OPTIMAL ERROR FEEDBACK FILTERS FOR UNIFORM QUANTIZERS AT REMOTE SENSORS

Shuichi Ohno⋆*, Yuji Wakasa*†*, and Makoto Nagata*⋆

⋆Hiroshima University, 1-4-1 Kagamiyama, Higashi-Hiroshima, 739-8527, JAPAN
†Yamaguchi University, 2-16-1 Tokiwadai, Ube 755-8611, JAPAN

ABSTRACT

Optimal error feedback filters for A/D converters of remote sensors in networked control systems are studied. It is shown that if the transfer function from the quantization error to the output of interest has a minimum phase, then its inverse is the optimal error feedback filter in terms of any kind of norms of errors. For general LTI systems, the design of error feedback filters is formulated as an optimization problem, which can be numerically solved. A design example is provided to demonstrate the effectiveness of our proposed method.

***Index Terms*—** quantization, uniform quantizer, network control system, error feedback filter

1. INTRODUCTION

The networked control system is a system composed of multiple (sub-)systems that are connected via wirelne and wireless communications to exchange information (e.g., see [1] and the references therein). In a networked control system, control signals from the controller are transmitted to the plant to be operated. Signals observed by sensors for the plant are transmitted to the controller, from which the control signals are generated.

If one adopts a digital communication for the transmission, then continuous-valued signals have to be discretized and quantized. If the systems are connected through a wireline and have sufficient resources, then one may avoid the effects of quantization, since one can increase the transmission rate easily and hence use a sufficient number of bits to represent the continuous-valued signals. However, if not, one may have to consider the effects of quantization, since wireless communications are not that robust.

Quantization is an old yet important topic, which has been studied well since the advent of digital processors. Quantization is the mapping from the continuous-valued signal to the discrete-valued signal. If the distribution of the continuous-valued signal is known, then the vector quantization [2, Sec.3.4] is effective, which determines the discrete-valued signal based on the distribution of the continuous-valued signal. When the knowledge on the system and the signals are not available, uniformly distributed values are assigned to the discrete-valued signal. Since the error affects the performance of the system, the error has been analyzed by many researchers (see [3] and the references therein).

The quantization error signal is often modeled as white noise. Under the white noise assumption, there have been many proposals to decrease the effects of quantization errors. In [4], the optimal realization of the digital filter that minimizes the output noise variance due to the roundoff is developed. Error spectrum shaping, which is also called error feedback, is a technique to reduce the effects of quantization errors by feeding back filtered errors to the quantizer [5, 6]. The optimal FIR error feedback filters are designed in [7]. Most of them are FIR filters, since it is difficult to assure the stability of designed filters.

Without the white noise assumption, based on the l_∞ norm, the optimal error feedback filter for A/D converters of the controller is presented for systems having minimum phases in [8]. To obtain optimal error feedback filters at the controller, a numerical design using linear programming is proposed in [9]. To guarantee the stability of error feedback filters, [10] makes a good use of the invariant set theory for discrete-time LTI systems [11] and provides a numerical design of error feedback filters based on linear matrix inequalities (LMIs).

This paper studies the optimal IIR error feedback filters for A/D converters of remote sensors, where each sensor transmits its observed signals to the controller but works independently, i.e, it does not utilize the signals from other sensors. First, we show that if the transfer function from the quantization error to the interested output has a minimum phase, then its inverse is the optimal error feedback filter in terms of any kind of norms of the errors. Then, for general LTI systems, we formulate our design of error feedback filters as an optimization problem. For a fixed scalar parameter variable, the optimization problem is converted into a convex optimization problem, which can be efficiently solved by existing convex optimization solvers. Then, the optimization is achieved by finding the optimal scalar variable. A design example is provided to demonstrate the effectiveness of error feedback filters designed by our proposed method.

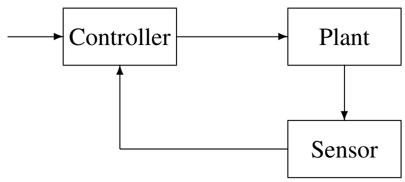

Fig. 1. A network control system.

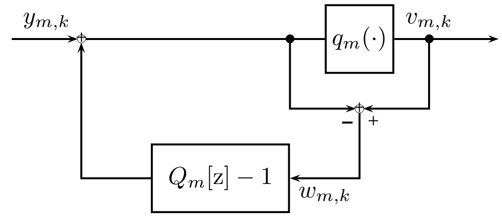

Fig. 2. Quantizer with an error feedback filter

2. NETWORKED SYSTEMS AND QUANTIZATION

Let us consider a networked control system depicted in Fig. 1. In this paper, we only deal with SISO plants.

We consider discretized systems. Let the plant be an LTI linear system that is controllable and observable. We denote the output of the mth sensor as $y_{m,k}$. Since the control input u_k is generated based on the observed signals in a feedback system, one may express u_k as $u_k = \sum_{m=1}^{M} K_m[\mathrm{z}]y_{m,k}$, where $K_m[\mathrm{z}]$ is the feedback system.

The output of each sensor is quantized by a quantizer that maps a continuous value to a discrete value. Simple static uniform quantizers are often utilized. For the continuous-valued input x, let the output of the static uniform quantizer be $q(x) = \lfloor \frac{x}{d} + \frac{1}{2} \rfloor d$, where d is the quantization interval and $\lfloor a \rfloor$ denotes the largest integer not exceeding a.

We evaluate the effect of quantization, expressing the quantized signal of $y_{m,k}$ as $v_{m,k}$. If one denotes the quantization error of the mth quantizer as $e_{m,k}$, then one can express $v_{m,k} = y_{m,k} + e_{m,k}$.

Let z_k be the output of the system without quantization and $z_{Q,k}$ be the output of the plant with quantization. Their difference

$$\epsilon_k = z_{Q,k} - z_k \tag{1}$$

can be used to evaluate the effect of the quantization.

We express the transfer function from the quantization error of the mth quantizer to the output z_k as $H_m[\mathrm{z}]$. Since the plant is linear, the effect on the output of the error $e_{m,k}$ is given by

$$\epsilon_{m,k} = H_m[\mathrm{z}]e_{m,k}. \tag{2}$$

Then, we have $\epsilon_k = \sum_{m=1}^{M} \epsilon_{m,k}$. When sensors can not communicate with each other, it is reasonable to independently minimize $\epsilon_{m,k}$ for each m to mitigate the effect of the quantization.

To measure the error signal, we use the l_p norm defined for a finite p as

$$||x_k||_p = \left[\sum_{k=0}^{\infty} |x_k|^p\right]^{\frac{1}{p}} \tag{3}$$

and for $p = \infty$ as

$$||x_k||_\infty = \sup_k |x_k|. \tag{4}$$

3. QUANTIZATION WITH AN ERROR FEEDBACK FILTER

To mitigate the effect of the quantization by a static uniform quantizer, we utilize quantizer having an error feedback filter, which is originally developed to reduce quantization errors in digital filters and is also called *error spectrum shaping* [5–7].

Fig. 2 depicts a schematic diagram of our quantizer. Our quantizer is composed of the static uniform quantizer $q_m(\cdot)$ and an error feedback filter $Q_m[\mathrm{z}] - 1$, where $Q_m[\mathrm{z}]$ is the transfer function with $Q_m[\infty] = 1$.

The difference signal $w_{m,k}$ between the input and the output of the static quantizer is filtered by a filter $Q_m[\mathrm{z}] - 1$ and then is fed back to $y_{m,k}$. It should be remarked that if $Q_m[\mathrm{z}] \neq 1$, then $w_{m,k}$ is not equal to the quantization error $v_{m,k} - y_{m,k}$. One can show that

$$\epsilon_{m,k} = H_m[\mathrm{z}]Q_m[\mathrm{z}]w_{m,k}. \tag{5}$$

Suppose that $H_m[\mathrm{z}]$ can be expressed with a product of delay z^{-D_m} and a proper function $\tilde{H}_m[\mathrm{z}] = \sum_{k=0}^{\infty} \tilde{h}_{m,k}\mathrm{z}^{-k}$ having $\tilde{h}_{m,k} \neq 0$ as $H_m[\mathrm{z}] = \mathrm{z}^{-D_m}\tilde{H}_m[\mathrm{z}]$. The induced norm of the system $H[\mathrm{z}] = \sum_{k=0}^{\infty} h_k\mathrm{z}^{-k}$ by the l_p norm is defined

$$||H[\mathrm{z}]||_{ip} = \sup_{\{x_k\}\neq 0} \frac{||H[\mathrm{z}]x_k||_p}{||x_k||_p} \tag{6}$$

where $||H[\mathrm{z}]x_k||_p$ is the l_p norm of the output of the system.

From the property of norms, we have

$$\begin{aligned} ||\epsilon_{m,k}||_p &\leq ||H_m[\mathrm{z}]Q_m[\mathrm{z}]||_{ip}||w_{m,k}||_p \\ &\leq |\tilde{h}_{m,0}|||w_{m,k}||_p \\ &\quad +||H_m[\mathrm{z}]Q_m[\mathrm{z}] - \tilde{h}_{m,0}\mathrm{z}^{-D_m}||_{ip}||w_{m,k}||_p. \end{aligned} \tag{7}$$

The last term vanishes if $H_m[\mathrm{z}]Q_m[\mathrm{z}] - \tilde{h}_{m,0}\mathrm{z}^{-D_m} = 0$, i.e., $Q_m[\mathrm{z}] = \tilde{h}_{m,0}\mathrm{z}^{-D_m}H_m^{-1}[\mathrm{z}] = \tilde{h}_{m,0}\tilde{H}_m^{-1}[\mathrm{z}]$. If $H_m[\mathrm{z}]Q_m[\mathrm{z}] = \tilde{h}_{m,0}\mathrm{z}^{-D_m}$, then $||\epsilon_{m,k}||_p = |\tilde{h}_{m,0}|||w_{m,k}||_p$. Thus, the equality holds true for (7). This shows that the norm of the error is minimized if and only if $H_m[\mathrm{z}]Q_m[\mathrm{z}] = \tilde{h}_{m,0}\mathrm{z}^{-D_m}$.

The optimal quantizer for minimum-phase systems in terms of the l_∞ norm has been provided in [8] based on state-space expressions. We have proven that the inverse of the minimum-phase system is the optimal error feedback filter in terms of any kind of norm, using not state-space expressions but transfer functions.

Since $\tilde{H}_m[\mathrm{z}]$ is not guaranteed to have minimum-phase in practice, we will consider the design of a stable feedback filter which minimizes $||\epsilon_{m,k}||_\infty$. We omit the subscript m since filters can be designed separately.

4. DESIGN OF ERROR FEEDBACK FILTERS FOR QUANTIZATION

Our objective is to design a stable error feedback filter that minimizes the effect of the quantization. Mathematically, we can formulate our design problem as

$$\min_{Q[\mathrm{z}]\in RH_\infty} ||\epsilon_k||_\infty \tag{8}$$

subject to $Q[0]=1$.

The signal w_k which is the difference between the input and the output of the mth static uniform quantizer satisfies $|w_k| \leq \frac{d}{2}$, which means $||w_k||_\infty \leq \frac{d}{2}$. Since the transfer function from w_k to ϵ_k is linear, we can put $d=2$ without loss of generality so that $|w_k| \leq 1$ and hence $|w_k|^2 \leq 1$.

The composite system $H[\mathrm{z}]Q[\mathrm{z}]$ has to be internally stable. Let us denote the (A,B,C,D) matrices of a state-space realization of $H[\mathrm{z}]$ as $(A_h,B_h,C_h,0)$. Let the order of $Q[\mathrm{z}]$ be n and let $(A_q,B_q,C_q,1)$ be (A,B,C,D) matrices of a state-space realization of $Q[\mathrm{z}]$. Then, one can express the state-space realization of $H[\mathrm{z}]Q[\mathrm{z}]$ as

$$\begin{aligned} x_{k+1} &= \mathcal{A}x_k + \mathcal{B}w_k && (9)\\ \epsilon_k &= \mathcal{C}x_k && (10) \end{aligned}$$

where

$$\mathcal{A} = \begin{bmatrix} A_h & B_hC_q \\ \mathbf{0} & A_q \end{bmatrix},\ \mathcal{B} = \begin{bmatrix} B_h \\ B_q \end{bmatrix},\ \mathcal{C} = \begin{bmatrix} C_h & \mathbf{0} \end{bmatrix}$$

We borrow the idea of [10] to design our quantizer, which utilizes the invariant set of a discrete-time system [11]:

Definition 1 *Let $x_k \in \mathbb{R}^n$ be the state vector of the LTI system given by*

$$x_{k+1} = Ax_k + Bw_k \tag{11}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$ and $w_k \in \mathbb{R}^m$. A set $\mathcal{X}$ that satisfies $x_{k+1} \in \mathcal{X}$ if $x_k \in \mathcal{X}$ and $||w_k||_2 \leq 1$ is called an invariant set of the system (11).

The following proposition describes how to obtain an ellipsoid which is an invariant set of the system (11) [11]:

Proposition 1 *Let $\mathcal{E}(\mathcal{P})$ be the ellipsoid defined by an $n\times n$ real symmetric matrix $\mathcal{P} \succeq \mathbf{0}$ as $\mathcal{E}(\mathcal{P}) = \{x \in \mathbb{R}^n : x^T\mathcal{P}x \leq 1\}$.*

The ellipsoid $\mathcal{E}(\mathcal{P})$ is an invariant set of the system (11) *if and only if there exists a scalar $\alpha \in [0, 1-\rho^2(A)]$ which satisfies*

$$\begin{bmatrix} A^T\mathcal{P}A - (1-\alpha)\mathcal{P} & A^T\mathcal{P}B \\ B^T\mathcal{P}A & B^T\mathcal{P}B - \alpha I \end{bmatrix} \succeq \mathbf{0} \tag{12}$$

where $\rho(A)$ is the spectral radius of A.

For $\mathcal{P} \succeq \mathbf{0}$, using the Schur complement, we can express (12) as

$$\begin{bmatrix} (1-\alpha)\mathcal{P} & \mathbf{0} & A^T\mathcal{P} \\ \mathbf{0} & \alpha & B^T\mathcal{P} \\ \mathcal{P}A & \mathcal{P}B & \mathcal{P} \end{bmatrix} \succeq \mathbf{0}. \tag{13}$$

For $\mathcal{P}$ that satisfies (13), we have $\sup_{x_k\in\mathcal{E}(\mathcal{P})} |\epsilon_k| = |\mathcal{C}\mathcal{P}^{-1}\mathcal{C}^T|^{\frac{1}{2}}$. Since $x_k \in \mathcal{E}(\mathcal{P})$ is satisfied if (A_q,B_q,C_q) and $\mathcal{P}$ meet (13) for given (A_h,B_h,C_h) and for a fixed α, we can obtain the optimal (A_q,B_q,C_q) by solving the minimization problem:

$$\min_{(A_q,B_q,C_q),\mathcal{P}} |\mathcal{C}\mathcal{P}^{-1}\mathcal{C}^T|^{\frac{1}{2}} \tag{14}$$

subject to (13). It should be remarked that the resultant quantizer is stable, since $\mathcal{A}$ has to be a Schur matrix to satisfy (13).

Introducing a variable γ, we can describe our problem as

$$\min_{(A_q,B_q,C_q),\mathcal{P},\gamma} \gamma \tag{15}$$

subject to $\mathcal{C}\mathcal{P}^{-1}\mathcal{C}^T \leq \gamma$ and (13). Using the Schur complement, we can express $\mathcal{C}\mathcal{P}^{-1}\mathcal{C}^T \leq \gamma$ as an LMI given by

$$\begin{bmatrix} \mathcal{P} & \mathcal{C}^T \\ \mathcal{C} & \gamma \end{bmatrix} \succeq \mathbf{0}. \tag{16}$$

Eq. (16) is convex in the optimization variables, while Eq. (13) is a bilinear matrix inequality (BMI) of the variables. In general, BMIs are not convex and NP-hard to solve numerically. Fortunately, we can covert the BMI (13) to an LMI by using the change of variables proposed in [13].

Let the order of $Q[\mathrm{z}]$ be equivalent to the system, which is denoted by n. Let us define matrices $\{M_A, M_B, M_C, M_\mathcal{P}\}$ as

$$M_A = \begin{bmatrix} A_hP_f + B_hW_f & A_h \\ L & P_gA_h \end{bmatrix},\quad M_B = \begin{bmatrix} B_h \\ W_g \end{bmatrix}$$

$$M_C = \begin{bmatrix} C_hP_f & \mathbf{0} \end{bmatrix},\quad M_\mathcal{P} = \begin{bmatrix} P_f & I_n \\ I_n & P_g \end{bmatrix}$$

where P_f and P_g are $n\times n$ positive definite matrices, $W_f \in \mathbb{R}^{1\times n}$, $W_g \in \mathbb{R}^{n\times 1}$, and $L \in \mathbb{R}^{n\times n}$.

Theorem 1 [13] proves that (13) for the original variables $\{\mathcal{A}, \mathcal{B}, \mathcal{C}\}$ is equivalent to the matrix inequality for the new variables given by

$$\begin{bmatrix} (1-\alpha)M_\mathcal{P} & \mathbf{0} & M_A^T \\ \mathbf{0} & \alpha & M_B^T \\ M_A & M_B & M_\mathcal{P} \end{bmatrix} \succeq \mathbf{0}, \tag{17}$$

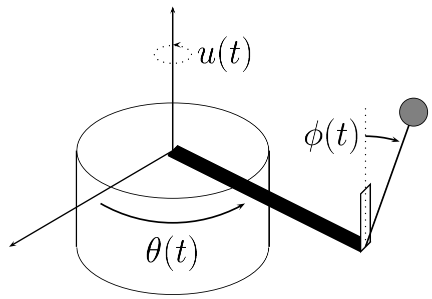

Fig. 3. A rotary inverted pendulum.

which is an LMI.

Moreover, (16) can be expressed by the new variables as

$$\begin{bmatrix} M_{\mathcal{P}} & M_C^T \\ M_C & \gamma \end{bmatrix} \succeq \mathbf{0}. \tag{18}$$

Then, the minimization problem

$$\min_{P_f, P_g, P_h, W_f, W_g, L, \gamma} \gamma \tag{19}$$

subject to (17) and (18), gives the minimum of the original minimization problem for a given α. Once the optimal $P_f, P_g, P_h, W_f, W_g, L$ are given, from Lemma 2 in [13], the optimal (A_h, B_h, C_h) can be obtained by

$$A_q = [B_p W_f - P_g^{-1}(L - P_g A_p P_f)](P_f - P_g^{-1})^{-1}, \tag{20}$$
$$B_q = B_h - P_g^{-1} W_g, \quad C_q = W_f(P_f - P_g^{-1})^{-1}. \tag{21}$$

For a fixed α, the minimization problem is a semidefinite program, which can be numerically solved by existing optimization packages, e.g., CVX [14]. Then, all we have to do is to find α which gives the minimum. Since $\mathcal{A}$ is our design parameter, a line search for $\alpha \in (0,1)$ is required to obtain the minimum.

5. NUMERICAL EXAMPLE

For our plant, let us consider the rotary inverted pendulum depicted in Fig. 3. A motor rotates the main body in the horizontal plane to control the pendulum connected at the end of the rotary arm. The torque $u(t)$ is applied to actuate the pendulum. Let the yaw angle of the arm be $\phi(t)$. The pendulum freely swings about a pitch angle $\theta(t)$ in the vertical plane to the arm. If $\theta(t) = 0$, then the pendulum is balanced in the inverted position.

The state of the rotary inverted pendulum is defined as $x^T(t) = [\phi(t), \theta(t), \dot{\phi}(t), \dot{\theta}(t)]$. We discretize the continuous system with the sampling period $T_s = 0.01$ to obtain the discrete-time system.

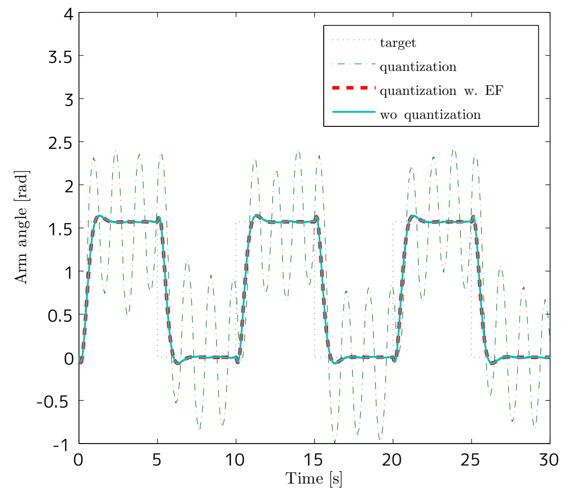

Fig. 4. Arm angles without quantization and with quantization by our error feedback quantizer (dashed line) as well as by the static quantizer (long dashed short dashed line).

Our control objective is to periodically change the arm angle, while keeping the stability of the rotary inverted pendulum. The target value of the arm angle $\bar{\phi}(t)$ is for $k = 0, 1, \ldots$

$$\bar{\phi}(t) = \begin{cases} \frac{\pi}{2} & (10k \le t < 5 + 10k) \\ 0 & (5 + 10k \le t < 10(k+1)) \end{cases}. \tag{22}$$

To control the system, we employ the state feedback control whose gain is determined by the linear quadratic regulator (LQR) technique to minimize $\sum_{k=0}^{\infty} \left(x_k^T Q_{lqr} x_k + r|u_k|^2 \right)$ where the weights are $Q_{lqr} = \mathrm{diag}[10, 2, 0.5, 0]$ and $r = 0.05$.

All of the state variables are assumed to be available. The state variables are quantized and transmitted to the controller. The transfer function $H_m[\mathrm{z}]$ is given by $C(zI - A - BK)^{-1} BK_m$. Since $K_m[\mathrm{z}]$ for $m = 1, \ldots, 4$ are scalars, we design a quantizer for $C(zI - A - BK)^{-1} B$. In this case, we have $(A_h, B_h, C_h) = (A - BK, B, C)$. Then, our optimization problem is numerically solved by CVX [14], an efficient convex optimization solver.

For the quantization interval $d = 0.5$, Fig. 4 compares the arm angle without quantization and with quantization by our error feedback quantizer (dashed line) as well as by the static quantizer (long dashed short dashed line). The dotted line denotes the target value. The arm angle of our designed quantizer is almost the same as the arm angle of the control without quantization, while the arm angle with the static quantizer is fluctuated by quantization errors, which shows the importance and the effectiveness of the optimal error feedback filter for the quantization with the knowledge of the system.

6. REFERENCES

[1] N. Ploplys, P. Kawka, and A. Alleyne, "Closed-loop control over wireless networks," *IEEE Control Systems*, vol. 24, no. 3, pp. 58–71, Jun 2004.

[2] J. Proakis, *Digital Communications*, 4th ed. New York: McGraw-Hill, 2001.

[3] B. Widrow and I. Kollár, *Quantization Noise: Roundoff Error in Digital Computation, Signal Processing, Control, and Communications*. Cambridge University Press New York, 2008.

[4] C. Mullis and R. Roberts, "Synthesis of minimum roundoff noise fixed point digital filters," *IEEE Transactions on Circuits and Systems*, vol. 23, no. 9, pp. 551–562, Sep 1976.

[5] Tran-Thong and B. Liu, "Error spectrum shaping in narrow-band recursive filters," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 25, no. 2, pp. 200–203, Apr 1977.

[6] W. Higgins and D. C. J. Munson, "Noise reduction strategies for digital filters: Error spectrum shaping versus the optimal linear state-space formulation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 30, no. 6, pp. 963–973, Dec 1982.

[7] T. Laakso and I. Hartimo, "Noise reduction in recursive digital filters using high-order error feedback," *IEEE Transactions on Signal Processing*, vol. 40, no. 5, pp. 1096–1107, May 1992.

[8] S. Azuma and T. Sugie, "Optimal dynamic quantizers for discrete-valued input control," *Automatica*, vol. 44, no. 2, pp. 396–406, Feb. 2008. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0005109807003068

[9] ——, "Synthesis of Optimal Dynamic Quantizers for Discrete-Valued Input Control," *IEEE Transactions on Automatic Control*, vol. 53, no. 9, pp. 2064–2075, Oct. 2008. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4639491

[10] K. Sawada and S. Shin, "Dynamic quantizer synthesis based on invariant set analysis for SISO systems with discrete-valued input," in *the 19th International Symposium on Mathematical Theory of Networks and Systems*, 2010, pp. 1385–1390.

[11] H. Shingin and Y. Ohta, "Optimal invariant sets for discrete-time systems: Approximation of reachable sets for bounded inputs," in *10th IFAC/IFORS/IMACS/IFIP Symposium on Large Scale Systems: Theory and Applications (LSS)*, 2004, pp. 401–406.

[12] B. L. Ho and R. E. Kalman, "Effective construction of linear, state-variable models from input/output functions," *Automatisierungstechnik*, vol. 14, no. 1-12, pp. 545–548, 1966.

[13] I. Masubuchi, A. Ohara, and N. Suda, "LMI-based controller synthesis: A unified formulation and solution," *International Journal of Robust and Nonlinear Control*, vol. 8, no. 8, p. 669686, July 1998.

[14] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.0 beta," http://cvxr.com/cvx, Sep. 2012.