Global Optimal Realizations of Finite Precision Digital Controllers

Jun Wu †, Sheng Chen ‡, Gang Li § and Jian Chu †

† National Key Laboratory of Industrial Control Technology
Institute of Advanced Process Control, Zhejiang University
Hangzhou, 310027, P. R. China

‡ Department of Electronics and Computer Science
University of Southampton, Southampton SO17 1BJ, U.K.

§ School of Electrical and Electronic Engineering
Nanyang Technological University, Singapore

Abstract

The paper analyzes global solutions to the optimal digital controller realization problem based on maximizing a finite word length (FWL) closed-loop stability measure. For each closed-loop eigenvalue, a single-pole FWL stability function is first introduced, and a single-pole FWL stability measure is then defined as the maximum of the corresponding single-pole stability function over all the controller realizations. It is shown that the minimum of the single-pole stability measures for all the closed-loop eigenvalues is an upper bound of the optimal value for the optimal realization problem. An analytical method to compute a single-pole stability measure is developed, and an expression for all the realizations which achieve a given single-pole measure is derived. When a realization, which is a solution of the minimum single-pole measure, further satisfies the condition that the values of its all the single-pole stability functions are not less than the minimum single-pole measure, the minimum single-pole measure is the optimal value of the optimal realization problem and this realization is the solution for the optimal realization problem. An algorithm is presented to compute an optimal FWL controller realization. Unlike most of the existing methods relying on numerical optimization search algorithms, which can be computationally expensive and may easily be trapped at local optimal solutions, the proposed analytical approach guarantees to find a global optimal controller realization.

Index Terms – digital controller, finite word length, closed-loop stability, fragility, optimization.

I. Introduction

The classical control system design often assumes that the controller is implemented exactly. This assumption is usually justified on the ground that the plant uncertainty is the most significant source of uncertainty in the control system. It has been realized, however, that the controller uncertainties caused by finite-precision implementation also have significant influence on the performance of the control system. Failures to take into account the uncertainties in the controller implementation may result in a controller that is fragile [1]. The fragility issues are strongly related to and interconnected with the FWL controller implementation issues [2],[3]. This paper considers the case that the controller is implemented using a fixed-point processor. In real-time applications where computational efficiency is critical, a digital controller implemented with fixed-point arithmetic has advantages over floating-point implementation. However, the detrimental FWL effects are markedly increased in fixed-point implementation due to a reduced precision.

It has been noted that a controller design can be implemented with different realizations and the FWL effect on the closed-loop stability depends on the controller realization. This property can be utilized to select controller realization in order to improve the robustness of FWL closed-loop stability, and many studies have investigated digital controller realizations with FWL considerations [4]–[9]. A basic idea in these studies has been to define some FWL closed-loop stability measure which depends on the controller realization and to search for an "optimal" realization by optimizing the measure. In the work [5], an FWL stability measure based on closed-loop eigenvalue sensitivity was derived and the optimal controller realization problem was defined as the maximization of this measure over all the possible controller realizations. An analytical solution to this optimal realization problem was attempted in [5]. However, the conditions presented in [5] were not sufficient to provide an optimal realization that maximizes the FWL stability measure [10].

Due to the lack of analytical solutions to optimal FWL controller realization problems, numerical optimization methods have been adopted to search for optimal realizations [6]–[9]. A numerical optimization approach can be effective if the dimension of the problem is small. In general, however, an optimal FWL realization problem is

Contact author: S. Chen, Tel/Fax: +44 (0)23 8059 6660/4508, Email: sqc@ecs.soton.ac.uk

a highly complicated nonlinear and non-convex optimization problem, especially when the order of the controller is large. Methods based on numerical optimization algorithms are then computationally expensive, and chances of search being trapped at some bad local solutions increase for large-scale problems. The main contribution of this paper is to derive an analytical solution for the optimal FWL realization problem defined in [5], which guarantees to achieve global optimal solutions.

II. Problem definition

Consider the discrete-time closed-loop control system, consisting of a linear time-invariant plant $P(z)$ and a digital controller $C(z)$. $P(z)$ is assumed to be strictly proper with a state-space description $(\mathbf{A}_P, \mathbf{B}_P, \mathbf{C}_P)$, where $\mathbf{A}_P \in \mathcal{R}^{m\times m}$, $\mathbf{B}_P \in \mathcal{R}^{m\times l}$ and $\mathbf{C}_P \in \mathcal{R}^{q\times m}$. Let $(\mathbf{A}_C, \mathbf{B}_C, \mathbf{C}_C, \mathbf{D}_C)$ be a state-space description of $C(z)$, with $\mathbf{A}_C \in \mathcal{R}^{n\times n}$, $\mathbf{B}_C \in \mathcal{R}^{n\times q}$, $\mathbf{C}_C \in \mathcal{R}^{l\times n}$ and $\mathbf{D}_C \in \mathcal{R}^{l\times q}$. A linear system with a given transfer function matrix has an infinite number of state-space descriptions. In fact, if $(\mathbf{A}_C^0, \mathbf{B}_C^0, \mathbf{C}_C^0, \mathbf{D}_C^0)$ is a state-space description of $C(z)$, all the state-space descriptions of $C(z)$ form a realization set

$$\mathcal{S}_C \triangleq \{(\mathbf{A}_C, \mathbf{B}_C, \mathbf{C}_C, \mathbf{D}_C) | \mathbf{A}_C = \mathbf{T}^{-1}\mathbf{A}_C^0\mathbf{T}, \\ \mathbf{B}_C = \mathbf{T}^{-1}\mathbf{B}_C^0, \mathbf{C}_C = \mathbf{C}_C^0\mathbf{T}, \mathbf{D}_C = \mathbf{D}_C^0\} \tag{1}$$

where $\mathbf{T} \in \mathcal{R}^{n\times n}$ is an arbitrary non-singular matrix. Denote $N \triangleq (l+n)(q+n)$ and

$$\mathbf{X} \triangleq \begin{bmatrix} \mathbf{D}_C & \mathbf{C}_C \\ \mathbf{B}_C & \mathbf{A}_C \end{bmatrix} \\ = \begin{bmatrix} x_1 & x_{l+n+1} & \cdots & x_{N-l-n+1} \\ x_2 & x_{l+n+2} & \cdots & x_{N-l-n+2} \\ \vdots & \vdots & \cdots & \vdots \\ x_{l+n} & x_{2l+2n} & \cdots & x_N \end{bmatrix}. \tag{2}$$

The stability of the closed-loop control system depends on the eigenvalues of the closed-loop transition matrix

$$\overline{\mathbf{A}}(\mathbf{X}) = \begin{bmatrix} \mathbf{A}_P + \mathbf{B}_P\mathbf{D}_C\mathbf{C}_P & \mathbf{B}_P\mathbf{C}_C \\ \mathbf{B}_C\mathbf{C}_P & \mathbf{A}_C \end{bmatrix} \\ = \begin{bmatrix} \mathbf{A}_P & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{B}_P & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n \end{bmatrix} \mathbf{X} \begin{bmatrix} \mathbf{C}_P & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n \end{bmatrix} \\ \triangleq \mathbf{M}_0 + \mathbf{M}_1\mathbf{X}\mathbf{M}_2 \tag{3}$$

where the zero matrix $\mathbf{0}$ has an appropriate dimension. All the different realizations $\mathbf{X}$ in $\mathcal{S}_C$ have exactly the same set of closed-loop poles if they are implemented with infinite precision. Since the closed-loop system has been designed to be stable, all the eigenvalues $\lambda_i(\overline{\mathbf{A}}(\mathbf{X}))$, $1 \le i \le m+n$, are within the unit disk.

When a controller $\mathbf{X}$ is implemented with a fixed-point processor, it is perturbed to $\mathbf{X} + \Delta\mathbf{X}$ due to the FWL effect. Each element of $\Delta\mathbf{X}$ is bounded by $\pm\varepsilon$, that is,

$$\max_{j\in\{1,\cdots,N\}} |\Delta x_j| \le \varepsilon \tag{4}$$

where ε is the maximum round-off error of the fixed-point processor. With the perturbation $\Delta\mathbf{X}$, $\lambda_i(\overline{\mathbf{A}}(\mathbf{X}))$ is moved to $\lambda_i(\overline{\mathbf{A}}(\mathbf{X}+\Delta\mathbf{X}))$. If an eigenvalue of $\overline{\mathbf{A}}(\mathbf{X}+\Delta\mathbf{X})$ is outside the open unit disk, the closed-loop system, designed to be stable, becomes unstable with the finite-precision implemented $\mathbf{X}$. How easily the FWL error $\Delta\mathbf{X}$ can cause a stable control system to become unstable is determined by how close $\lambda_i(\overline{\mathbf{A}}(\mathbf{X}))$ are to the unit circle and how sensitive they are to the controller perturbations. The following FWL closed-loop stability measure [5] is considered in this study:

$$f(\mathbf{X}) \triangleq \min_{i\in\{1,\cdots,m+n\}} \frac{1-|\lambda_i|}{\sqrt{N}\left\|\frac{\partial\lambda_i}{\partial\mathbf{X}}\right\|_F} \tag{5}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, that is, for any complex-valued matrix $\mathbf{M}$,

$$\|\mathbf{M}\|_F \triangleq \sqrt{\mathrm{tr}(\mathbf{M}^H\mathbf{M})} \tag{6}$$

with ^{H} denoting the conjugate transpose operator.

The measure $f(\mathbf{X})$ describes the "robustness" of closed-loop stability to FWL controller perturbations. As different controller realizations $\mathbf{X}$ result in different values of $f(\mathbf{X})$, it is natural to search for "optimal" controller realizations that maximize the measure defined in (5). This leads to the following optimal FWL realization problem [5]:

$$\upsilon \triangleq \max_{\mathbf{X}\in\mathcal{S}_C} f(\mathbf{X}). \tag{7}$$

III. Single-pole FWL stability measure

Define the single-pole FWL stability function for the closed-loop eigenvalue λ_i as

$$g(\mathbf{X}, i) \triangleq \frac{1-|\lambda_i|}{\sqrt{N}\left\|\frac{\partial\lambda_i}{\partial\mathbf{X}}\right\|_F}. \tag{8}$$

The optimal FWL realization problem (7) can be written as

$$\upsilon = \max_{\mathbf{X}\in\mathcal{S}_C} \min_{i\in\{1,\cdots,m+n\}} g(\mathbf{X}, i). \tag{9}$$

Lemma 1: $\max_{\mathbf{X}\in\mathcal{S}_C} \min_{i\in\{1,\cdots,m+n\}} g(\mathbf{X}, i) \le \min_{i\in\{1,\cdots,m+n\}} \max_{\mathbf{X}\in\mathcal{S}_C} g(\mathbf{X}, i)$.

Lemma 1 is obvious. For the eigenvalue λ_i, define the single-pole FWL stability measure as

$$\rho_i \triangleq \max_{\mathbf{X}\in\mathcal{S}_C} g(\mathbf{X}, i) = \max_{\mathbf{X}\in\mathcal{S}_C} \frac{1-|\lambda_i|}{\sqrt{N}\left\|\frac{\partial\lambda_i}{\partial\mathbf{X}}\right\|_F}. \tag{10}$$

From Lemma 1, it can be seen that the minimum of all ρ_is is an upper bound of the optimal value v in (7). We now discuss how to attain the measure ρ_i for the pole λ_i, in other words, how to solve for the minimization problem of single-pole sensitivity $\min_{\mathbf{X}\in\mathcal{S}_C} \|\frac{\partial\lambda_i}{\partial\mathbf{X}}\|_F$, since

$$\rho_i = \frac{(1-|\lambda_i|)/\sqrt{N}}{\min_{\mathbf{X}\in\mathcal{S}_C} \|\frac{\partial\lambda_i}{\partial\mathbf{X}}\|_F}. \tag{11}$$

Denote $\mathbf{p}_i$ a right eigenvector of $\overline{\mathbf{A}}(\mathbf{X})$ corresponding to the eigenvalue λ_i and $\mathbf{y}_i$ the related reciprocal left eigenvector. The following lemma is due to [5].

Lemma 2: Let $\overline{\mathbf{A}}(\mathbf{X}) = \mathbf{M}_0 + \mathbf{M}_1\mathbf{X}\mathbf{M}_2$ given in (3) be diagonalisable. Then

$$\frac{\partial\lambda_i}{\partial\mathbf{X}} \triangleq \begin{bmatrix} \frac{\partial\lambda_i}{\partial x_1} & \cdots & \frac{\partial\lambda_i}{\partial x_{N-l-n+1}} \\ \vdots & \cdots & \vdots \\ \frac{\partial\lambda_i}{\partial x_{l+n}} & \cdots & \frac{\partial\lambda_i}{\partial x_N} \end{bmatrix} = \mathbf{M}_1^T\mathbf{y}_i^*\mathbf{p}_i^T\mathbf{M}_2^T \tag{12}$$

where * denotes the conjugate operation and ^{T} the transpose operator.

Combining Lemma 2 with the definition of $\|\cdot\|_F$ in (6) leads to

$$\left\|\frac{\partial\lambda_i}{\partial\mathbf{X}}\right\|_F = \|\mathbf{M}_1^T\mathbf{y}_i\|_F\,\|\mathbf{M}_2\mathbf{p}_i\|_F\,. \tag{13}$$

Let $\mathbf{p}_i \in \mathcal{C}^{m+n}$ and $\mathbf{y}_i \in \mathcal{C}^{m+n}$ be partitioned into

$$\mathbf{p}_i = \begin{bmatrix}\mathbf{p}_i(1)\\ \mathbf{p}_i(2)\end{bmatrix},\ \ \mathbf{y}_i = \begin{bmatrix}\mathbf{y}_i(1)\\ \mathbf{y}_i(2)\end{bmatrix} \tag{14}$$

where $\mathbf{p}_i(1), \mathbf{y}_i(1) \in \mathcal{C}^m$ and $\mathbf{p}_i(2), \mathbf{y}_i(2) \in \mathcal{C}^n$. For the initial controller realization

$$\mathbf{X}_0 \triangleq \begin{bmatrix}\mathbf{D}_C^0 & \mathbf{C}_C^0\\ \mathbf{B}_C^0 & \mathbf{A}_C^0\end{bmatrix}, \tag{15}$$

let

$$\mathbf{p}_{0i} = \begin{bmatrix}\mathbf{p}_{0i}(1)\\ \mathbf{p}_{0i}(2)\end{bmatrix},\ \ \mathbf{y}_{0i} = \begin{bmatrix}\mathbf{y}_{0i}(1)\\ \mathbf{y}_{0i}(2)\end{bmatrix} \tag{16}$$

be a right eigenvector and the related reciprocal left eigenvector of $\overline{\mathbf{A}}(\mathbf{X}_0)$ corresponding to the eigenvalue λ_i, respectively. The definition of $\mathcal{S}_C$ in (1) means that

$$\mathbf{X} \triangleq \mathbf{X}(\mathbf{T}) = \begin{bmatrix}\mathbf{I}_l & \mathbf{0}\\ \mathbf{0} & \mathbf{T}^{-1}\end{bmatrix}\mathbf{X}_0\begin{bmatrix}\mathbf{I}_q & \mathbf{0}\\ \mathbf{0} & \mathbf{T}\end{bmatrix}. \tag{17}$$

It can then be shown that

$$\overline{\mathbf{A}}(\mathbf{X}) = \begin{bmatrix}\mathbf{I}_m & \mathbf{0}\\ \mathbf{0} & \mathbf{T}^{-1}\end{bmatrix}\overline{\mathbf{A}}(\mathbf{X}_0)\begin{bmatrix}\mathbf{I}_m & \mathbf{0}\\ \mathbf{0} & \mathbf{T}\end{bmatrix} \tag{18}$$

which implies that

$$\begin{bmatrix}\mathbf{p}_i(1)\\ \mathbf{p}_i(2)\end{bmatrix} = \begin{bmatrix}\mathbf{I}_m & \mathbf{0}\\ \mathbf{0} & \mathbf{T}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{p}_{0i}(1)\\ \mathbf{p}_{0i}(2)\end{bmatrix}, \tag{19}$$

$$\begin{bmatrix}\mathbf{y}_i(1)\\ \mathbf{y}_i(2)\end{bmatrix} = \begin{bmatrix}\mathbf{I}_m & \mathbf{0}\\ \mathbf{0} & \mathbf{T}^T\end{bmatrix}\begin{bmatrix}\mathbf{y}_{0i}(1)\\ \mathbf{y}_{0i}(2)\end{bmatrix}. \tag{20}$$

Hence

$$\left\|\frac{\partial\lambda_i}{\partial\mathbf{X}}\right\|_F^2 = \left\|\begin{matrix}\mathbf{B}_P^T\mathbf{y}_{0i}(1)\\ \mathbf{T}^T\mathbf{y}_{0i}(2)\end{matrix}\right\|_F^2\left\|\begin{matrix}\mathbf{C}_P\mathbf{p}_{0i}(1)\\ \mathbf{T}^{-1}\mathbf{p}_{0i}(2)\end{matrix}\right\|_F^2 =$$

$$\|\mathbf{T}^{-1}\mathbf{p}_{0i}(2)\|_F^2\|\mathbf{T}^T\mathbf{y}_{0i}(2)\|_F^2 + \alpha_i^2\|\mathbf{T}^T\mathbf{y}_{0i}(2)\|_F^2 + \beta_i^2\|\mathbf{T}^{-1}\mathbf{p}_{0i}(2)\|_F^2 + \alpha_i^2\beta_i^2 \tag{21}$$

where $\alpha_i = \|\mathbf{C}_P\mathbf{p}_{0i}(1)\|_F$ and $\beta_i = \|\mathbf{B}_P^T\mathbf{y}_{0i}(1)\|_F$. In order to attain ρ_i, we need to minimize the function

$$\xi(\mathbf{T},\delta,\eta,\mathbf{q},\mathbf{z}) \triangleq \|\mathbf{T}^{-1}\mathbf{q}\|_F^2\|\mathbf{T}^T\mathbf{z}\|_F^2 + \delta^2\|\mathbf{T}^T\mathbf{z}\|_F^2 + \eta^2\|\mathbf{T}^{-1}\mathbf{q}\|_F^2 + \delta^2\eta^2 \tag{22}$$

where $\mathbf{T} \in \mathcal{R}^{n\times n}$ is nonsingular, $\delta, \eta \in \mathcal{R}$ are positive, and $\mathbf{q}, \mathbf{z} \in \mathcal{C}^n$ are nonzero vectors. Let $\mathbf{e}_i$ denote the ith coordinate vector. For different cases of $\mathbf{q}$ and $\mathbf{z}$, the following theorems give the results on minimizing $\xi(\mathbf{T},\delta,\eta,\mathbf{q},\mathbf{z})$.

Theorem 1: Given positive $\delta,\eta \in \mathcal{R}$, $\mathbf{q},\mathbf{z} \in \mathcal{R}^n$, and $\mathbf{z}^T\mathbf{q} \neq 0$, we have

$$\min_{\substack{\mathbf{T}\in\mathcal{R}^{n\times n}\\ \det\mathbf{T}\neq 0}} \xi(\mathbf{T},\delta,\eta,\mathbf{q},\mathbf{z}) = (|\mathbf{z}^T\mathbf{q}| + \delta\eta)^2, \tag{23}$$

and $\xi(\mathbf{T},\delta,\eta,\mathbf{q},\mathbf{z})$ attains the minimum if and only if

$$\mathbf{T} = \mathbf{Q}\begin{bmatrix}\sqrt{h} & \mathbf{0}\\ \frac{1}{\sqrt{h}}\mathbf{F} & \mathbf{\Omega}\end{bmatrix}\mathbf{V} \tag{24}$$

where the orthogonal matrix $\mathbf{Q}$ can be obtained from the QR factorization of $\mathbf{z}$, i.e.

$$\mathbf{z} = \mathbf{Q}\left[\gamma\ 0\cdots 0\right]^T \tag{25}$$

with a nonzero $\gamma \in \mathcal{R}$,

$$h = \frac{\eta}{\delta\gamma^2}|\mathbf{z}^T\mathbf{q}|, \tag{26}$$

$$\mathbf{F} = \mathrm{sign}(\mathbf{z}^T\mathbf{q})\frac{\eta}{\delta\gamma}\begin{bmatrix}\mathbf{e}_2^T\\ \vdots\\ \mathbf{e}_n^T\end{bmatrix}\mathbf{Q}^T\mathbf{q}, \tag{27}$$

$\mathbf{\Omega} \in \mathcal{R}^{(n-1)\times(n-1)}$ is an arbitrary nonsingular matrix, and $\mathbf{V} \in \mathcal{R}^{n\times n}$ is an arbitrary orthogonal matrix.

For any $\mathbf{y} \in \mathcal{C}^n$, define $\Upsilon(\mathbf{y}) \triangleq [\mathrm{Re}(\mathbf{y})\ \mathrm{Im}(\mathbf{y})]$, where $\mathrm{Re}(\mathbf{y})$ and $\mathrm{Im}(\mathbf{y})$ denotes the real and the imaginary parts of $\mathbf{y}$, respectively.

Theorem 2: Given positive $\delta,\eta \in \mathcal{R}$, $\mathbf{q},\mathbf{z} \in \mathcal{C}^n$ and $\det((\Upsilon(\mathbf{z}))^T\Upsilon(\mathbf{q})) > 0$, we have

$$\min_{\substack{\mathbf{T}\in\mathcal{R}^{n\times n}\\ \det\mathbf{T}\neq 0}} \xi(\mathbf{T},\delta,\eta,\mathbf{q},\mathbf{z}) = (|\mathbf{z}^H\mathbf{q}| + \delta\eta)^2, \tag{28}$$

and $\xi(\mathbf{T}, \delta, \eta, \mathbf{q}, \mathbf{z})$ achieves the minimum if and only if

$$\mathbf{T} = \mathbf{Q}\begin{bmatrix} \mathbf{H}^{1/2} & \mathbf{0} \\ \mathbf{F}(\mathbf{H}^{1/2})^{-T} & \mathbf{\Omega} \end{bmatrix}\mathbf{V} \tag{29}$$

where the orthogonal matrix $\mathbf{Q}$ can be obtained from the QR factorization of $\Upsilon(\mathbf{z})$, i.e.

$$\Upsilon(\mathbf{z}) = \mathbf{Q}\begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} \tag{30}$$

with nonzero $\gamma_{11}, \gamma_{22} \in \mathcal{R}$,

$$\mathbf{H} = \frac{\eta}{\delta}\begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \end{bmatrix}^{-T} (\Upsilon(\mathbf{z}))^T \Upsilon(\mathbf{q}) \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \times \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \end{bmatrix}^{-1}, \tag{31}$$

$$\mathbf{F} = \frac{\eta}{\delta}\begin{bmatrix} \mathbf{e}_3^T \\ \vdots \\ \mathbf{e}_n^T \end{bmatrix} \mathbf{Q}^T \Upsilon(\mathbf{q}) \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \end{bmatrix}^{-1}, \tag{32}$$

θ is the solution of

$$\begin{cases} \tan\theta = \frac{a_{21}-a_{12}}{a_{11}+a_{22}} \\ a_{11}\cos\theta - a_{12}\sin\theta > 0 \end{cases} \tag{33}$$

with

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = (\Upsilon(\mathbf{z}))^T \Upsilon(\mathbf{q}), \tag{34}$$

$\mathbf{\Omega} \in \mathcal{R}^{(n-2)\times(n-2)}$ is an arbitrary nonsingular matrix, and $\mathbf{V} \in \mathcal{R}^{n\times n}$ is an arbitrary orthogonal matrix.

Theorem 3: Given positive $\delta, \eta \in \mathcal{R}$, $\mathbf{q}, \mathbf{z} \in \mathcal{C}^n$ and $\det((\Upsilon(\mathbf{z}))^T \Upsilon(\mathbf{q})) < 0$, we have

$$\min_{\substack{\mathbf{T}\in\mathcal{R}^{n\times n} \\ \det\mathbf{T}\neq 0}} \xi(\mathbf{T}, \delta, \eta, \mathbf{q}, \mathbf{z}) = (|\mathbf{z}^T\mathbf{q}| + \delta\eta)^2, \tag{35}$$

and $\xi(\mathbf{T}, \delta, \eta, \mathbf{q}, \mathbf{z})$ achieves the minimum if and only if

$$\mathbf{T} = \mathbf{Q}\begin{bmatrix} \mathbf{H}^{1/2} & \mathbf{0} \\ \mathbf{F}(\mathbf{H}^{1/2})^{-T} & \mathbf{\Omega} \end{bmatrix}\mathbf{V} \tag{36}$$

where the orthogonal matrix $\mathbf{Q}$ can be obtained from the QR factorization of $\Upsilon(\mathbf{z}^*)$, i.e.

$$\Upsilon(\mathbf{z}^*) = \mathbf{Q}\begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} \tag{37}$$

with nonzero $\gamma_{11}, \gamma_{22} \in \mathcal{R}$,

$$\mathbf{H} = \frac{\eta}{\delta}\begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \end{bmatrix}^{-T} (\Upsilon(\mathbf{z}^*))^T \Upsilon(\mathbf{q}) \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \times \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \end{bmatrix}^{-1}, \tag{38}$$

$$\mathbf{F} = \frac{\eta}{\delta}\begin{bmatrix} \mathbf{e}_3^T \\ \vdots \\ \mathbf{e}_n^T \end{bmatrix} \mathbf{Q}^T \Upsilon(\mathbf{q}) \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \end{bmatrix}^{-1}, \tag{39}$$

θ is the solution of

$$\begin{cases} \tan\theta = \frac{a_{21}-a_{12}}{a_{11}+a_{22}} \\ a_{11}\cos\theta - a_{12}\sin\theta > 0 \end{cases} \tag{40}$$

with

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = (\Upsilon(\mathbf{z}^*))^T \Upsilon(\mathbf{q}), \tag{41}$$

$\mathbf{\Omega} \in \mathcal{R}^{(n-2)\times(n-2)}$ is an arbitrary nonsingular matrix, and $\mathbf{V} \in \mathcal{R}^{n\times n}$ is an arbitrary orthogonal matrix.

Detailed proofs of these theorems can be found in [11].

IV. Computing optimal controller realizations

In the previous section, the problem of the single-pole FWL stability measure ρ_i is solved and hence the minimum of all the single-pole measures can be found. Define

$$i_1 \triangleq \arg\min_{i\in\{1,\cdots,m+n\}} \rho_i. \tag{42}$$

It is straightforward to verify the following lemma.

Lemma 3: A controller realization $\mathbf{X}_*$ is a solution of the optimal realization problem (7) and

$$\rho_{i_1} = v \tag{43}$$

if and only if $\mathbf{X}_*$ meets the conditions

$$g(\mathbf{X}_*, i_1) = \rho_{i_1} \tag{44}$$

and

$$g(\mathbf{X}_*, i) \geq \rho_{i_1}, \ \forall i \in \{1, \cdots, m+n\} \setminus \{i_1\}. \tag{45}$$

Without the loss of generality, we will assume that λ_{i_1} is a complex-valued eigenvalue, $\lambda_{i_1+1} = \lambda_{i_1}^*$ and $\det((\Upsilon(\mathbf{y}_{0i_1}(2)))^T \Upsilon(\mathbf{p}_{0i_1}(2))) > 0$. From Theorem 2, all the transformation matrices achieving ρ_{i_1} form the set

$$\mathcal{T}_1 \triangleq \left\{ \mathbf{T} \,\middle|\, \mathbf{T} = \mathbf{Q}_1 \begin{bmatrix} \mathbf{H}_1^{1/2} & \mathbf{0} \\ \mathbf{F}_1(\mathbf{H}_1^{1/2})^{-T} & \mathbf{\Omega}_1 \end{bmatrix} \mathbf{V}_1 \right\} \tag{46}$$

where $\mathbf{Q}_1$, $\mathbf{H}_1$ and $\mathbf{F}_1$ are determined in Theorem 2, $\mathbf{\Omega}_1 \in \mathcal{R}^{(n-2)\times(n-2)}$ is an arbitrary nonsingular matrix and $\mathbf{V}_1 \in \mathcal{R}^{n\times n}$ is an arbitrary orthogonal matrix. Lemma 3 shows that if there exists $\mathbf{T}_* \in \mathcal{T}_1$ satisfying

$$g(\mathbf{X}(\mathbf{T}_*), i) \geq \rho_{i_1}, \ \forall i \in \{1, \cdots, m+n\} \setminus \{i_1, i_1+1\}, \tag{47}$$

then $\rho_{i_1} = \upsilon$ and $\mathbf{X}(\mathbf{T}_*)$ is an optimal controller realization for the optimization problem (7). Thus, searching in $\mathcal{T}_1$ for $\mathbf{T}_*$ which satisfy (47) is a method of solving for the optimal realization problem (7). We present an algorithm for computing $\mathbf{X}(\mathbf{T}_*)$ in this way.

By setting $\mathbf{V}_1 = \mathbf{I}_n$ in (46), we have

$$\mathbf{T}(\mathbf{\Omega}_1) = \mathbf{Q}_1 \begin{bmatrix} \mathbf{H}_1^{1/2} & \mathbf{0} \\ \mathbf{F}_1(\mathbf{H}_1^{1/2})^{-T} & \mathbf{\Omega}_1 \end{bmatrix}. \tag{48}$$

Notice that, $\forall i \in \{1, \cdots, m+n\}$, the single-pole FWL stability function $g(\mathbf{X}(\mathbf{T}(\mathbf{\Omega}_1)), i)$ is differentiable with respect to $\mathbf{\Omega}_1$. With the derivative, we know how to change $\mathbf{\Omega}_1$ so that $g(\mathbf{X}(\mathbf{T}(\mathbf{\Omega}_1)), i)$ increases. Hence, for those $g(\mathbf{X}(\mathbf{T}(\mathbf{\Omega}_1)), i) < \rho_{i_1}$, $\mathbf{\Omega}_1$ can appropriately be changed step by step until $g(\mathbf{X}(\mathbf{T}(\mathbf{\Omega}_1)), i) \geq \rho_{i_1}$. With the derivative, we also know the direction of change which will decrease $g(\mathbf{X}(\mathbf{T}(\mathbf{\Omega}_1)), i)$. Hence, by avoiding to change $\mathbf{\Omega}_1$ in some directions, those $g(\mathbf{X}(\mathbf{T}(\mathbf{\Omega}_1)), i) \geq \rho_{i_1}$ can be made to hold their values. The basic idea of the algorithm is to search for an optimal transformation matrix $\mathbf{T}_*$ through increasing those $g(\mathbf{X}, i)$ which are smaller than ρ_{i_1} while not decreasing those $g(\mathbf{X}, i)$ which are larger than or equal to ρ_{i_1}. The detailed algorithm is as follows.

Initialization: Arbitrarily select a nonsingular $\mathbf{\Omega}_1$ to obtain an initial realization $\mathbf{X}(\mathbf{T}(\mathbf{\Omega}_1))$, set σ to a proper positive number and τ a small positive number.
Step 1: Find out every elements of the set

$$\mathcal{S}_I(\mathbf{\Omega}_1) = \{i | \rho_{i_1} \leq g(\mathbf{X}(\mathbf{T}(\mathbf{\Omega}_1)), i) < \rho_{i_1} + \sigma, \quad i \neq i_1, i \neq i_1 + 1\}. \tag{49}$$

Step 2: Find out

$$i_* = \arg \min_{i \in \{1, \cdots, m+n\}} g(\mathbf{X}(\mathbf{T}(\mathbf{\Omega}_1)), i). \tag{50}$$

If $\min_{i \in \{1, \cdots, m+n\}} g(\mathbf{X}(\mathbf{T}(\mathbf{\Omega}_1)), i) = \rho_{i_1}$, $\mathbf{T}_* = \mathbf{T}(\mathbf{\Omega}_1)$ and terminate the algorithm.
Step 3: Choose $\mathbf{J} \in \mathcal{R}^{(n-2)\times(n-2)}$ such that
i) $\forall i \in \mathcal{S}_I(\mathbf{\Omega}_1)$, $g(\mathbf{X}(\mathbf{T}(\mathbf{\Omega}_1 + \tau\mathbf{J})), i)$ is not less than $g(\mathbf{X}(\mathbf{T}(\mathbf{\Omega}_1)), i)$.
ii) $g(\mathbf{X}(\mathbf{T}(\mathbf{\Omega}_1)), i_*)$ increases as fast as possible.
iii) $\|\mathbf{J}\|_F = 1$.
Step 4: $\mathbf{\Omega}_1 = \mathbf{\Omega}_1 + \tau\mathbf{J}$, and go to *Step 1*.

The key of this algorithm is how to obtain $\mathbf{J}$. Denote $\mathrm{Vec}(\cdot)$ the column stacking operator. With a small τ, condition i) means that

$$\left(\mathrm{Vec}\left(\frac{dg(\mathbf{X}(\mathbf{T}(\mathbf{\Omega}_1)), i)}{d\mathbf{\Omega}_1}\right)\right)^T \mathrm{Vec}(\mathbf{J}) \geq 0, \ \forall i \in \mathcal{S}_I(\mathbf{\Omega}_1). \tag{51}$$

Condition ii) requires to improve $g(\mathbf{X}(\mathbf{T}(\mathbf{\Omega}_1)), i_*)$ as fast as possible and, therefore, the best direction is $\frac{dg(\mathbf{X}(\mathbf{T}(\mathbf{\Omega}_1)), i_*)}{d\mathbf{\Omega}_1}$. Combining all the three conditions in *Step 3*, $\mathrm{Vec}(\mathbf{J})$ can be chosen as in the following inner loop:

Step 3-1: Initially let $\mathcal{S}_n$ be an empty set and $\mathbf{J}_t = \frac{dg(\mathbf{X}(\mathbf{T}(\mathbf{\Omega}_1)), i_*)}{d\mathbf{\Omega}_1} / \left\| \frac{dg(\mathbf{X}(\mathbf{T}(\mathbf{\Omega}_1)), i_*)}{d\mathbf{\Omega}_1} \right\|_F$.
Step 3-2: Find every elements of the set

$$\mathcal{S}_t = \{i \,| i \in \mathcal{S}_I(\mathbf{\Omega}_1), \quad \left(\mathrm{Vec}\left(\frac{dg(\mathbf{X}(\mathbf{T}(\mathbf{\Omega}_1)), i)}{d\mathbf{\Omega}_1}\right)\right)^T \mathrm{Vec}(\mathbf{J}_t) < 0\}. \tag{52}$$

If $\mathcal{S}_t$ is empty, $\mathbf{J} = \mathbf{J}_t$ and terminate the inner loop.
Step 3-3: Find the index in $\mathcal{S}_t$ with which the derivative direction has the largest "angle" with respect to $\frac{dg(\mathbf{X}(\mathbf{T}(\mathbf{\Omega}_1)), i_*)}{d\mathbf{\Omega}_1}$, i.e.

$$i_f = \arg \min_{i \in \mathcal{S}_t} \frac{\left(\mathrm{Vec}\left(\frac{dg(\mathbf{X}(\mathbf{T}(\mathbf{\Omega}_1)), i)}{d\mathbf{\Omega}_1}\right)\right)^T \mathrm{Vec}\left(\frac{dg(\mathbf{X}(\mathbf{T}(\mathbf{\Omega}_1)), i_*)}{d\mathbf{\Omega}_1}\right)}{\left\|\frac{dg(\mathbf{X}(\mathbf{T}(\mathbf{\Omega}_1)), i)}{d\mathbf{\Omega}_1}\right\|_F \left\|\frac{dg(\mathbf{X}(\mathbf{T}(\mathbf{\Omega}_1)), i_*)}{d\mathbf{\Omega}_1}\right\|_F} \tag{53}$$

and let $\mathcal{S}_n = \mathcal{S}_n \cup \{i_f\}$.
Step 3-4: Suppose that $\mathcal{S}_n$ contains r elements which are numbered by $k_1, \cdots, k_r$. Compute the matrix

$$\mathbf{E} = \begin{bmatrix} \left(\mathrm{Vec}\left(\frac{dg(\mathbf{X}(\mathbf{T}(\mathbf{\Omega}_1)), k_1)}{d\mathbf{\Omega}_1}\right)\right)^T \\ \vdots \\ \left(\mathrm{Vec}\left(\frac{dg(\mathbf{X}(\mathbf{T}(\mathbf{\Omega}_1)), k_r)}{d\mathbf{\Omega}_1}\right)\right)^T \end{bmatrix} \in \mathcal{R}^{r\times(n-2)^2}. \tag{54}$$

Step 3-5: Choose $\mathrm{Vec}(\mathbf{J}_t)$ as the orthogonal projection of $\mathrm{Vec}\left(\frac{dg(\mathbf{X}(\mathbf{T}(\mathbf{\Omega}_1)), i_*)}{d\mathbf{\Omega}_1}\right)$ onto the null space $\mathcal{N}(\mathbf{E})$ of $\mathbf{E}$. Suppose that $\mathcal{N}(\mathbf{E})$, of dimension t, has the bases $\{\mathbf{b}_1, \cdots, \mathbf{b}_t\}$. We can compute $\mathrm{Vec}(\mathbf{J}_t)$ as follows:

$$\left.\begin{array}{c} a_j = \mathbf{b}_j^T \mathrm{Vec}\left(\frac{dg(\mathbf{X}(\mathbf{T}(\mathbf{\Omega}_1)), i_*)}{d\mathbf{\Omega}_1}\right) \in \mathcal{R}, \\ \forall j \in \{1, \cdots, t\}, \\ \mathbf{v} = \sum_{j=1}^{t} a_j \mathbf{b}_j \in \mathcal{R}^{(n-2)^2}, \\ \mathrm{Vec}(\mathbf{J}_t) = \frac{\mathbf{v}}{\sqrt{\mathbf{v}^T\mathbf{v}}} \in \mathcal{R}^{(n-2)^2}. \end{array}\right\} \tag{55}$$

Step 3-6: Go to *Step 3-2*.

In this algorithm, the derivatives $\frac{dg(\mathbf{X}(\mathbf{T}(\mathbf{\Omega}_1)), i)}{d\mathbf{\Omega}_1}$ are needed. Based on (3), (8), (19), (20), (48) and Lemma 2, these derivatives can be computed easily.

V. A DESIGN EXAMPLE

The example considered in [5] was used to illustrate the effectiveness of the proposed analytical design approach. This example had 5 pairs of complex-valued conjugate closed-loop eigenvalues $\lambda_{1,2}$, $\lambda_{3,4}$, $\lambda_{5,6}$, $\lambda_{7,8}$ and $\lambda_{9,10}$ along with one real-valued eigenvalue λ_{11}. Using the method described in Section III, we attained the minimum of all the single-pole measures $\rho_{i_1} = 6.7344e - 6$ with $i_1 = 5$ together with the corresponding $\mathbf{Q}_1$, $\mathbf{H}_1$ and $\mathbf{F}_1$. The algorithm of Section IV was then applied with $\tau = 0.1$ and the initial $\mathbf{\Omega}_1 = \mathbf{I}_4$ to find a global optimal realization. Fig. 1 illustrates the changes of all the single-pole FWL stability functions in each iteration stage. It can be seen that at the 37th stage, the global optimal controller realization $\mathbf{X}_{\rm opt}$ was found, since at this stage the conditions of Lemma 3 were met and the algorithm terminated. Table I summarizes the values of the FWL closed-loop stability measure for $\mathbf{X}_0$ and $\mathbf{X}_{\rm opt}$. It can be seen that the optimal controller realizations improve the FWL closed-loop stability measure by a factor of 2×10^5 over the initially designed controller realization.

VI. CONCLUSIONS

We have developed an analytic approach to solve for the optimal controller realization problem based on an FWL closed-loop stability measure, which avoids the drawbacks usually associated with using numerical optimization methods to tackle this problem. For each closed-loop eigenvalue, a single-pole stability measure has been defined, and an analytical method has been derived to compute all the realizations which achieve the single-pole measure. It has been shown that the minimum of all the single-pole measures is an upper bound of the optimal value of the optimal FWL realization problem. The necessary and sufficient conditions have been given for a realization which attains the minimum single-pole measure to be a global solution of the optimal realization problem. An algorithm have been presented to compute global optimal realizations.

ACKNOWLEDGEMENTS

J. Wu and S. Chen wish to thank the support of the U.K. Royal Society under a KC Wong fellowship (RL/ART/CN/XFI/KCW/11949). J. Wu and J. Chu wish to thank the support of the National Natural Science Foundation of China under Grant 60174026.

realization $\mathbf{X}$	FWL stability measure $f(\mathbf{X})$
$\mathbf{X}_0$	$3.1797e - 11$
$\mathbf{X}_{\rm opt}$	$6.7344e - 6$

TABLE I

COMPARISON OF THE FWL CLOSED-LOOP STABILITY MEASURES FOR CONTROLLER REALIZATIONS $\mathbf{X}_0$ AND $\mathbf{X}_{\rm opt}$.

REFERENCES

[1] L.H. Keel and S.P. Bhattacharryya, "Robust, fragile, or optimal?" *IEEE Trans. Automatic Control*, Vol.42, No.8, pp.1098–1105, 1997.

[2] M. Gevers and G. Li, *Parameterizations in Control, Estimation and Filtering Problems: Accuracy Aspects*. London: Springer Verlag, 1993.

[3] R.S.H. Istepanian and J.F. Whidborne, eds., *Digital Controller Implementation and Fragility: A Modern Perspective*. London: Springer Verlag, 2001.

[4] I.J. Fialho and T.T. Georgiou, "On stability and performance of sampled-data systems subject to wordlength constraint," *IEEE Trans. Automatic Control*, Vol.39, No.12, pp.2476–2481, 1994.

[5] G. Li, "On the structure of digital controllers with finite word length consideration," *IEEE Trans. Automatic Control*, Vol.43, No.5, pp.689–693, 1998.

[6] J.F. Whidborne, J. Wu and R.S.H. Istepanian, "Finite word length stability issues in an l_1 framework," *Int. J. Control*, Vol.73, No.2, pp.166–176, 2000.

[7] J. Wu, S. Chen, G. Li, R.S.H. Istepanian and J. Chu, "An improved closed-loop stability related measure for finite-precision digital controller realizations," *IEEE Trans. Automatic Control*, Vol.46, No.7, pp.1162–1166, 2001.

[8] R.S.H. Istepanian, J. Wu and J.F. Whidborne, "Controller realizations of a teleoperated dual-wrist assembly system with finite word length considerations," *IEEE Trans. Control Systems Technology*, Vol.9, No.4, pp.624–628, 2001.

[9] J.F. Whidborne, R.S.H. Istepanian and J. Wu, "Reduction of controller fragility by pole sensitivity minimization," *IEEE Trans. Automatic Control*, Vol.46, No.2, pp.320–325, 2001.

[10] J.F. Whidborne, J. Wu, R.S.H. Istepanian and J. Chu, "Comments on 'On the structure of digital controllers with finite word length consideration'," *IEEE Trans. Automatic Control*, Vol.45, No.2, pp.344–344, 2000.

[11] J. Wu, S. Chen, G. Li and J. Chu, "Analysis of global solutions to an optimal finite precision digital controller realization problem," submitted to *IEEE Trans. Automatic Control*, 2001.

[12] S. Boyd, L.El Ghaoui, E. Feron and V. Balakrishnan, *Linear Matrix Inequalities in Systems and Control Theory*. Philadelphia, PA: SIAM, 1994.

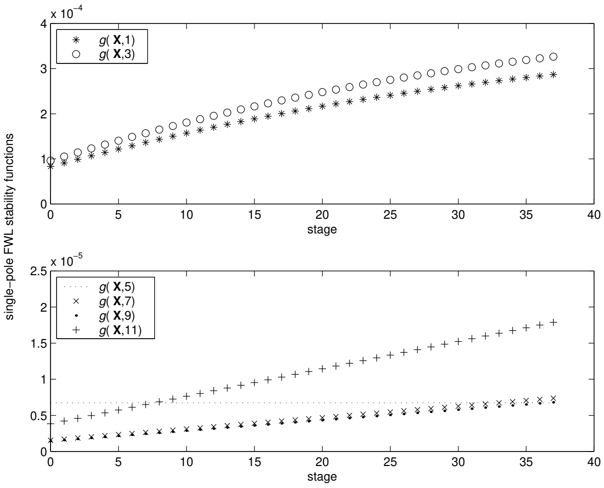

Fig. 1. Single-pole FWL stability functions in each iteration stage of the algorithm.

(For the benefits of review, we give the proof of Theorem 2. The proofs of the other two theorems are similar.)

APPENDIX PROOF OF THEOREM 2

Lemma 4: (See [12]). Let real-valued matrices $\mathbf{M}_{22}, \mathbf{M}_{21}$ and $\mathbf{M}_{11} > 0$ be given with appropriate dimensions. Then

$$\begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{21}^T \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{bmatrix} > 0 \tag{56}$$

if and only if

$$\mathbf{M}_{22} - \mathbf{M}_{21}\mathbf{M}_{11}^{-1}\mathbf{M}_{21}^T > 0. \tag{57}$$

Lemma 5: Given positive $\delta, \eta \in \mathcal{R}$, $\mathbf{q}, \mathbf{z} \in \mathcal{C}^n$, and for any nonsingular $\mathbf{T} \in \mathcal{R}^{n\times n}$, we have

$$\xi(\mathbf{T}, \delta, \eta, \mathbf{q}, \mathbf{z}) \geq (|\mathbf{z}^H\mathbf{q}| + \delta\eta)^2. \tag{58}$$

The equality occurs if and only if there exist $\mathbf{W} \in \mathcal{R}^{n\times n}$, $\mathbf{W} > 0$ and non-negative $\theta \in \mathcal{R}$ satisfying the following condition:

$$\mathbf{W}\mathbf{z} = (\cos\theta + j\sin\theta)\frac{\eta}{\delta}\mathbf{q}. \tag{59}$$

When the above equation has solutions, the equality in (58) occurs only at the transformation matrix

$$\mathbf{T} = \mathbf{W}^{1/2}\mathbf{V} \tag{60}$$

where $\mathbf{V} \in \mathcal{R}^{n\times n}$ is an arbitrary orthogonal matrix.

Proof: First of all,

$$\begin{aligned} &\|\mathbf{T}^{-1}\mathbf{q}\|_F^2\|\mathbf{T}^T\mathbf{z}\|_F^2 + \delta^2\|\mathbf{T}^T\mathbf{z}\|_F^2 + \eta^2\|\mathbf{T}^{-1}\mathbf{q}\|_F^2 + \delta^2\eta^2 \geq \\ &\|\mathbf{T}^{-1}\mathbf{q}\|_F^2\|\mathbf{T}^T\mathbf{z}\|_F^2 + 2\delta\eta\|\mathbf{T}^T\mathbf{z}\|_F\|\mathbf{T}^{-1}\mathbf{q}\|_F + \delta^2\eta^2 \geq \\ &(\|\mathbf{T}^{-1}\mathbf{q}\|_F\|\mathbf{T}^T\mathbf{z}\|_F + \delta\eta)^2. \end{aligned} \tag{61}$$

The equality holds if and only if

$$\delta\|\mathbf{T}^T\mathbf{z}\|_F = \eta\|\mathbf{T}^{-1}\mathbf{q}\|_F. \tag{62}$$

Using the Cauchy-Schwartz inequality, we have

$$\begin{aligned} (\|\mathbf{T}^{-1}\mathbf{q}\|_F\|\mathbf{T}^T\mathbf{z}\|_F + \delta\eta)^2 &\geq (\|(\mathbf{T}^T\mathbf{z})^H\mathbf{T}^{-1}\mathbf{q}\|_F + \delta\eta)^2 \\ &\geq (|\mathbf{z}^H\mathbf{q}| + \delta\eta)^2. \end{aligned} \tag{63}$$

The equality holds if and only if

$$\mathbf{T}^T\mathbf{z} = c\mathbf{T}^{-1}\mathbf{q} \tag{64}$$

for some complex number c.

To achieve (61) and (63) with equality, one needs to satisfy both the conditions (62) and (64). This implies that $c = (\cos\theta + j\sin\theta)\frac{\eta}{\delta}$ and $0 \leq \theta \in \mathcal{R}$. Thus,

$$\mathbf{T}^T\mathbf{z} = (\cos\theta + j\sin\theta)\frac{\eta}{\delta}\mathbf{T}^{-1}\mathbf{q}. \tag{65}$$

As $\mathbf{T}$ is nonsingular, equality (65) is equivalent to

$$\mathbf{W}\mathbf{z} = (\cos\theta + j\sin\theta)\frac{\eta}{\delta}\mathbf{q} \tag{66}$$

with $\mathbf{W} > 0$ and $\mathbf{T} = \mathbf{W}^{1/2}\mathbf{V}$.

Comments: With the map $\Upsilon(\mathbf{y}) = [\mathrm{Re}(\mathbf{y})\ \mathrm{Im}(\mathbf{y})]$, conditions (59) can be viewed as

$$\mathbf{W}\Upsilon(\mathbf{z}) = \frac{\eta}{\delta}\Upsilon(\mathbf{q})\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}. \tag{67}$$

Lemma 6: Given positive $\delta, \eta \in \mathcal{R}$, $\mathbf{q}, \mathbf{z} \in \mathcal{C}^n$ and $\mathrm{rank}(\Upsilon(\mathbf{z})) = 2$, equation (59) has solutions if and only if $\det((\Upsilon(\mathbf{z}))^T\Upsilon(\mathbf{q})) > 0$. Moreover, any solution to equation (59) can be expressed as

$$\tan\theta = \frac{a_{21} - a_{12}}{a_{11} + a_{22}}, \tag{68}$$

$$a_{11}\cos\theta - a_{12}\sin\theta > 0, \tag{69}$$

$$\mathbf{W} = \mathbf{Q}\begin{bmatrix} \mathbf{H} & \mathbf{F}^T \\ \mathbf{F} & \mathbf{G} \end{bmatrix}\mathbf{Q}^T, \tag{70}$$

where a_{11}, a_{12}, a_{21}, a_{22}, $\mathbf{Q}$, $\mathbf{H}$ and $\mathbf{F}$ are as determined in Theorem 2, and

$$\mathbf{G} = \mathbf{F}\mathbf{H}^{-1}\mathbf{F}^T + \mathbf{U} \tag{71}$$

with $\mathbf{U} \in \mathcal{R}^{(n-2)\times(n-2)}$ being an arbitrary positive definite matrix.

Proof: If $\det((\Upsilon(\mathbf{z}))^T\Upsilon(\mathbf{q})) > 0$, it is easy to verify that $\mathbf{W}$ and θ given by (68)–(70) are a solution to equation (59). If on the other hand equation (59) has a solution $\mathbf{W}$ and θ, $\mathbf{W}$ and θ also satisfies condition (67). From (67), we have

$$(\Upsilon(\mathbf{z}))^T\mathbf{W}\Upsilon(\mathbf{z}) = \frac{\eta}{\delta}(\Upsilon(\mathbf{z}))^T\Upsilon(\mathbf{q})\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}. \tag{72}$$

On account of $(\Upsilon(\mathbf{z}))^T\mathbf{W}\Upsilon(\mathbf{z}) > 0$, it can be seen that

$$(\Upsilon(\mathbf{z}))^T\Upsilon(\mathbf{q})\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} > 0. \tag{73}$$

A necessary condition to satisfy (73) is that

$$\det\left((\Upsilon(\mathbf{z}))^T\Upsilon(\mathbf{q})\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}\right) > 0. \tag{74}$$

Since

$$\begin{aligned} &\det\left((\Upsilon(\mathbf{z}))^T\Upsilon(\mathbf{q})\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}\right) = \\ &\det((\Upsilon(\mathbf{z}))^T\Upsilon(\mathbf{q}))\det\left(\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}\right) \\ &= \det((\Upsilon(\mathbf{z}))^T\Upsilon(\mathbf{q})), \end{aligned} \tag{75}$$

the condition (74) becomes

$$\det((\Upsilon(\mathbf{z}))^T \Upsilon(\mathbf{q})) > 0. \tag{76}$$

This completes the proof of the first part of Lemma 6.

Now, (73) holds if and only if all of the following three conditions are satisfied

$$a_{21} \cos\theta - a_{22} \sin\theta = a_{11} \sin\theta + a_{12} \cos\theta, \tag{77}$$

$$a_{11} \cos\theta - a_{12} \sin\theta > 0, \tag{78}$$

$$\det((\Upsilon(\mathbf{z}))^T \Upsilon(\mathbf{q})) > 0. \tag{79}$$

From (77), we directly obtain

$$\tan\theta = \frac{a_{21} - a_{12}}{a_{11} + a_{22}}. \tag{80}$$

Denote

$$\mathbf{S} = \mathbf{Q}^T \mathbf{W} \mathbf{Q}. \tag{81}$$

Then

$$\mathbf{S}\,[\,\mathbf{e}_1 \quad \mathbf{e}_2\,] \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \end{bmatrix} = \mathbf{S} \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix}$$

$$= \frac{\eta}{\delta} \mathbf{Q}^T \Upsilon(\mathbf{q}) \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \tag{82}$$

and $\mathbf{S} > 0$. Partition $\mathbf{S}$ into

$$\mathbf{S} = \begin{bmatrix} \mathbf{H} & \mathbf{F}^T \\ \mathbf{F} & \mathbf{G} \end{bmatrix} \tag{83}$$

where $\mathbf{H} \in \mathcal{R}^{2\times 2}$, $\mathbf{F} \in \mathcal{R}^{(n-2)\times 2}$ and $\mathbf{G} \in \mathcal{R}^{(n-2)\times(n-2)}$. Then from (82) and noticing

$$(\Upsilon(\mathbf{z}))^T = \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \end{bmatrix}^T [\,\mathbf{e}_1 \quad \mathbf{e}_2\,]^T \mathbf{Q}^T, \tag{84}$$

we have

$$\mathbf{H} = \begin{bmatrix} \mathbf{e}_1^T \\ \mathbf{e}_2^T \end{bmatrix} \mathbf{S}\,[\,\mathbf{e}_1 \quad \mathbf{e}_2\,] = \frac{\eta}{\delta} \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \end{bmatrix}^{-T} \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \end{bmatrix}^T$$

$$\times [\,\mathbf{e}_1 \quad \mathbf{e}_2\,]^T \mathbf{Q}^T \Upsilon(\mathbf{q}) \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \end{bmatrix}^{-1}$$

$$= \frac{\eta}{\delta} \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \end{bmatrix}^{-T} (\Upsilon(\mathbf{z}))^T \Upsilon(\mathbf{q}) \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

$$\times \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \end{bmatrix}^{-1}, \tag{85}$$

$$\mathbf{F} = \begin{bmatrix} \mathbf{e}_3^T \\ \vdots \\ \mathbf{e}_n^T \end{bmatrix} \mathbf{S}\,[\,\mathbf{e}_1 \quad \mathbf{e}_2\,] = \frac{\eta}{\delta} \begin{bmatrix} \mathbf{e}_3^T \\ \vdots \\ \mathbf{e}_n^T \end{bmatrix} \mathbf{Q}^T \Upsilon(\mathbf{q})$$

$$\times \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \gamma_{11} & \gamma_{12} \\ 0 & \gamma_{22} \end{bmatrix}^{-1}. \tag{86}$$

From Lemma 4, it is known that

$$\mathbf{G} = \mathbf{F}\mathbf{H}^{-1}\mathbf{F}^T + \mathbf{U} \tag{87}$$

where $\mathbf{U} \in \mathcal{R}^{(n-2)\times(n-2)}$ is an arbitrary positive definite matrix. This completes the proof of Lemma 6.

Combining Lemmas 5 and 6 directly leads to Theorem 2.

Comments: It should be pointed out that there always exists θ satisfying (68) and (69) in Lemma 6. Firstly, both the constraint (77) and the constraint

$$a_{11} \cos\theta - a_{12} \sin\theta = 0 \tag{88}$$

can not be met simultaneously by any $0 \leq \theta \in \mathcal{R}$. In fact, combining (77) and (88) forms

$$\begin{bmatrix} a_{11} + a_{22} & a_{12} - a_{21} \\ -a_{12} & a_{11} \end{bmatrix} \begin{bmatrix} \sin\theta \\ \cos\theta \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{89}$$

We observe that

$$\begin{vmatrix} a_{11} + a_{22} & a_{12} - a_{21} \\ -a_{12} & a_{11} \end{vmatrix} = \begin{vmatrix} a_{11} & a_{12} \\ -a_{12} & a_{11} \end{vmatrix} + \begin{vmatrix} a_{22} & -a_{21} \\ -a_{12} & a_{11} \end{vmatrix}$$

$$= a_{11}^2 + a_{12}^2 + \det((\Upsilon(\mathbf{z}))^T \Upsilon(\mathbf{q})) > 0. \tag{90}$$

This leads to the impossible situation $\sin\theta = \cos\theta = 0$. Thus, for any θ satisfying (77), there will be $a_{11} \cos\theta - a_{12} \sin\theta \neq 0$. Secondly, after we have determined θ by (68) or (77), in the case that $a_{11} \cos\theta - a_{12} \sin\theta < 0$, we can alter θ to $\theta + \pi$ which also satisfies (68) and $a_{11} \cos(\theta + \pi) - a_{12} \sin(\theta + \pi) > 0$. Therefore, θ can always be attained by (68) and (69).