Roundoff Noise Minimization
for 2-D Separable-Denominator Digital Filters
Using Jointly Optimal High-Order Error Feedback and Realization

Takao Hinamoto	Akimitsu Doi	Wu-Sheng Lu
Hiroshima University	Hiroshima Institute of Technology	University of Victoria
Higashi-Hiroshima, Japan	Hiroshima, Japan	Victoria, Canada

May 31, 2017

Outline

- Early Work and Objectives
- Model
- Noise Gain and Its Minimization
- Numerical Example

1. Early Work and Objectives

Error Feedback (EF)

- Higgins and Munson, 1984.
- Vaidyanathan, 1985.
- Laakso and Hartimo, 1992.

EF Combined with State-Space Realization

- Hinamoto, Doi & Lu, 2003, 2005, 2013.

The Problem of Studies

Jointly optimizing *high-order* EF and realization for minimizing roundoff noise subject to l_2 -scaling constraints for 2-D separable denominator digital filters.

◊ Closed-form solutions for the controllability and observability Grammians;

◊ EF order is now made to be independent from the dimension of the state.

2. Model

$$\boldsymbol{x}_1(i, j) = \boldsymbol{A}\boldsymbol{x}_0(i, j) + \boldsymbol{b}u(i, j)$$
$$y(i, j) = c\boldsymbol{x}_0(i, j) + du(i, j)$$

where the filter's separable denominator means that

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_2 \\ \boldsymbol{0} & \boldsymbol{A}_4 \end{bmatrix}$$

- Taking error feedback and finite-word-length (FWL) effect into account, the model becomes

$$\tilde{\boldsymbol{x}}_1(i, j) = \boldsymbol{A}Q[\tilde{\boldsymbol{x}}_0(i, j)] + \boldsymbol{b}u(i, j)$$
$$+\sum\nolimits_{k=0}^{M-1}[\boldsymbol{D}_{1k} \oplus \boldsymbol{0}]\boldsymbol{e}_{-k}(i, j) + \sum\nolimits_{l=0}^{N-1}[\boldsymbol{0} \oplus \boldsymbol{D}_{4l}]\boldsymbol{e}_{-l}(i, j)$$
$$\tilde{y}(i, j) = \boldsymbol{c}Q[\tilde{\boldsymbol{x}}_0(i, j)] + du(i, j) + \boldsymbol{h}\boldsymbol{e}_0(i, j)$$

where $\boldsymbol{h} \in R^{1\times(m+n)}$ is an error-feedforward vector, $\boldsymbol{D}_{1k} \in R^{m\times m}$ and $\boldsymbol{D}_{4l} \in R^{n\times n}$ are high-order EF matrices, and

$$\boldsymbol{e}_k(i, j) = \tilde{\boldsymbol{x}}_k(i, j) - Q[\tilde{\boldsymbol{x}}_k(i, j)]$$

- From above equations, we have

$$\Delta \boldsymbol{x}_1(i,j) = \boldsymbol{A}\Delta \boldsymbol{x}_0(i,j) + \boldsymbol{A}\boldsymbol{e}_0(i,j) - \sum\nolimits_{k=0}^{M-1}[\boldsymbol{D}_{1k} \oplus \boldsymbol{0}]\boldsymbol{e}_{-k}(i,j) - \sum\nolimits_{l=0}^{N-1}[\boldsymbol{0} \oplus \boldsymbol{D}_{4l}]\boldsymbol{e}_{-l}(i,j)$$

$$\Delta y(i,j) = \boldsymbol{c}\Delta \boldsymbol{x}_0(i,j)] + (\boldsymbol{c} - \boldsymbol{h})\boldsymbol{e}_0(i,j)$$

where $\Delta \boldsymbol{x}_k(i,j) = \boldsymbol{x}_k(i,j) - \tilde{\boldsymbol{x}}_k(i,j)$ and $\Delta y(i,j) = y(i,j) - \tilde{y}(i,j)$.

- Based on this, we obtain the frequency-domain model

$$\Delta Y(z_1,z_2) = \left[\boldsymbol{H}_e^h(z_1) \quad \boldsymbol{H}_e^v(z_1,z_2) \right] \boldsymbol{E}_0(z_1,z_2)$$

where

$$\boldsymbol{H}_e^h(z_1) = \boldsymbol{c}_1 \sum_{i=0}^{\infty} \left(\boldsymbol{A}_1^i - \sum_{k=0}^{M-1} \boldsymbol{A}_1^{i-k-1} \boldsymbol{D}_{1k} \right) z_1^{-i} - \boldsymbol{h}_1$$

$$\boldsymbol{H}_e^v(z_1,z_2) = \left[\boldsymbol{c}_2 + \boldsymbol{c}_1 (z_1 \boldsymbol{I}_m - \boldsymbol{A}_1)^{-1} \boldsymbol{A}_2 \right] \sum_{i=0}^{\infty} \left(\boldsymbol{A}_4^i - \sum_{k=0}^{N-1} \boldsymbol{A}_4^{i-k-1} \boldsymbol{D}_{4k} \right) z_2^{-i} - \boldsymbol{h}_2$$

3. Noise Gain and Its Minimization

3.A Noise Gain

- Based on above analysis, the ***normalized noise gain*** which is a scaled noise variance at the filter output is found to be

$$J_0(\boldsymbol{h}, \boldsymbol{D}_1, \boldsymbol{D}_4) = J_0^h(\boldsymbol{h}_1, \boldsymbol{D}_1) + J_0^v(\boldsymbol{h}_2, \boldsymbol{D}_4)$$

where

$$J_0^h(\boldsymbol{h}_1, \boldsymbol{D}_1) = \text{trace}\left[\frac{1}{2\pi j}\oint_{|z_1|=1} \boldsymbol{H}_e^h(z_1)^* \boldsymbol{H}_e^h(z_1)\frac{dz_1}{z_1}\right]$$

$$J_0^v(\boldsymbol{h}_2, \boldsymbol{D}_4) = \text{trace}\left[\frac{1}{(2\pi j)^2}\oint_{|z_1|=1}\oint_{|z_2|=1} \boldsymbol{H}_e^v(z_1, z_2)^* \boldsymbol{H}_e^v(z_1, z_2)\frac{dz_1 dz_2}{z_1 z_2}\right]$$

with

$$\boldsymbol{D}_1 = \begin{bmatrix} \boldsymbol{D}_{10} & \boldsymbol{D}_{11} & \cdots & \boldsymbol{D}_{1,M-1} \end{bmatrix}, \quad \boldsymbol{D}_4 = \begin{bmatrix} \boldsymbol{D}_{40} & \boldsymbol{D}_{41} & \cdots & \boldsymbol{D}_{4,N-1} \end{bmatrix}$$

- Utilizing the expression of $\boldsymbol{H}_e^h(z_1)$ and $\boldsymbol{H}_e^v(z_1, z_2)$ in terms of the state-space model parameters, the noise gain can be expressed explicitly as follows:

$$J_0^h(\boldsymbol{h}_1, \boldsymbol{D}_1) = \text{trace}\Bigg[\boldsymbol{W}^h - \sum_{k=0}^{M-1}\left\{\boldsymbol{D}_{1k}^T\boldsymbol{W}^h\boldsymbol{A}_1^{k+1} + (\boldsymbol{A}_1^T)^{k+1}\boldsymbol{W}^h\boldsymbol{D}_{1k}\right\}$$
$$+\sum_{k=0}^{M-1}\sum_{l=0}^{M-1}\boldsymbol{D}_{1k}^T\left\{\boldsymbol{W}^h\boldsymbol{A}_1^{k-l} + (\boldsymbol{A}_1^T)^{l-k}\boldsymbol{W}^h\right\}\boldsymbol{D}_{1l}$$
$$-\sum_{k=0}^{M-1}\boldsymbol{D}_{1k}^T\boldsymbol{W}^h\boldsymbol{D}_{1k} - 2\boldsymbol{h}_1^T\boldsymbol{c}_1 + \boldsymbol{h}_1^T\boldsymbol{h}_1\Bigg]$$

and

$$J_0^v(\boldsymbol{h}_2, \boldsymbol{D}_4) = \text{trace}\Bigg[\boldsymbol{W}^v - \sum_{k=0}^{N-1}\left\{\boldsymbol{D}_{4k}^T\boldsymbol{W}^v\boldsymbol{A}_4^{k+1} + (\boldsymbol{A}_4^T)^{k+1}\boldsymbol{W}^v\boldsymbol{D}_{4k}\right\}$$
$$+\sum_{k=0}^{N-1}\sum_{l=0}^{N-1}\boldsymbol{D}_{4k}^T\left\{\boldsymbol{W}^v\boldsymbol{A}_4^{k-l} + (\boldsymbol{A}_4^T)^{l-k}\boldsymbol{W}^v\right\}\boldsymbol{D}_{4l}$$
$$-\sum_{k=0}^{N-1}\boldsymbol{D}_{4k}^T\boldsymbol{W}^v\boldsymbol{D}_{4k} - 2\boldsymbol{h}_2^T\boldsymbol{c}_2 + \boldsymbol{h}_2^T\boldsymbol{h}_2\Bigg]$$

where $\boldsymbol{W}^h$ and $\boldsymbol{W}^v$ are the horizontal and vertical observability Grammians of the state-space filter which can be found by solving the Lyapunov equations:

$$\boldsymbol{W}^h = \boldsymbol{A}_1^T \boldsymbol{W}^h \boldsymbol{A}_1 + \boldsymbol{c}_1^T \boldsymbol{c}_1$$

and

$$\boldsymbol{W}^v = \boldsymbol{A}_4^T \boldsymbol{W}^v \boldsymbol{A}_4 + \boldsymbol{A}_2^T \boldsymbol{W}^h \boldsymbol{A}_2 + \boldsymbol{c}_2^T \boldsymbol{c}_2$$

- The noise gain can be considerably simplified if the element EF matrices in $\boldsymbol{D}_1$ and $\boldsymbol{D}_4$ are diagonal:

$$J^h(\boldsymbol{h}_1, \boldsymbol{D}_1) = \mathrm{tr}\left[\boldsymbol{W}^h - \boldsymbol{c}_1^T \boldsymbol{c}_1 - 2\sum_{k=0}^{M-1} \boldsymbol{W}^h \boldsymbol{A}_1^{k+1} \boldsymbol{D}_{1k} + \sum_{k=0}^{M-1}\sum_{l=0}^{M-1} \boldsymbol{W}^h \boldsymbol{A}_1^{|k-l|} \boldsymbol{D}_{1k} \boldsymbol{D}_{1l}\right] + (\boldsymbol{c}_1 - \boldsymbol{h}_1)(\boldsymbol{c}_1 - \boldsymbol{h}_1)^T$$

and

$$J^v(\boldsymbol{h}_2, \boldsymbol{D}_4) = \mathrm{tr}\left[\boldsymbol{W}^v - \boldsymbol{c}_2^T \boldsymbol{c}_2 - 2\sum_{k=0}^{N-1} \boldsymbol{W}^v \boldsymbol{A}_4^{k+1} \boldsymbol{D}_{4k} + \sum_{k=0}^{N-1}\sum_{l=0}^{N-1} \boldsymbol{W}^v \boldsymbol{A}_4^{|k-l|} \boldsymbol{D}_{4k} \boldsymbol{D}_{4l}\right] + (\boldsymbol{c}_2 - \boldsymbol{h}_2)(\boldsymbol{c}_2 - \boldsymbol{h}_2)^T$$

3.B Minimization of Noise Gain

We now seek to find a state-space realization of a given state-space filter, $(\bar{\boldsymbol{A}}, \bar{\boldsymbol{b}}, \bar{\boldsymbol{c}}, d)$ where

$$\bar{\boldsymbol{A}} = \boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T},\ \bar{\boldsymbol{b}} = \boldsymbol{T}^{-1}\boldsymbol{b},\ \bar{\boldsymbol{c}} = \boldsymbol{c}\boldsymbol{T}$$

with $\boldsymbol{T} = \boldsymbol{T}_1 \oplus \boldsymbol{T}_4$, such that its noise gain is minimized subject to the l_2-scaling constraints

$$(\bar{\boldsymbol{K}})_{ii} = (\boldsymbol{T}^{-1}\boldsymbol{K}\boldsymbol{T})_{ii} = 1 \text{ for } 1 \le i \le m+n$$

where $\boldsymbol{K} = \boldsymbol{K}^h \oplus \boldsymbol{K}^v$ is the controllability Grammians which can be found by solving the Laypunov equations

$$\boldsymbol{K}^v = \boldsymbol{A}_4\boldsymbol{K}^v\boldsymbol{A}_4^T + \boldsymbol{b}_2\boldsymbol{b}_2^T$$

and

$$\boldsymbol{K}^h = \boldsymbol{A}_1\boldsymbol{K}^h\boldsymbol{A}_1^T + \boldsymbol{A}_2\boldsymbol{K}^v\boldsymbol{A}_2^T + \boldsymbol{b}_1\boldsymbol{b}_1^T$$

- Evidently, the variables involved in this problem are coordinate transformation matrix $\boldsymbol{T}$ as well as diagonal EF matrices in $\boldsymbol{D}_1$ and $\boldsymbol{D}_4$ which are jointly optimized by minimizing the objective

$$\bar{J}(\boldsymbol{T}, \boldsymbol{D}_1, \boldsymbol{D}_4) = \bar{J}^h(\boldsymbol{T}_1, \boldsymbol{D}_1) + \bar{J}^v(\boldsymbol{T}_4, \boldsymbol{D}_4)$$

where

$$\bar{J}^h(\boldsymbol{T}_1,\boldsymbol{D}_1)=\mathrm{tr}\Big[\boldsymbol{T}_1^T\left(\boldsymbol{W}^h-\boldsymbol{c}_1^T\boldsymbol{c}_1\right)\boldsymbol{T}_1-2\sum_{k=0}^{M-1}\boldsymbol{T}_1^T\boldsymbol{W}^h\boldsymbol{A}_1^{k+1}\boldsymbol{T}_1\boldsymbol{D}_{1k}$$
$$+\sum_{k=0}^{M-1}\sum_{l=0}^{M-1}\boldsymbol{T}_1^T\boldsymbol{W}^h\boldsymbol{A}_1^{|k-l|}\boldsymbol{T}_1\boldsymbol{D}_{1k}\boldsymbol{D}_{1l}\Big]$$

and

$$\bar{J}^v(\boldsymbol{T}_4,\boldsymbol{D}_4)=\mathrm{tr}\Big[\boldsymbol{T}_4^T\left(\boldsymbol{W}^v-\boldsymbol{c}_2^T\boldsymbol{c}_2\right)\boldsymbol{T}_4-2\sum_{k=0}^{N-1}\boldsymbol{T}_4^T\boldsymbol{W}^v\boldsymbol{A}_4^{k+1}\boldsymbol{T}_4\boldsymbol{D}_{4k}$$
$$+\sum_{k=0}^{N-1}\sum_{l=0}^{N-1}\boldsymbol{T}_4^T\boldsymbol{W}^v\boldsymbol{A}_4^{|k-l|}\boldsymbol{T}_4\boldsymbol{D}_{4k}\boldsymbol{D}_{4l}\Big]$$

- First, the l_2 constraints $(\boldsymbol{T}^{-1}\boldsymbol{K}\boldsymbol{T})_{ii}=1$ are eliminated in two simple steps:
  (1) Variable change $\hat{\boldsymbol{T}}=\hat{\boldsymbol{T}}_1\oplus\hat{\boldsymbol{T}}_4=(\boldsymbol{T}_1\oplus\boldsymbol{T}_4)^T(\boldsymbol{K}^h\oplus\boldsymbol{K}^v)^{-1/2}$ simplifies the constraints to $(\hat{\boldsymbol{T}}^{-T}\hat{\boldsymbol{T}}^{-1})_{ii}=1$;
  (2) The above constraints are automatically satisfied by assuming

$$\hat{\boldsymbol{T}}_1^{-1}=\begin{bmatrix}\dfrac{\boldsymbol{t}_{11}}{\|\boldsymbol{t}_{11}\|} & \dfrac{\boldsymbol{t}_{12}}{\|\boldsymbol{t}_{12}\|} & \cdots & \dfrac{\boldsymbol{t}_{1m}}{\|\boldsymbol{t}_{1m}\|}\end{bmatrix} \quad \text{and} \quad \hat{\boldsymbol{T}}_4^{-1}=\begin{bmatrix}\dfrac{\boldsymbol{t}_{41}}{\|\boldsymbol{t}_{41}\|} & \dfrac{\boldsymbol{t}_{42}}{\|\boldsymbol{t}_{42}\|} & \cdots & \dfrac{\boldsymbol{t}_{4n}}{\|\boldsymbol{t}_{4n}\|}\end{bmatrix}$$

- Now the design variables in the gain minimization problem are $\boldsymbol{t}_{11}$, …, $\boldsymbol{t}_{1m}$, $\boldsymbol{t}_{41}$, …, $\boldsymbol{t}_{4n}$, $\boldsymbol{D}_{10}$, …, $\boldsymbol{D}_{1,M-1}$, $\boldsymbol{D}_{40}$, …, $\boldsymbol{D}_{4,N-1}$. These variables are arranged to form a column vector $\boldsymbol{x}$, so the noise gain is a function of $\boldsymbol{x}$, hence is denoted by $J(\boldsymbol{x})$.
- A quasi-Newton algorithm with BFGS updates was applied to minimize $J(\boldsymbol{x})$ iteratively.

◊ It starts with a trivial initial point $\boldsymbol{x}_0$ which is obtained by assigning $\boldsymbol{T}_1$, $\boldsymbol{T}_4$, and all $\boldsymbol{D}_{1k}$ and $\boldsymbol{D}_{4l}$ to be identity matrices.

◊ It updates iterate $\boldsymbol{x}_k$ to $\boldsymbol{x}_{k+1}$ using $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{d}_k$ where

$$\boldsymbol{d}_k = -\boldsymbol{S}_k \nabla J(\boldsymbol{x}_k), \quad \alpha_k = \arg\left[\min_{\alpha} J(\boldsymbol{x}_k + \alpha \boldsymbol{d}_k)\right]$$

$$\boldsymbol{S}_{k+1} = \boldsymbol{S}_k + \left(1 + \frac{\boldsymbol{\gamma}_k^T \boldsymbol{S}_k \boldsymbol{\gamma}_k}{\boldsymbol{\gamma}_k^T \boldsymbol{\delta}_k}\right) \frac{\boldsymbol{\delta}_k \boldsymbol{\delta}_k^T}{\boldsymbol{\gamma}_k^T \boldsymbol{\delta}_k} - \frac{\boldsymbol{\delta}_k \boldsymbol{\gamma}_k^T \boldsymbol{S}_k + \boldsymbol{S}_k \boldsymbol{\gamma}_k \boldsymbol{\delta}_k^T}{\boldsymbol{\gamma}_k^T \boldsymbol{\delta}_k}$$

$$\boldsymbol{S}_0 = \boldsymbol{I}, \ \boldsymbol{\delta}_k = \boldsymbol{x}_{k+1} - \boldsymbol{x}_k, \ \boldsymbol{\gamma}_k = \nabla J(\boldsymbol{x}_{k+1}) - \nabla J(\boldsymbol{x}_k)$$

◊ Closed-form formula for $\nabla J(\boldsymbol{x}_k)$ for fast and accurate computation has been derived .

4. Numerical Example

We consider a 2-D separable-denominator digital filter $(\boldsymbol{A},\boldsymbol{b},\boldsymbol{c},d)_{3+3}$ with

$$\boldsymbol{A}_1^T = \begin{bmatrix} 0 & 0 & 0.599655 \\ 1 & 0 & -1.836929 \\ 0 & 1 & 2.173645 \end{bmatrix}, \quad \boldsymbol{A}_2 = \begin{bmatrix} 0.064564 & 0.033034 & 0.012881 \\ 0.091213 & 0.110512 & 0.102759 \\ 0.097256 & 0.151864 & 0.172460 \end{bmatrix}$$

$$\boldsymbol{A}_4 = \begin{bmatrix} 0 & 0 & 0.564961 \\ 1 & 0 & -1.887939 \\ 0 & 1 & 2.280029 \end{bmatrix}, \quad \boldsymbol{b}_1 = \begin{bmatrix} 0.047053 \\ 0.062274 \\ 0.060436 \end{bmatrix}, \quad \boldsymbol{b}_2 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \boldsymbol{c}_1^T = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

$$\boldsymbol{c}_2 = \begin{bmatrix} 0.016556 & 0.012550 & 0.008243 \end{bmatrix}, \quad d = 0.019421.$$

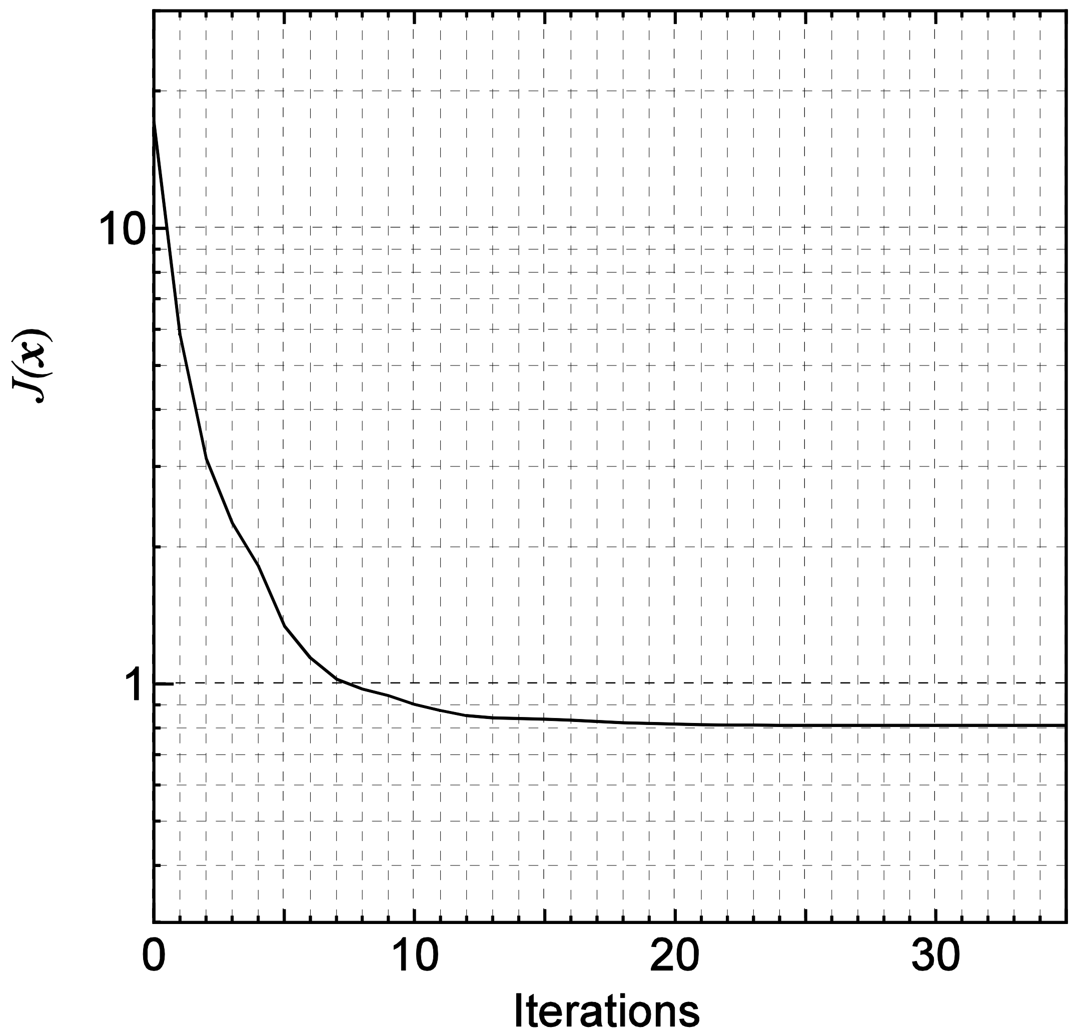
Case 1: Minimum noise gain without error feedforward and feedback :

$$\hat{J}_{\min}(\hat{\boldsymbol{T}}, \mathbf{0}, \mathbf{0}) = 7.936533$$

Case 2: Minimum noise gain with error feedforward and feedback:

(M, N)	Infinite Precision	3-Bit Quantization	Integer Quantization
(1, 1)	1.3004	1.3518	3.0832
(2, 1)	1.0217	1.0672	2.2677
(1, 2)	1.0899	1.1349	3.0247
(2, 2)	0.8112	0.8502	2.2091

- The figure below depicts the profile of the noise gain $J(\boldsymbol{x}_k)$ in first 35 iterations for the case of $(M, N) = (2, 2)$.



Thank you.

Q & A