a square matrix. Given a symmetric matrix $\boldsymbol{A}$, $\boldsymbol{A} > \boldsymbol{0}$ means that $\boldsymbol{A}$ is a positive definite matrix, and $\boldsymbol{A} \geq \boldsymbol{0}$ means that $\boldsymbol{A}$ is a positive semi-definite matrix. Given two positive semi-definite matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{A} > \boldsymbol{B}$ means that $\boldsymbol{A} - \boldsymbol{B} > \boldsymbol{0}$, and $\boldsymbol{A} \geq \boldsymbol{B}$ means that $\boldsymbol{A} - \boldsymbol{B} \geq \boldsymbol{0}$.

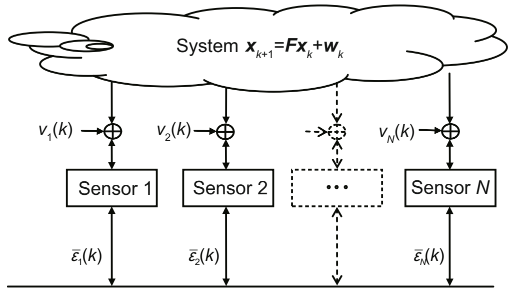

Fig. 1 State estimation with quantized innovations

2 Modeling assumptions and preliminaries

In this study, to highlight the impact of quantification on the filtering performance, we focus on the state estimation problem of a linear time-invariant stochastic system.

Consider the discrete linear time-invariant stochastic system

$$\boldsymbol{x}_{k+1} = \boldsymbol{F}\boldsymbol{x}_k + \boldsymbol{w}_k, \tag{1}$$

$$y_k = \boldsymbol{H}\boldsymbol{x}_k + v_k, \tag{2}$$

where $\boldsymbol{x}_k \in \mathbb{R}^n$ is a state vector to be estimated at time $t_k = k\Delta t$, Δt is the time step of the sample, $\boldsymbol{F}$ is a time-invariant matrix with a suitable dimensionality. $y_k \in \mathbb{R}$ is the scalar observation of the sensor, and $\boldsymbol{H}$ is the measurement coefficient vector. $\boldsymbol{w}_k \in \mathbb{R}^n$ and $v_k \in \mathbb{R}$ are uncorrelated Gaussian noises with zero mean and positive covariance matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$. The initial value $\boldsymbol{x}_0$ with mean μ_0 and variance $\boldsymbol{P}_0$ is independent of $\boldsymbol{w}_k$ and v_k.

For the system (1)–(2), the discrete minimum-variance linear estimate of $\boldsymbol{x}_k$, i.e., the standard Kalman filter, is determined by Algorithm 1.

It is known that $\widehat{\boldsymbol{x}}^*_{k|k}$ is the minimum-variance estimate, $\boldsymbol{P}^*_{k|k}$ is the estimation error covariance matrix after processing the measurement y_k, and $\boldsymbol{P}^*_{k|k-1}$ is the extrapolated error covariance matrix.

Algorithm 1 Standard Kalman filter

1: Initialization. Give $\hat{\boldsymbol{x}}_{0|0}$ ($\mathrm{E}(\hat{\boldsymbol{x}}_{0|0})=\mu_0$) and $\boldsymbol{P}_{0|0}$.

2: Time update:

$$\widehat{\boldsymbol{x}}^*_{k|k-1} = \boldsymbol{F}\widehat{\boldsymbol{x}}^*_{k-1|k-1}, \quad (3)$$

$$\boldsymbol{P}^*_{k|k-1} = \boldsymbol{F}\boldsymbol{P}^*_{k-1|k-1}\boldsymbol{F}^{\mathrm{T}} + \boldsymbol{Q}, \quad (4)$$

$$\boldsymbol{K}^*_k = \boldsymbol{P}^*_{k|k-1}\boldsymbol{H}^{\mathrm{T}}(\boldsymbol{H}\boldsymbol{P}^*_{k|k-1}\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{R})^{-1}. \quad (5)$$

3: Measurement update:

$$\widehat{\boldsymbol{x}}^*_{k|k} = \widehat{\boldsymbol{x}}^*_{k|k-1} + \boldsymbol{K}^*_k(y_k - \boldsymbol{H}\widehat{\boldsymbol{x}}^*_{k|k-1}), \quad (6)$$

$$\boldsymbol{P}^*_{k|k} = \boldsymbol{P}^*_{k|k-1} - \boldsymbol{K}^*_k\boldsymbol{H}\boldsymbol{P}^*_{k|k-1}, \quad (7)$$

where $\widehat{\boldsymbol{x}}^*_{k|k-1}$ and $\widehat{\boldsymbol{x}}^*_{k|k}$ are called the priori and posteriori estimates, respectively.

3 Kalman filter using quantized innovations

In this section, we briefly review the state estimation based on quantized innovations.

3.1 Quantized innovations

Recently, there has been a lot of resurgent interest in the research on quantized estimation, which has applications in networked systems such as WSNs (You *et al.*, 2009). Because of the bandwidth constraint, each sensor is only able to transmit a finite number of bits. Observations have to be quantized before transmission.

The activated sensor makes an observation, and computes the innovation $\boldsymbol{\varepsilon}(k) = \boldsymbol{Y}(k) - \widehat{\boldsymbol{Y}}_{k|k-1}$. The one-step predictor $\widehat{\boldsymbol{Y}}_{k|k-1}$ of the observation and the inverse $(\boldsymbol{S}_k^{1/2})^{-1}$ are received by the sensor from the estimator center, where $\boldsymbol{S}_k$ is the innovation covariance (cf. Eq. (14)). Denote the normalized innovation as

$$\overline{\overline{\boldsymbol{\varepsilon}}}(k) = (\boldsymbol{S}_k^{1/2})^{-1}\boldsymbol{\varepsilon}(k). \quad (8)$$

Then each component $\overline{\overline{\varepsilon}}_i(k)$ $(i=1,2,\cdots,d)$ of the normalized innovation $\overline{\overline{\boldsymbol{\varepsilon}}}(k)$ is quantized to produce a quantized innovation $\overline{\boldsymbol{\varepsilon}}(k) = q_L(\overline{\overline{\boldsymbol{\varepsilon}}}(k))$. We consider a symmetric $L=(2l+1)$ levels quantizer $\overline{\varepsilon}_i(k) = q_L(\overline{\overline{\varepsilon}}_i(k))$, where L is the number of quantization levels for every component of the innovation vector. More specifically, the symmetric quantizer $\overline{\varepsilon}_i(k) = q_L(\overline{\overline{\varepsilon}}_i(k))$ for $\overline{\overline{\varepsilon}}_i(k)$ is given by

$$q_L(\overline{\overline{\varepsilon}}_i(k))=\begin{cases} a_l, & \overline{\overline{\varepsilon}}_i(k) \in (b_l,+\infty), \\ \cdots & \cdots \\ a_2, & \overline{\overline{\varepsilon}}_i(k) \in (b_2,b_3], \\ a_1, & \overline{\overline{\varepsilon}}_i(k) \in (b_1,b_2], \\ 0, & \overline{\overline{\varepsilon}}_i(k) \in (-b_1,b_1], \\ -a_1, & \overline{\overline{\varepsilon}}_i(k) \in (-b_2,-b_1], \\ -a_2, & \overline{\overline{\varepsilon}}_i(k) \in (-b_3,-b_2], \\ \cdots & \cdots \\ -a_l, & \overline{\overline{\varepsilon}}_i(k) \in (-\infty,-b_l], \end{cases} \quad (9)$$

where $b_1, b_2, \cdots, b_l$ are the thresholds of the minimum distortion quantization (Max, 1960). Thus, the normalized innovation $\overline{\overline{\boldsymbol{\varepsilon}}}(k)$ is quantized to produce a quantized innovation $\overline{\boldsymbol{\varepsilon}}(k) = [\overline{\varepsilon}_1(k), \overline{\varepsilon}_2(k), \cdots, \overline{\varepsilon}_d(k)]^{\mathrm{T}}$. For the same number of quantization levels, different transmission strategies will lead to different numbers of bits (Xu and Li, 2011). In this study, we focus only on the number of quantization levels, L.

3.2 Kalman filtering using quantized innovations

Research on the state estimation using quantization innovations has attracted considerable attention (Ribeiro *et al.*, 2006; You *et al.*, 2008; 2009; 2011; Xu and Li, 2011). According to You *et al.* (2009), the QIKF for a linear system can be described as Algorithm 2. We denote the quanti-

Algorithm 2 Kalman filtering with quantization innovations

1: Initialization. Give $\hat{\boldsymbol{x}}_{0|0}$ ($\mathrm{E}(\hat{\boldsymbol{x}}_{0|0})=\mu_0$) and $P_{0|0}$.

2: Time update:

$$\widehat{\boldsymbol{x}}_{k|k-1} = \boldsymbol{F}\widehat{\boldsymbol{x}}_{k-1|k-1}, \quad (10)$$

$$\boldsymbol{P}_{k|k-1} = \boldsymbol{F}\boldsymbol{P}_{k-1|k-1}\boldsymbol{F}^{\mathrm{T}} + \boldsymbol{Q}\,, \quad (11)$$

$$\boldsymbol{K}_k = \boldsymbol{P}_{k|k-1}\boldsymbol{H}^{\mathrm{T}}(\boldsymbol{H}\boldsymbol{P}_{k|k-1}\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{R})^{-1}. \quad (12)$$

3: Quantization:

$$\boldsymbol{\varepsilon}_k = \boldsymbol{y}_k - \boldsymbol{H}\widehat{\boldsymbol{x}}_{k|k-1}, \quad (13)$$

$$\boldsymbol{S}_k = \boldsymbol{H}\boldsymbol{P}_{k|k-1}\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{R}, \quad (14)$$

$$\overline{\boldsymbol{\varepsilon}}_k = q_L((\boldsymbol{S}_k^{1/2})^{-1}\boldsymbol{\varepsilon}_k). \quad (15)$$

4: Measurement update:

$$\widehat{\boldsymbol{x}}_{k|k} = \widehat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{K}_k\boldsymbol{S}_k^{1/2}\overline{\boldsymbol{\varepsilon}}_k, \quad (16)$$

$$\boldsymbol{P}_{k|k} = \boldsymbol{P}_{k|k-1} - \boldsymbol{K}_k\boldsymbol{S}_k\boldsymbol{K}_k^{\mathrm{T}} + \boldsymbol{K}_k\boldsymbol{C}_k\boldsymbol{K}_k^{\mathrm{T}}. \quad (17)$$

zation error as

$$\boldsymbol{u}_k \equiv \boldsymbol{S}_k^{1/2}\overline{\boldsymbol{\varepsilon}}_k - \boldsymbol{\varepsilon}_k. \tag{18}$$

Denote the covariance matrix of quantization errors as

$$\boldsymbol{C}_k \equiv \mathrm{E}[\boldsymbol{u}_k\boldsymbol{u}_k^{\mathrm{T}}]. \tag{19}$$

The calculation of $\boldsymbol{C}_k$ will be discussed in Section 3.3.

From Eq. (19), we have

$$\begin{aligned}\boldsymbol{S}_k^{1/2}\overline{\boldsymbol{\varepsilon}}_k &\equiv \boldsymbol{\varepsilon}_k + \boldsymbol{u}_k \\ &= \boldsymbol{H}\boldsymbol{x}_k + \boldsymbol{u}_k + \boldsymbol{v}_k - \boldsymbol{H}\widehat{\boldsymbol{x}}_{k|k-1}.\end{aligned} \tag{20}$$

Thus, introducing the quantization error at time k is equivalent to introducing a perturbation $\boldsymbol{u}_k$ into observation system (2), which makes the measurement innovation equal to $\boldsymbol{S}_k^{1/2}\overline{\boldsymbol{\varepsilon}}_k$. Hence, the original state-observer system (1)–(2) becomes the equivalent state-observer system

$$\overline{\boldsymbol{x}}_{k+1} = \boldsymbol{F}\overline{\boldsymbol{x}}_k + \boldsymbol{w}_k, \tag{21}$$

$$\overline{\boldsymbol{y}}_k = \boldsymbol{H}\overline{\boldsymbol{x}}_k + \boldsymbol{u}_k + \boldsymbol{v}_k. \tag{22}$$

Remark 1 Since $\boldsymbol{u}_k$ is non-Gaussian, and is correlated to $\boldsymbol{\varepsilon}_k$, $\widehat{\boldsymbol{x}}_{k|k}$ is no longer the minimum variance estimation of state $\boldsymbol{x}_k$. Accordingly, $\boldsymbol{P}_{k|k}$ is no longer the actual estimation error covariance matrix. Thus, in addition to system noise $\boldsymbol{w}_k$ and observation noise $\boldsymbol{v}_k$, the state estimation error also comes from two parts: (1) the introduction of 'undetermined' random perturbation $\boldsymbol{u}_k$ in the observation system; (2) non-Gaussian $\boldsymbol{u}_k$ and the correlation between $\boldsymbol{u}_k$ and $\boldsymbol{\varepsilon}_k$. In QIKF (Algorithm 2), the state estimation error from the first part is considered, and the second part is ignored, especially the correlation between $\boldsymbol{u}_k$ and $\boldsymbol{\varepsilon}_k$. Hence, the covariance matrix system of the estimate errors in Algorithm 2 is no longer the case. In fact, when the number of quantization levels L is small, the actual covariance matrix of estimation errors should be slightly larger than that obtained from Eq. (17) in Algorithm 2. Accordingly, $\boldsymbol{S}_k$ is no longer the actual covariance matrix of innovation $\boldsymbol{\varepsilon}_k$. This easily leads to the divergence of QIKF, especially when the system (1) is very unstable, i.e., the eigenvalue of the system state transition matrix is large.

Remark 2 It is known that, if there had been no quantization before time k, then $\widetilde{\boldsymbol{x}}_{k-1|k-1}$ would be uncorrelated with $\boldsymbol{\varepsilon}_k$, and thus independent of $\boldsymbol{\varepsilon}_k$, because both would be Gaussian distributed (Fu and Xie, 2009). Hence, if $\boldsymbol{\varepsilon}_k$ were quantized, its quantization error $\boldsymbol{u}_k$ would be uncorrelated with $\widetilde{\boldsymbol{x}}_{k|k-1} = \boldsymbol{F}\widetilde{\boldsymbol{x}}_{k-1|k-1}$. Because quantization occurred before time k, $\widetilde{\boldsymbol{x}}_{k|k-1}$ and $\boldsymbol{u}_k$ are correlated in general. However, the correlation is typically weak. In particular, the effect of past quantization errors should be negligible and it is thus fair to ignore the correlation between $\widetilde{\boldsymbol{x}}_{k|k-1}$ and $\boldsymbol{u}_k$. Only the correlations between $\boldsymbol{\varepsilon}_k$, $\boldsymbol{v}_k$, and $\boldsymbol{u}_k$ are considered in this study. For the same reason, the impact of the non-Gaussian of $\boldsymbol{u}_{k-1}$ on the Gaussian of $\boldsymbol{\varepsilon}_k$ is also slight. In fact, the following conditions hold very well in numerical simulations:

Assumption 1 Asymptotically, the measurement innovation $\boldsymbol{\varepsilon}_k$ is approximately Gaussian distributed with zero mean and variance $\boldsymbol{S}_k$.

Assumption 2 The quantization error $\boldsymbol{u}_{k-1}$ is uncorrelated with $\boldsymbol{H}\widetilde{\boldsymbol{x}}_{k|k-1}$.

Note that, besides the Gaussianity assumption, Ribeiro *et al.* and You *et al.* also assumed that the quantizing error $\boldsymbol{u}_k$ is uncorrelated with all other random variables. Because of this, they can easily obtain the QIKF. For the same reason, the stability of QIKF needs to be improved. The main difference between our Assumptions 1 and 2 and the assumption in Ribeiro (2005), Ribeiro *et al.* (2006), and You *et al.* (2008), is that we consider the correlation between quantization error $\boldsymbol{u}_k$ and measurement noise $\boldsymbol{v}_k$. Thus, the effect of the quantization noise can be fully considered in the algorithm. The stability of the resulting filter can thus be improved.

3.3 Some important properties

For later application, we provide some properties of the quantization error $\boldsymbol{u}_k$ and the estimation algorithm.

Theorem 1 For the estimation error dynamics

$$\widetilde{\boldsymbol{x}}_{k|k-1} \equiv \boldsymbol{x}_k - \widehat{\boldsymbol{x}}_{k|k-1} = \boldsymbol{F}\widetilde{\boldsymbol{x}}_{k-1|k-1} + \boldsymbol{w}_{k-1}, \tag{23}$$

$$\widetilde{\boldsymbol{x}}_{k|k} \equiv \boldsymbol{x}_k - \widehat{\boldsymbol{x}}_{k|k} = \boldsymbol{H}\widetilde{\boldsymbol{x}}_{k|k-1} + \boldsymbol{v}_k + \boldsymbol{u}_k, \tag{24}$$

the estimation error $\widetilde{\boldsymbol{x}}_{k|k}$, the prediction error $\widetilde{\boldsymbol{x}}_{k|k-1}$, and the quantization error $\boldsymbol{u}_k$ have zero-mean and an even probability density function for all $k > 0$.

Proof As in Fu and de Souza (2009), the statement can be easily shown by induction. Since

$\mathrm{E}(\boldsymbol{x}_{0|0}) = \mu_0$, $\widetilde{\boldsymbol{x}}_{0|0}$ is zero-mean with an even probability density function. Note that L-level quantizer $q_L(\cdot)$ is an odd function. Suppose $\widetilde{\boldsymbol{x}}_{k-1|k-1}$ is zero-mean with an even probability density function for some $k-1$. Then it follows from Eqs. (23) and (24) that $\widetilde{\boldsymbol{x}}_{k|k}$ is also zero-mean with an even probability density function. Hence, by induction, $\widetilde{\boldsymbol{x}}_{k|k}$ is zero-mean with an even probability density for all $k>0$. In addition, in view of Eq. (23), it follows that $\widetilde{\boldsymbol{x}}_{k|k-1}$ has zero-mean and an even probability density function for all $k>0$. In view of Eqs. (13)–(15) and (19), we have

$$\begin{aligned}\boldsymbol{\varepsilon}_k &= \boldsymbol{y}_k - \boldsymbol{H}\widehat{\boldsymbol{x}}_{k|k-1}\\ &= \boldsymbol{H}\widetilde{\boldsymbol{x}}_{k|k-1} + \boldsymbol{v}_k,\\ \overline{\boldsymbol{\varepsilon}}_k &= q_L((\boldsymbol{S}_k^{1/2})^{-1}\boldsymbol{\varepsilon}_k),\\ u_k &= \boldsymbol{S}_k^{1/2}\overline{\boldsymbol{\varepsilon}}_k - (\boldsymbol{H}_k\widetilde{\boldsymbol{x}}_{k|k-1} + \boldsymbol{v}_k)\\ &= \boldsymbol{S}_k^{1/2}q_L((\boldsymbol{S}_k^{1/2})^{-1}\boldsymbol{\varepsilon}_k) - (\boldsymbol{H}_k\widetilde{\boldsymbol{x}}_{k|k-1} + \boldsymbol{v}_k).\end{aligned}$$

It follows that $\boldsymbol{u}_k$ has zero-mean and an even probability density function for all $k>0$.

Lemma 1 Let $\boldsymbol{\varepsilon}$ be a random variable with zero mean and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$. Consider the L-level minimum distortion quantizer (9). There is a real number $\alpha_L \in (0,1)$, named the 'distortion rate', such that the covariance matrix $\boldsymbol{\Sigma}_{q_L}$ of the quantization error satisfies

$$\begin{aligned}\boldsymbol{\Sigma}_{q_L} &\equiv \mathrm{E}\{(\boldsymbol{\varepsilon} - q_L(\boldsymbol{\varepsilon}))(\boldsymbol{\varepsilon} - q_L(\boldsymbol{\varepsilon}))^{\mathrm{T}}\}\\ &\leq \alpha_L\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} \to \boldsymbol{0}, \quad \text{as } L\to\infty.\end{aligned} \tag{25}$$

Equivalently, $\alpha_L \to 0$, as $L\to\infty$, where L is the number of quantization levels.

Proof It follows from Gray and Neuhoff (1998) that there is a real number $0<\alpha_L<1$, such that

$$\boldsymbol{\Sigma}_{q_L} = \alpha_L\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}} \to \boldsymbol{0}, \quad \text{as } L\to\infty.$$

Remark 3 In this paper, the quantizer we consider is the minimum distortion quantization in Max (1960). The distortion rates of 2–17 levels minimum distortion quantization (Max, 1960) are given in Table 1. For more details, one can refer to Max (1960) and the references therein.

For later application, we indicate some relationships here, concerning the gain matrix $\boldsymbol{K}_k$, the forecast error variance matrix $\boldsymbol{P}_{k|k-1}$, and the estimate error variance matrix $\boldsymbol{P}_{k|k}$. In Eqs. (12) and (17), $\boldsymbol{K}_k$ and $\boldsymbol{P}_{k|k}$ can also be expressed as

$$\begin{aligned}\boldsymbol{K}_k &= (\boldsymbol{P}_{k|k-1}^{-1} + \boldsymbol{H}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{H})^{-1}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{R}^{-1}\\ &= (\boldsymbol{P}_{k|k} - \boldsymbol{K}_k\boldsymbol{C}_k\boldsymbol{K}_k^{\mathrm{T}})\boldsymbol{H}^{\mathrm{T}}\boldsymbol{R}^{-1},\end{aligned} \tag{26}$$

$$\begin{aligned}\boldsymbol{P}_{k|k} &= (\boldsymbol{I} - \boldsymbol{K}_k\boldsymbol{H})\boldsymbol{P}_{k|k-1}(\boldsymbol{I} - \boldsymbol{K}_k\boldsymbol{H})^{\mathrm{T}}\\ &\quad + \boldsymbol{K}_k\boldsymbol{R}\boldsymbol{K}_k^{\mathrm{T}} + \boldsymbol{K}_k\boldsymbol{C}_k\boldsymbol{K}_k^{\mathrm{T}}\\ &= (\boldsymbol{I} - \boldsymbol{K}_k\boldsymbol{H})\boldsymbol{P}_{k|k-1} + \boldsymbol{K}_k\boldsymbol{C}_k\boldsymbol{K}_k^{\mathrm{T}}\\ &= (\boldsymbol{P}_{k|k-1}^{-1} + \boldsymbol{H}\boldsymbol{R}^{-1}\boldsymbol{H}^{\mathrm{T}})^{-1} + \boldsymbol{K}_k\boldsymbol{C}_k\boldsymbol{K}_k^{\mathrm{T}}.\end{aligned} \tag{27}$$

The derivation is straightforward and is hence omitted.

Table 1 The distortion rate of the minimum distortion quantization in Max (1960)

Number of quantization levels, L	Distortion rate, α_L	Number of quantization levels, L	Distortion rate, α_L
2	0.3634	10	0.02293
3	0.1902	11	0.01922
4	0.1175	12	0.01634
5	0.07994	13	0.01406
6	0.05798	14	0.01223
7	0.04400	15	0.01073
8	0.03454	16	0.009497
9	0.02785	17	0.008463

4 Performance evaluation of QIKF

Just as in the analysis in Remark 1, for an unstable system (1), quantization easily leads to the divergence of QIKF. Hence, in this section, we discuss the stability of QIKF for system (1)–(2) under Assumptions 1 and 2.

4.1 Preliminaries

To provide complete proofs of our main theorems, we need to introduce some conclusions about the boundedness of $\boldsymbol{P}_{k|k-1}$, $\boldsymbol{K}_k$, and $\boldsymbol{P}_{k|k}$ in Algorithm 2.

For $\boldsymbol{P}_{k|k-1}$, we consider the modified algebraic Riccati equation (MARE)

$$\begin{aligned}\boldsymbol{P}_k &= \boldsymbol{F}\boldsymbol{P}_{k-1}\boldsymbol{F}^{\mathrm{T}} + \boldsymbol{Q} - \lambda\boldsymbol{F}\boldsymbol{P}_{k-1}\boldsymbol{H}^{\mathrm{T}}\\ &\quad \cdot(\boldsymbol{H}\boldsymbol{P}_{k-1}\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{R})^{-1}\boldsymbol{H}\boldsymbol{P}_{k-1}\boldsymbol{F}^{\mathrm{T}},\end{aligned} \tag{28}$$

where $\boldsymbol{P}_k = \boldsymbol{P}_{k|k-1}$. Research on the stability of the MARE has attracted considerable attention as it is vital to the stability problems related to packet dropout and the quantization of networked control systems (You *et al.*, 2008).

For unstable $\boldsymbol{F}$, Sinopoli *et al.* (2004) proved that there is a critical value

$$\overline{\lambda} = \arg\inf_{\lambda}\{\boldsymbol{P}|\boldsymbol{P} > \boldsymbol{FPF}^{\mathrm{T}} + \boldsymbol{Q} - \lambda\boldsymbol{FPH}^{\mathrm{T}} \cdot(\boldsymbol{HPH}^{\mathrm{T}} + \boldsymbol{R})^{-1}\boldsymbol{HPF}^{\mathrm{T}}\}. \quad (29)$$

When $\lambda > \overline{\lambda}$, the MARE (28) has a unique positive definite solution $\boldsymbol{P}_k$ satisfying

$$\lim_{k\to\infty} \boldsymbol{P}_k \to \boldsymbol{P}, \quad (30)$$

where $\boldsymbol{P}$ is the unique positive definite solution of the algebraic Riccati equation

$$\boldsymbol{P} = \boldsymbol{FPF}^{\mathrm{T}} + \boldsymbol{Q} - \lambda\boldsymbol{FPH}^{\mathrm{T}}(\boldsymbol{HPH}^{\mathrm{T}} +\boldsymbol{R})^{-1}\boldsymbol{HPF}^{\mathrm{T}}. \quad (31)$$

Meanwhile, we know that the $\boldsymbol{P}_k = \boldsymbol{P}_{k|k-1}$ is bounded; i.e., there are two positive real numbers, α_1 and β_1, such that

$$\alpha_1\boldsymbol{I}_n \le \boldsymbol{P}_{k|k-1} \le \beta_1\boldsymbol{I}_n, \quad \text{for all } k \ge 1, \quad (32)$$

where $\boldsymbol{I}_n$ is an identity matrix with n dimensions. It follows from Eqs (11), (12), and (14) that the $\boldsymbol{P}_{k|k}$, $\boldsymbol{K}_k$, and $\boldsymbol{S}_k$ are also bounded when $\lambda > \overline{\lambda}$. Hence, when $\lambda > \overline{\lambda}$, there are positive numbers α, β, μ, and ν, such that

$$\alpha\boldsymbol{I}_n \le \boldsymbol{P}_{k|k} \le \beta\boldsymbol{I}_n, \quad (33)$$
$$\|\boldsymbol{K}_k\boldsymbol{K}_k^{\mathrm{T}}\| \le \mu, \quad (34)$$
$$\boldsymbol{0} \le \boldsymbol{S}_k \le \nu\boldsymbol{I}_d, \quad (35)$$

for all $k \ge 1$, where $\boldsymbol{I}_n$ and $\boldsymbol{I}_d$ are identity matrices with n and d dimensions, respectively.

Remark 4 When rank$\boldsymbol{H} = 1$, it is shown in Sinopoli *et al.* (2004) that the closed form of $\overline{\lambda}$ is

$$\overline{\lambda} = 1 - \frac{1}{\Pi_i|\sigma_i^u|^2}. \quad (36)$$

When $\boldsymbol{H}$ is invertible, $\overline{\lambda}$ depends only on the maximum eigenvalue of $\boldsymbol{F}$, i.e.,

$$\overline{\lambda} = 1 - \frac{1}{\max_i|\sigma_i^u|^2}, \quad (37)$$

where σ_i^u is the unstable eigenvalue of $\boldsymbol{F}$. For more details about $\overline{\lambda}$, one can refer to Sinopoli *et al.* (2004) and the references therein.

The following lemma comes from the results of Sinopoli *et al.* (2004). It demonstrates some useful properties of the $\boldsymbol{P}_{k|k}$.

Lemma 2 Suppose that $(\boldsymbol{F}, \boldsymbol{Q}^{1/2})$ is controllable and $(\boldsymbol{F}, \boldsymbol{H})$ is detectable. Then for an unstable $\boldsymbol{F}$, there exists a $\overline{\lambda} \in (0, 1)$ such that the algebraic equation

$$\boldsymbol{P} = \boldsymbol{FPF}^{\mathrm{T}} + \boldsymbol{Q} - \lambda\boldsymbol{K}[\boldsymbol{H}(\boldsymbol{FPF}^{\mathrm{T}} +\boldsymbol{Q})\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{R}]\boldsymbol{K}^{\mathrm{T}} \quad (38)$$

has a positive definite solution $\boldsymbol{P}$ and the equation

$$\boldsymbol{P}_{k|k} = \boldsymbol{FP}_{k-1|k-1}\boldsymbol{F}^{\mathrm{T}} - \lambda\boldsymbol{K}_k[\boldsymbol{H}(\boldsymbol{FP}_{k-1|k-1}\boldsymbol{F}^{\mathrm{T}} +\boldsymbol{Q})\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{R}]\boldsymbol{K}_k^{\mathrm{T}} + \boldsymbol{Q} \quad (39)$$

admits a unique positive definite solution $\boldsymbol{P}_{k|k}$ satisfying $\boldsymbol{P}_{k|k} \to \boldsymbol{P}$ as $k \to \infty$ for any nonnegative $\boldsymbol{P}_{0|0} \ge \boldsymbol{0}$ if and only if $\lambda > \overline{\lambda}$, where $\boldsymbol{K} = \boldsymbol{PH}^{\mathrm{T}}\boldsymbol{R}^{-1}$ and $\boldsymbol{K}_k = \boldsymbol{P}_{k-1|k-1}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{R}^{-1}$.

Proof It can be obtained directly from Theorems 3–5 in Sinopoli *et al.* (2004).

4.2 Stability of QIKF

The following theorem provides the upper and lower bounds for the estimation error covariance matrix $\widetilde{\boldsymbol{P}}_{k|k}$ of QIKF.

Theorem 2 For the scalar observation system (1)–(2), suppose that $(\boldsymbol{F}, \boldsymbol{Q}^{1/2})$ is controllable and $(\boldsymbol{F}, \boldsymbol{H})$ is detectable. If

$$\alpha_L + 2\sqrt{\alpha_L} < 1 - \overline{\lambda}, \quad (40)$$

then for the estimation error covariance matrix $\widetilde{\boldsymbol{P}}_{k|k}$ of QIKF, we have

$$\boldsymbol{P}^*_{k|k} \le \widetilde{\boldsymbol{P}}_{k|k} \le \overline{\boldsymbol{P}}_{k|k} \quad (41)$$

for any nonnegative $\boldsymbol{P}_0 \ge \boldsymbol{0}$, where $\boldsymbol{P}^*_{k|k}$ is the unique positive definite solution of Eq. (7) and $\overline{\boldsymbol{P}}_{k|k}$ is the unique positive definite solution of equation

$$\overline{\boldsymbol{P}}_{k|k} = \boldsymbol{F}\overline{\boldsymbol{P}}_{k-1|k-1}\boldsymbol{F}^{\mathrm{T}} + \boldsymbol{Q} - (1-\alpha_L-2\sqrt{\alpha_L})\boldsymbol{K}_k \cdot[\boldsymbol{H}(\boldsymbol{F}\overline{\boldsymbol{P}}_{k-1|k-1}\boldsymbol{F}^{\mathrm{T}} + \boldsymbol{Q})\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{R}]\boldsymbol{K}_k^{\mathrm{T}}. \quad (42)$$

Proof The proof is shown in the Appendix.

Remark 5 Let $\overline{\lambda}$ denote the critical value in Eq. (29). It follows from Theorem 2 that, besides controllability and observability, to guarantee that the QIKF is stable, we need

$$\alpha_L + 2\sqrt{\alpha_L} < 1 - \overline{\lambda}, \tag{43}$$

which leads to

$$\alpha_L < 3 - \overline{\lambda} - 2\sqrt{2 - \overline{\lambda}}. \tag{44}$$

Thus, the number of quantization levels to guarantee a stable QIKF can be obtained according to Table 1. Note that the condition (40) is only sufficient, but not necessary. This means that the minimum number of quantization levels obtained from condition (40) is not the smallest one. In the process of theorem derivation, the Cauchy-Schwartz inequality is used to estimate the upper bound of the correlation matrix, and hence the corresponding condition (40) obtained is conservative.

The next theorem gives an estimate of the limit of the covariance matrix $\widetilde{\boldsymbol{P}}_{k|k}$, when it is bounded.

Theorem 3 For the scalar observation system (1)–(2), suppose that $(\boldsymbol{F}, \boldsymbol{Q}^{1/2})$ is controllable and $(\boldsymbol{F}, \boldsymbol{H})$ is detectable. If

$$\alpha_L + 2\sqrt{\alpha_L} < 1 - \overline{\lambda},$$

then for the estimation error covariance matrix $\widetilde{\boldsymbol{P}}_{k|k}$, we have

$$\overline{\boldsymbol{P}^*} \leq \underline{\lim}_{k\to+\infty} \widetilde{\boldsymbol{P}}_{k|k} \leq \overline{\lim}_{k\to+\infty} \widetilde{\boldsymbol{P}}_{k|k} \leq \overline{\boldsymbol{P}} \tag{45}$$

for any $\boldsymbol{P}_0 \geq \boldsymbol{0}$, where $\overline{\boldsymbol{P}^*}$ and $\overline{\boldsymbol{P}}$ are solutions of the respective algebraic equations

$$\overline{\boldsymbol{P}^*} = \boldsymbol{F}\overline{\boldsymbol{P}^*}\boldsymbol{F}^{\mathrm{T}} + \boldsymbol{Q} - \overline{\boldsymbol{P}^*}\boldsymbol{H}^{\mathrm{T}}(\boldsymbol{H}\overline{\boldsymbol{P}^*}\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{R})^{-1}\boldsymbol{H}\overline{\boldsymbol{P}^*} \tag{46}$$

and

$$\overline{\boldsymbol{P}} = \boldsymbol{F}\overline{\boldsymbol{P}}\boldsymbol{F}^{\mathrm{T}} + \boldsymbol{Q} - (1-\overline{\alpha}_L)\overline{\boldsymbol{P}}\boldsymbol{H}^{\mathrm{T}}(\boldsymbol{H}\overline{\boldsymbol{P}}\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{R})^{-1}\boldsymbol{H}\overline{\boldsymbol{P}}. \tag{47}$$

Proof First, it follows from Theorem 1 that

$$\boldsymbol{P}^*_{k|k} \leq \widetilde{\boldsymbol{P}}_{k|k} \leq \overline{\boldsymbol{P}}_{k|k},$$

where $\boldsymbol{P}^*_{k|k}$ and $\overline{\boldsymbol{P}}_{k|k}$ are the solutions of Eqs. (7) and (42), respectively.

On the one hand, it follows from the convergence of Kalman filter (Walrand, 1972) that

$$\lim_{k\to+\infty} \boldsymbol{P}^*_{k|k} = \overline{\boldsymbol{P}^*},$$

where $\overline{\boldsymbol{P}^*}$ is the solution of Eq. (46). On the other hand, by Theorem 1 and Lemma 3, Eq. (42) admits a unique positive definite solution $\overline{\boldsymbol{P}}_{k|k}$ satisfying

$$\lim_{k\to+\infty} \overline{\boldsymbol{P}}_{k|k} = \overline{\boldsymbol{P}}$$

for any nonnegative $\boldsymbol{P}_0 \geq \boldsymbol{0}$, where $\overline{\boldsymbol{P}}$ is the solution of Eq. (47). This completes the proof.

5 Kalman filter using scaling quantization innovations

In some cases, communication bandwidth limits are very strict. It is not feasible to improve the stability of QIKF by increasing the number of quantization levels. Hence, in this section, we discuss how to deal with the instability and divergence issues of QIKF when the number of quantization levels is small.

5.1 Modified Kalman filtering using scaling quantized innovations

On the one hand, due to the independent assumption in QIKF, in Eq. (17), the influence of correlations between $\boldsymbol{\varepsilon}_k$, $\boldsymbol{v}_k$, and $\boldsymbol{u}_k$ is not considered. From Remark 2 and the derivation of Theorem 1, it is easy to determine that when the number of quantization levels is small, such as $L \leq 4$, the effect of correlations between $\boldsymbol{\varepsilon}_k$, $\boldsymbol{v}_k$, and $\boldsymbol{u}_k$ cannot be ignored, especially when the system (1) is unstable. The correlations between $\boldsymbol{\varepsilon}_k$, $\boldsymbol{v}_k$, and $\boldsymbol{u}_k$ make the covariance matrix of the estimation error, $\widetilde{\boldsymbol{P}}_{k|k}$, larger than the one in Eq. (17). Correspondingly, the covariance matrix $\widetilde{\boldsymbol{S}}_k$ of innovation is larger than that in Eq. (14). Thus, the covariance matrix of the normalized innovation is larger than the identity matrix. However, the quantizer adopted is still the one in Max (1960), for a unit covariance. Then there will be too many values of normalized innovation located in the intervals $(-\infty, -b_l)$ and (b_l, ∞). The quantization error might also be very large. Once it is beyond a certain range, the QIKF will easily diverge.

Fig. 2 shows the schematic of the minimum distortion quantization with 2–3 levels for a standard normal and also the probability density diagram of a standard normal. It is well known that the probability of the value of a standard normal random variable falling in interval $(-3, 3)$ is not less than 0.997. For 2-level minimum distortion quantization, the value of the quantizer is $a_1 = 0.7980$ (Fig. 2a). This means that the absolute quantizing error $|\varepsilon - q_2(\varepsilon)|$ may be in interval $(1, 2.20)$ with a great probability. Obviously, a large absolute quantizing error will cause a great estimate error or even make the filter instable and divergent. For other multiple level quantifications, similar phenomena exist. In short, quantizer q_L can lead to a large absolute quantizing error, and cause instability and divergence of QIKF, although the distortion rate is minimal.

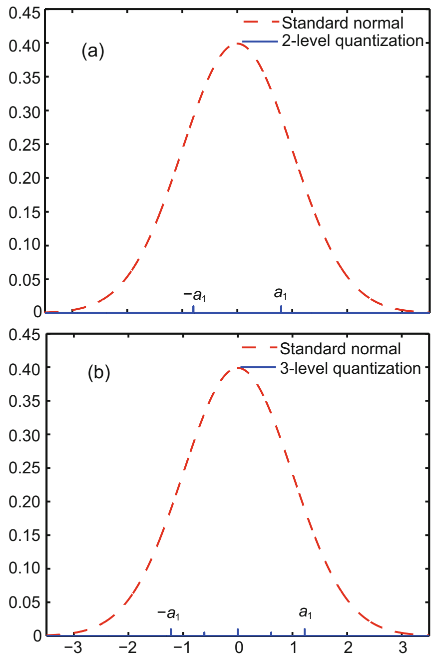

Fig. 2 Schematic of 2-level (a) and 3-level (b) quantization

From above analysis, it can be determined that the instability and divergence of QIKF are due not only to the correlations between $\boldsymbol{\varepsilon}_k$, $\boldsymbol{v}_k$, and $\boldsymbol{u}_k$, but also to the quantitative method. To overcome these limitations, two scaling factors τ_1 and τ_2 are introduced into QIKF. By enlarging the matrix $\boldsymbol{S}_k$, we can make the normalized innovation $\overline{\overline{\varepsilon}}_i(k)$ within the quantization range $(-b_l, b_l)$ as great as possible. Meanwhile, the absolute quantizing error can also be decreased by multiplying τ_2. Thus, a modified QIKF based on scaling quantizing innovations is obtained. The modified Kalman filtering using scaling quantized innovations (SQIKF) is provided by Algorithm 3.

Algorithm 3 Kalman filtering with scaling quantized innovations (SQIKF)

1: Initialization. Give $\hat{\boldsymbol{x}}_{0|0}$ ($\mathrm{E}(\hat{\boldsymbol{x}}_{0|0}) = \mu_0$) and $\boldsymbol{P}_{0|0}$.

2: Time update:

$$\begin{aligned} \widehat{\boldsymbol{x}}_{k|k-1} &= \boldsymbol{F}\widehat{\boldsymbol{x}}_{k-1|k-1}, \quad (48)\\ \boldsymbol{P}_{k|k-1} &= \boldsymbol{F}\boldsymbol{P}_{k-1|k-1}\boldsymbol{F}^{\mathrm{T}} + \boldsymbol{Q}, \quad (49)\\ \boldsymbol{K}_k &= \boldsymbol{P}_{k|k-1}\boldsymbol{H}^{\mathrm{T}}(\boldsymbol{H}\boldsymbol{P}_{k|k-1}\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{R})^{-1}. \quad (50) \end{aligned}$$

3: Quantization:

$$\begin{aligned} \boldsymbol{\varepsilon}_k &= \boldsymbol{y}_k - \boldsymbol{H}\widehat{\boldsymbol{x}}_{k|k-1}, \quad (51)\\ \boldsymbol{S}_k &= \boldsymbol{H}\boldsymbol{P}_{k|k-1}\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{R}, \quad (52)\\ \overline{\boldsymbol{\varepsilon}}_k &= q_L((\tau_1\boldsymbol{S}_k^{1/2})^{-1}\boldsymbol{\varepsilon}_k). \quad (53) \end{aligned}$$

4: Measurement update:

$$\begin{aligned} \widehat{\boldsymbol{x}}_{k|k} &= \widehat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{K}_k\tau_1\boldsymbol{S}_k^{1/2}\tau_2\overline{\boldsymbol{\varepsilon}}_k, \quad (54)\\ \boldsymbol{P}_{k|k} &= \boldsymbol{P}_{k|k-1} - \boldsymbol{K}_k\boldsymbol{S}_k\boldsymbol{K}_k^{\mathrm{T}} + \boldsymbol{K}_k\boldsymbol{C}_k\boldsymbol{K}_k^{\mathrm{T}}. \quad (55) \end{aligned}$$

5.2 Scaling factors τ_1 and τ_2

The key problem with SQIKF is how to design the scaling factors τ_1 and τ_2.

The main technical difficulty in designing τ_1 is that there is no effective calculation approach for the actual covariance matrix $\widetilde{\boldsymbol{S}}_k$ of innovation. Thus, we can provide only an approximative estimate of τ_1 as far as possible. From the derivation of Theorem 1 and the above analysis, we know that $\widetilde{\boldsymbol{S}}_k \geq \boldsymbol{S}_k$. At the same time, the smaller is L, the larger should be $\widetilde{\boldsymbol{S}}_k$ compared with $\boldsymbol{S}_k$. This means that the larger is α_L, the larger should be τ_1. Note that $\alpha_L \in (0, 1)$, and τ_1 should be larger than 1. It is known from the Chebyshev inequality that if τ_1 is too large, the normal innovations $\overline{\overline{\boldsymbol{\varepsilon}}}_k$ will be excessively compressed in interval $(-b_1, b_1)$. Thus, the measurement information in innovation $\boldsymbol{\varepsilon}_k$ cannot be fully utilized. From the above,

$$\tau_1^* = 1 + \alpha_L \quad (56)$$

should be a moderate choice.

On the other hand, the introduction of τ_2 is to reduce the absolute quantizing error $|\varepsilon - q_L(\varepsilon)|$ as far as possible. This implies that

$$\tau_2^*(L) = \arg\min_{\tau} \mathrm{E}[|\boldsymbol{\varepsilon} - \tau \cdot q_L(\boldsymbol{\varepsilon})|]. \tag{57}$$

Obviously, it is very difficult to solve this optimization problem. A suboptimal approach for the design of τ_2 is provided here. Specifically, we will provide the approximate scope of τ_2. Obviously, $\tau_2 > 1$; i.e., the lower bound of τ_2 is 1. We analyze the upper bound of τ_2. Fig. 2 shows that, for the quantizer q_L in Eq. (9), most of the quantitative values $q_L(\boldsymbol{\varepsilon})$ are concentrated in the area of the origin. It is well known that the probability of the values of standard normal random variables falling in interval $(-3, 3)$ is almost equal to 1. Fig. 2 shows that the quantization values a_i $(i = 1, 2, \cdots, l)$ are distributed evenly in the interval $(-a_l, a_l)$. However, the distance between a_l and 3 is significantly greater than the distance between any other two adjacent quantization values. Then if we amplify the quantization value a_i $(i = 1, 2, \cdots, l)$ and make them distributed in interval $(-3, 3)$ as evenly as possible, the absolute quantizing error $|\boldsymbol{\varepsilon} - q_L(\boldsymbol{\varepsilon})|$ can be reduced effectively. To do this, we can take the maximum of $\tau_2(L)$ such that

$$3 - \tau_2(L) \cdot a_l = (3 - (-3))/(2L),$$

i.e.,

$$\tau_2(L) \approx (3L - 3)/(L \cdot a_l). \tag{58}$$

The maxima of $\tau_2(L)$ for 2–5 levels quantization are provided in Table 2, where a_l is taken from Max (1960).

Table 2 The scaling factor $\boldsymbol{\tau_2}$

Number of quantization levels, L	Upper bound of τ_2
2	1.8797
3	1.6340
4	1.4901
5	1.3921

6 Numerical simulations

In this section, we present some cases in which the critical values $\overline{\lambda}$ are known, and give some examples.

Example 1 We consider a first-order system. The discrete time linear time-invariant (LTI) system used in this simulation has $\boldsymbol{F} = a$ $(a > 1)$, $\boldsymbol{H} = 1$, with $\boldsymbol{v}_k$ and $\boldsymbol{w}_k$ having zero mean and variances $\boldsymbol{Q} = 0.09$ and $\boldsymbol{R} = 2.5$, respectively. For all simulations in this example, $\boldsymbol{x}_{0|0} = 3$ and $\boldsymbol{P}_{0|0} = 9$.

When $a = 1.15$, it follows from Eqs. (36) and (44) that $\overline{\lambda} = 0.2439$ and $\alpha_L \le 0.1057$. Then it follows from Table 1 that $L \ge 5$. To show the effect of the stability of system (1) on the stability of QIKF, we consider a slightly larger a. When $a = 1.35$, by using Eq. (37), inequality (44), and Table 1, we have $\overline{\lambda} = 0.4513$, $\alpha_L \le 0.0598$, and $L \ge 6$.

Figs. 3 and 4 show the experimental and theoretical mean square errors (MSEs) of QIKF and standard KF for the system at $a = 1.15$ and $a = 1.35$, respectively. Here, the experimental MSE is obtained from $N = 200$ Monte Carlo simulations:

$$\mathrm{ExpMSE}(k) = \frac{1}{N}\sum_{i=1}^{N} [(\boldsymbol{x}_k - \boldsymbol{x}_{k|k})^{\mathrm{T}}(\boldsymbol{x}_k - \boldsymbol{x}_{k|k})], \tag{59}$$

and the theoretical MSE is obtained by

$$\mathrm{MSE}(k) = \sum_{j=1}^{d} P_{k|k}(j, j). \tag{60}$$

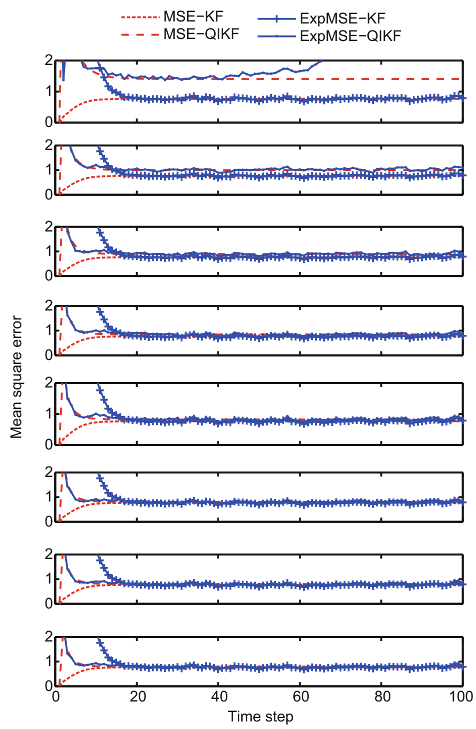

Fig. 3 Mean square errors (MSEs) of 2-, 3-, ..., and 9-level (in increasing order of the number of levels from top to bottom) KF and QIKF in Example 1 ($\boldsymbol{a} = 1.15$)

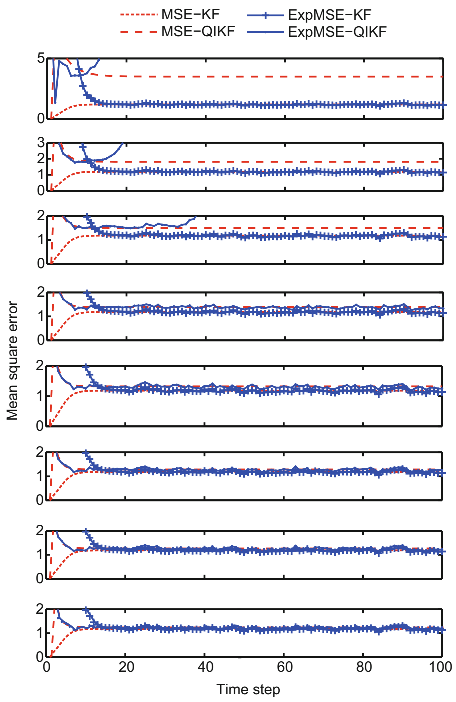

Fig. 4 Mean square errors (MSEs) of 2-, 3-, . . . , and 9-level (in increasing order of the number of levels from top to bottom) KF and QIKF in Example 1 ($a = 1.35$)

Fig. 3 shows that the 2-level QIKF fails to track the state, while QIKF works well when $L \geq 5$. Fig. 4 shows that QIKF diverges when $2 \leq L \leq 4$ and converges when $L \geq 6$. This means that QIKF actually may be diverged when it is applied to an unstable linear system. The more unstable is the system (1), the easier is the QIKF diverged. However, when the number of quantization levels meets the conditions of Theorem 2 and Remark 5 (inequality (44)), QIKF is stable. This shows the correctness of Theorem 1 and Remark 5.

Similar to Fu and de Souza (2009), we proceed to verify Assumptions 1 and 2 for our example. Fig. 5 shows the probability density function of the normalized innovations for 2–5 levels quantization, computed using simulated data and normalized to have a unity variance, along with a standard Gaussian probability density function. We see that the computed probability density function fits a Gaussian probability density well even for 2–5 levels of quantization. Hence, Assumption 1 holds well. Fig. 6 shows the correlation coefficients between quantization error $\boldsymbol{u}_k$ and predictive measurement error $\boldsymbol{H}\tilde{\boldsymbol{x}}_{k|k-1}$. The correlation coefficients are all between -0.04 and 0.04. This means that the quantization error $\boldsymbol{u}_k$ is almost uncorrelated with the predictive measurement error $\boldsymbol{H}\tilde{\boldsymbol{x}}_{k|k-1}$. Hence, Assumption 2 holds well too.

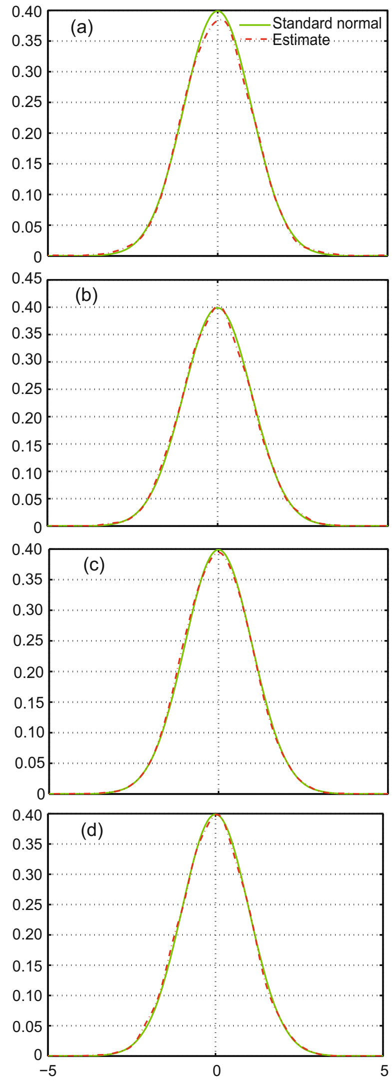

Fig. 5 Probability density estimation of normalized innovations for 2-level (a), 3-level (b), 4-level (c), and 5-level (d) quantization in Example 1 (time step=40)

Example 2 Consider the system (1)–(2) with

$$\boldsymbol{F} = \begin{pmatrix} 1.25 & 1 \\ 0 & 0.98 \end{pmatrix} \tag{61}$$

and

$$\boldsymbol{H} = \begin{pmatrix} 1 & 1 \end{pmatrix}. \tag{62}$$

In the system (1)–(2), $\boldsymbol{w}_k$ and $\boldsymbol{v}_k$ have zero mean and variance $\boldsymbol{Q} = 0.09\boldsymbol{I}_{2\times2}$ and $\boldsymbol{R} = 2.5$, $\boldsymbol{x}_{0|0} = (1, 0.3)$ and $\boldsymbol{P}_{0|0} = 9\boldsymbol{I}_{2\times2}$. This time $\overline{\lambda} = 0.36$, $\alpha_L = 0.0788$, and $L \geq 6$. In this example, Assumptions 1 and 2 still hold well.

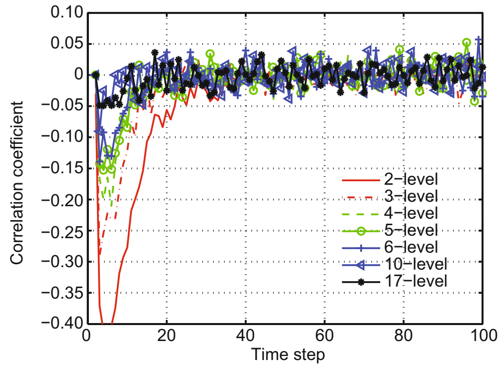

Fig. 6 Correlation coefficient between the quantization error and the predictive measurement in Example 1

Fig. 7 shows the MSEs of 2–9 levels quantized filters. The 2- and 3-level quantized filters fail to track the state over the entire course and the experimental MSEs diverge, although the theoretical MSEs are bounded. When $L > \underline{L} = 6$, the QIKF successfully fulfills the tracking duty, showing the significance of the additional quantization level. This shows the correctness of Theorem 1 and also demonstrates the correctness of the discussion on the number of quantizing levels in Section 4.2 (Remark 5).

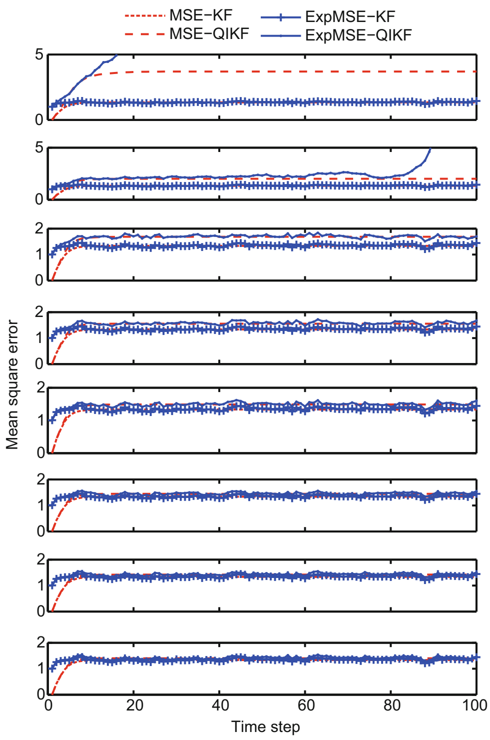

Fig. 7 Mean square errors (MSEs) of 2-, 3-, . . . , and 9-level (in increasing order of the number of levels from top to bottom) KF and QIKF in Example 2

To evaluate the performance of QIKF, the experimental and theoretical MSEs of standard KF are also shown in Fig. 7. With an increasing number of quantization levels, the MSEs of QIKF come close to the MSE of standard KF.

Fig. 8 shows the experimental and theoretical MSEs of 4–17 levels QIKF, when the conditions of Corollary 1 are met. Fig. 8 shows that the experimental MSE of QIKF is always slightly larger than the theoretical MSE. This is consistent with Corollary 1.

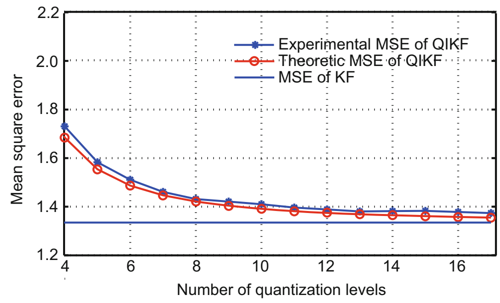

Fig. 8 Limits of the mean square error (MSE) in Example 2

We can also see that, with the increasing number of quantization levels, the experimental and theoretical MSEs of QIKF approach the MSE of standard KF. When $L = 8$, the MSEs of QIKF are very close to the MSE of standard KF. When $L = 17$, they are almost identical.

Example 3 The third example is aimed at demonstrating the stability of SQIKF. All the parameters are the same as those in Example 2. Fig. 9 shows the experimental and theoretical MSEs of 2- and 3-

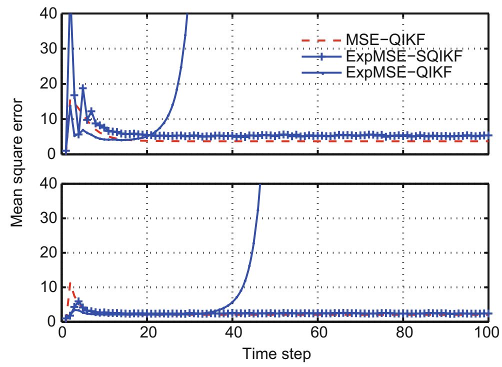

Fig. 9 Mean square errors (MSEs) of 2-level (a) and 3-level (b) QIKF and SQIKF in Example 3. For 2-level SQIKF, $\tau_1 = 1.3634$, $\tau_2 = 1.8$; for 3-level SQIKF, $\tau_1 = 1.1902$, $\tau_2 = 1.3$

level QIKF and SQIKF where, for 2-level SQIKF, $\tau_1 = 1.3634$, $\tau_2 = 1.8$, and for 3-level SQIKF, $\tau_1 = 1.1902$, $\tau_2 = 1.3$. This shows that the 2- and 3-level QIKF fail to track the state over the entire course, but the SQIKF works well. The validity of SQIKF is demonstrated by this example.

7 Conclusions

This work mainly deals with the stability and convergence of Kalman filtering based on quantized measurement innovations. By taking the quantization error as a random perturbation in the observation system, an estimate error analysis system is obtained. Through theoretical analysis, the conditions of stability of the filter are provided. To ensure that the estimated covariance matrix is bounded, a sufficient condition is provided. Then the number of quantization levels to guarantee the stability of QIKF is discussed. The asymptotic stability of the filtering error system is also obtained in the general vector case. We also discuss how to design the modified QIKF to guarantee the stability of QIKF. A scaling QIKF is given to overcome the divergence of QIKF when the number of quantization levels is small. Finally, numerical simulations demonstrate the validity of these theorems and algorithms.

Because of limited space, the stability of the modified Kalman filtering using scaling quantized innovations (SQIKF) for the general vector observation case will be discussed in another paper. Further work is necessary to discuss the existence of a stable state of QIKF. Moreover, an adaptive scheme is under investigation, which can adjust the number of quantization levels in response to the changing environment.

Acknowledgements

The authors would like to thank Prof. Min-yue FU, Prof. Li-hua XIE, and Dr. Ke-you YOU for their helpful advice.

References

Clements, K., Haddad, R., 1972. Approximate estimation for systems with quantized data. *IEEE Trans. Autom. Control*, **17**(2):235-239. [doi:10.1109/TAC.1972.1099954]

Curry, R.E., 1970. Estimation and Control with Quantized Measurements. MIT Research Monograph No. 60.

Curry, R.E., Velde, W.V., Potter, J., 1970. Nonlinear estimation with quantized measurements—PCM, predictive quantization, and data compression. *IEEE Trans. Inf. Theory*, **16**(2):152-161. [doi:10.1109/TIT.1970.1054436]

Duan, Z., Jilkov, V.P., Li, X.R., 2008. State Estimation with Quantized Measurements: Approximate MMSE Approach. Proc. 11th Int. Conf. on Information Fusion, p.1-6.

Fu, M., de Souza, C., 2009. State estimation for linear discrete-time systems using quantized measurements. *Automatica*, **45**(12):2937-2945. [doi:10.1016/j.automatica.2009.09.033]

Fu, M., Xie, L., 2009. Approximate estimation for systems with quantized data. *IEEE Trans. Autom. Control*, **54**(5):1165-1170. [doi:10.1109/TAC.2009.2017815]

Gray, R.M., Neuhoff, D.L., 1998. Quantization. *IEEE Trans. Inf. Theory*, **44**(6):2325-2383. [doi:10.1109/18.720541]

Karlsson, R., Gustafsson, F., 2005a. Filtering and Estimation for Quantized Sensor Information. Technical Report LiTH-ISY-R2674, Department of Electrical Engineering, Linkoping University, Linkoping, Sweden.

Karlsson, R., Gustafsson, F., 2005b. Particle Filtering for Quantized Sensor Information. Proc. 13th European Signal Processing Conf., p.1-4.

Max, J., 1960. Quantizing for minimum distortion. *IEEE Trans. Inf. Theory*, **6**(1):7-12. [doi:10.1109/TIT.1960.1057548]

Msechu, E.J., Roumeliotis, S.I., Ribeiro, A., Giannakis, G.B., 2008. Decentralized quantized Kalman filtering with scalable communication cost. *IEEE Trans. Signal Process.*, **56**(8):3727-3741. [doi:10.1109/TSP.2008.925931]

Ribeiro, A., 2005. Distributed Quantization-Estimation for Wireless Sensor Networks. Master Thesis, University of Minnesota.

Ribeiro, A., Giannakis, G.B., Roumeliotis, S.I., 2006. Soi-kf: distributed Kalman filtering with low-cost communications using the sign of innovations. *IEEE Trans. Signal Process.*, **54**(12):4782-4795. [doi:10.1109/TSP.2006.882059]

Sinopoli, B., Schenato, L., Franceschetti, M., Poolla, K., Jordan, M., Sastry, S., 2004. Kalman filtering with intermittent observations. *IEEE Trans. Autom. Control*, **49**(9):1453-1464. [doi:10.1109/TAC.2004.834121]

Sukhavasi, R.T., Hassibi, B., 2009. The Kalman Like Particle Filter: Optimal Estimation with Quantized Innovations/Measurements. Proc. Joint 48th IEEE Conf. on Decision and Control and 28th Chinese Control Conf., p.4446-4451. [doi:10.1109/CDC.2009.5400517]

Sun, S., Lin, J., Xie, L., Xiao, W., 2007. Approximate Estimation for Systems with Quantized Data. Proc. 22nd IEEE Int. Symp. on Intelligent Control, Part of IEEE Multi-conf. on Systems and Control, p.1-3.

Sviestins, E., Wigren, T., 2000. Optimal recursive state estimation with quantized measurements. *IEEE Trans. Autom. Control*, **45**(4):762-767. [doi:10.1109/9.847118]

Walrand, J., 1972. EE226a—Summary of Lecture 13 and 14 Kalman Filter: Convergence. Available from http://robotics.eecs.berkeley.edu/wlr/226aF05/L13.pdf [Accessed on Oct. 23, 2010].

Xiao, J.J., Cui, S.G., Luo, Z.Q., Goldsmith, A.J., 2006. Power scheduling of universal decentralized estimation in sensor networks. *IEEE Trans. Signal Process.*, **54**(2):413-422. [doi:10.1109/TSP.2005.861898]

Xu, J., Li, J.X., 2011. State estimation with quantized sensor information in wireless sensor networks. *IET Signal Process.*, **5**(1):16-26. [doi:10.1049/iet-spr.2009.0284]

You, K., Xie, L., Sun, S., Xiao, W., 2008. Multiple-Level Quantized Innovation Kalman Filter. Proc. 17th Int. Federation of Automatic Control, p.1420-1425.

You, K., Zhao, Y., Xie, L., 2009. Recursive Quantized State Estimation of Discrete-Time Linear Stochastic Systems. Proc. 7th Asian Control Conf., p.170-175.

You, K., Xie, L., Sun, S., Xiao, W., 2011. Quantized filtering of linear stochastic systems. *Trans. Inst. Meas. Control*, **33**(6):683-698. [doi:10.1177/0142331209342213]

Yu, K., Guo, Y.J., Hedley, M., 2009. TOA-based distributed localization with unknown internal delays and clock frequency offsets in wireless sensor networks. *IET Signal Process.*, **3**(2):106-118. [doi:10.1049/iet-spr:20080029]

Appendix: Proof of Theorem 2

From Algorithm 2,

$$\begin{aligned}\widetilde{\boldsymbol{x}}_{k|k-1} &= \boldsymbol{x}_k - \widehat{\boldsymbol{x}}_{k|k-1}\\ &= \boldsymbol{F}\boldsymbol{x}_{k-1} + \boldsymbol{w}_{k-1} - \boldsymbol{F}\widehat{\boldsymbol{x}}_{k-1|k-1}\\ &= \boldsymbol{F}\widetilde{\boldsymbol{x}}_{k-1|k-1} + \boldsymbol{w}_{k-1}, \qquad \text{(A1)}\\ \widetilde{\boldsymbol{x}}_{k|k} &= \boldsymbol{x}_k - \widehat{\boldsymbol{x}}_{k|k}\\ &= \boldsymbol{x}_k - \widehat{\boldsymbol{x}}_{k|k-1}\\ &\quad -\boldsymbol{K}_k(\boldsymbol{H}\boldsymbol{x}_k + u_k + v_k - \boldsymbol{H}\widehat{\boldsymbol{x}}_{k|k-1})\\ &= (\boldsymbol{I}-\boldsymbol{K}_k\boldsymbol{H})\widetilde{\boldsymbol{x}}_{k|k-1} - \boldsymbol{K}_k(u_k+v_k). \qquad \text{(A2)}\end{aligned}$$

Under Assumptions 1 and 2, and taking the mean square of Eqs. (A1) and (A2) on both sides, we obtain

$$\begin{aligned}\widetilde{\boldsymbol{P}}_{k|k-1} &= \mathrm{E}[\widetilde{\boldsymbol{x}}_{k|k-1}\widetilde{\boldsymbol{x}}_{k|k-1}^{\mathrm{T}}]\\ &= \mathrm{E}[(\boldsymbol{x}_k - \widehat{\boldsymbol{x}}_{k|k-1})(\boldsymbol{x}_k - \widehat{\boldsymbol{x}}_{k|k-1})^{\mathrm{T}}]\\ &= \boldsymbol{F}\widetilde{\boldsymbol{P}}_{k-1|k-1}\boldsymbol{F}^{\mathrm{T}} + \boldsymbol{Q}, \qquad \text{(A3)}\end{aligned}$$

$$\begin{aligned}\widetilde{\boldsymbol{P}}_{k|k} &= \mathrm{E}[\widetilde{\boldsymbol{x}}_{k|k}\widetilde{\boldsymbol{x}}_{k|k}^{\mathrm{T}}]\\ &= \mathrm{E}[(\boldsymbol{x}_k - \widehat{\boldsymbol{x}}_{k|k})(\boldsymbol{x}_k - \widehat{\boldsymbol{x}}_{k|k})^{\mathrm{T}}]\\ &= (\boldsymbol{I}-\boldsymbol{K}_k\boldsymbol{H})\widetilde{\boldsymbol{P}}_{k|k-1}(\boldsymbol{I}-\boldsymbol{K}_k\boldsymbol{H})^{\mathrm{T}}\\ &\quad +\boldsymbol{K}_k\mathrm{E}[(u_k+v_k)(u_k+v_k)]\boldsymbol{K}_k^{\mathrm{T}}\\ &= (\boldsymbol{I}-\boldsymbol{K}_k\boldsymbol{H})\widetilde{\boldsymbol{P}}_{k|k-1}(\boldsymbol{I}-\boldsymbol{K}_k\boldsymbol{H})^{\mathrm{T}}\\ &\quad +\boldsymbol{K}_k(\boldsymbol{C}_k + \boldsymbol{R} + \mathrm{E}[u_kv_k + v_ku_k])\boldsymbol{K}_k^{\mathrm{T}}. \qquad \text{(A4)}\end{aligned}$$

It is easy to verify that the above two equations can be rewritten as follows:

$$\begin{aligned}\widetilde{\boldsymbol{P}}_{k|k} &= (\boldsymbol{I}-\boldsymbol{K}_k\boldsymbol{H})(\boldsymbol{F}\widetilde{\boldsymbol{P}}_{k-1|k-1}\boldsymbol{F}^{\mathrm{T}} + \boldsymbol{Q})(\boldsymbol{I}-\boldsymbol{K}_k\boldsymbol{H})^{\mathrm{T}}\\ &\quad +\boldsymbol{K}_k(\boldsymbol{C}_k + \boldsymbol{R} + \mathrm{E}[u_kv_k + v_ku_k])\boldsymbol{K}_k^{\mathrm{T}}. \qquad \text{(A5)}\end{aligned}$$

Hence, it follows from Eqs. (27), (11), (12), and (17) that

$$\begin{aligned}\widetilde{\boldsymbol{P}}_{k|k} &= (\boldsymbol{I}-\boldsymbol{K}_k\boldsymbol{H})(\boldsymbol{F}\widetilde{\boldsymbol{P}}_{k-1|k-1}\boldsymbol{F}^{\mathrm{T}} + \boldsymbol{Q})(\boldsymbol{I}-\boldsymbol{K}_k\boldsymbol{H})^{\mathrm{T}}\\ &\quad +\boldsymbol{K}_k\boldsymbol{R}\boldsymbol{K}_k^{\mathrm{T}} + \boldsymbol{K}_k(\boldsymbol{C}_k + \mathrm{E}[u_kv_k + v_ku_k])\boldsymbol{K}_k^{\mathrm{T}}\\ &= \boldsymbol{F}\widetilde{\boldsymbol{P}}_{k-1|k-1}\boldsymbol{F}^{\mathrm{T}} + \boldsymbol{Q} - \boldsymbol{K}_k\boldsymbol{S}_k\boldsymbol{K}_k^{\mathrm{T}}\\ &\quad +\boldsymbol{K}_k(\boldsymbol{C}_k + \mathrm{E}[u_kv_k + v_ku_k])\boldsymbol{K}_k^{\mathrm{T}}. \qquad \text{(A6)}\end{aligned}$$

By Lemma 1, there is a number α_L, such that $\boldsymbol{C}_k \leq \alpha_L\boldsymbol{S}_k$. Because u_k and v_k are both scalar, by the Cauchy inequality and Theorem 1, we have

$$\begin{aligned}\mathrm{E}[u_kv_k] &\leq (\mathrm{E}[u_ku_k])^{1/2}(\mathrm{E}[v_kv_k])^{1/2}\\ &\leq (\alpha_L\boldsymbol{S}_k)^{1/2}\boldsymbol{R}^{1/2}\\ &\leq \sqrt{\alpha_L}\boldsymbol{S}_k. \qquad \text{(A7)}\end{aligned}$$

Thus, we have

$$\begin{aligned}\widetilde{\boldsymbol{P}}_{k|k} &= \boldsymbol{F}\widetilde{\boldsymbol{P}}_{k-1|k-1}\boldsymbol{F}^{\mathrm{T}} + \boldsymbol{Q} - \boldsymbol{K}_k\boldsymbol{S}_k\boldsymbol{K}_k^{\mathrm{T}}\\ &\quad +\boldsymbol{K}_k(\boldsymbol{C}_k + \mathrm{E}[u_kv_k + v_ku_k])\boldsymbol{K}_k^{\mathrm{T}}\\ &\leq \boldsymbol{F}\widetilde{\boldsymbol{P}}_{k-1|k-1}\boldsymbol{F}^{\mathrm{T}} + \boldsymbol{Q} - \boldsymbol{K}_k\boldsymbol{S}_k\boldsymbol{K}_k^{\mathrm{T}}\\ &\quad +(\alpha_L + 2\sqrt{\alpha_L})\boldsymbol{K}_k\boldsymbol{S}_k\boldsymbol{K}_k^{\mathrm{T}}\\ &= \boldsymbol{F}\widetilde{\boldsymbol{P}}_{k-1|k-1}\boldsymbol{F}^{\mathrm{T}} + \boldsymbol{Q}\\ &\quad -[1-(\alpha_L + 2\sqrt{\alpha_L})]\boldsymbol{K}_k\boldsymbol{S}_k\boldsymbol{K}_k^{\mathrm{T}}. \qquad \text{(A8)}\end{aligned}$$

If $\alpha_L + 2\sqrt{\alpha_L} \in (0, 1-\overline{\lambda})$, i.e., $1-(\alpha_L + 2\sqrt{\alpha_L}) > \overline{\lambda}$, then by Lemma 2 and Eqs. (11) and (14), the MARE

$$\begin{aligned}\overline{\boldsymbol{P}}_{k|k} &= \boldsymbol{F}\overline{\boldsymbol{P}}_{k-1|k-1}\boldsymbol{F}^{\mathrm{T}} + \boldsymbol{Q} - [1-(\alpha_L + 2\sqrt{\alpha_L})]\boldsymbol{K}_k\\ &\quad \cdot[\boldsymbol{H}(\boldsymbol{F}\overline{\boldsymbol{P}}_{k-1|k-1}\boldsymbol{F}^{\mathrm{T}} + \boldsymbol{Q})\boldsymbol{H}^{\mathrm{T}} + \boldsymbol{R}]\boldsymbol{K}_k^{\mathrm{T}} \qquad \text{(A9)}\end{aligned}$$

has a bounded positive definite solution $\boldsymbol{P}_{k|k}$. Finally, by the optimality of KF and inequality (A8), we obtain inequality (42). This completes the proof.