 CMES, vol.104, no.1, pp.1-39, 2015

Double Optimal Regularization Algorithms for Solving Ill-Posed Linear Problems under Large Noise

Chein-Shan Liu[1], **Satya N. Atluri**[2]

Abstract: A double optimal solution of an n-dimensional system of linear equations $\mathbf{Ax}=\mathbf{b}$ has been derived in an affine m-dimensional Krylov subspace with $m \ll n$. We further develop a *double optimal iterative algorithm* (DOIA), with the descent direction $\mathbf{z}$ being solved from the residual equation $\mathbf{Az}=\mathbf{r}_0$ by using its double optimal solution, to solve ill-posed linear problem under large noise. The DOIA is proven to be absolutely convergent step-by-step with the square residual error $\|\mathbf{r}\|^2=\|\mathbf{b}-\mathbf{Ax}\|^2$ being reduced by a positive quantity $\|\mathbf{Az}_k\|^2$ at each iteration step, which is found to be better than those algorithms based on the minimization of the square residual error in an m-dimensional Krylov subspace. In order to tackle the ill-posed linear problem under a large noise, we also propose a novel *double optimal regularization algorithm* (DORA) to solve it, which is an improvement of the Tikhonov regularization method. Some numerical tests reveal the high performance of DOIA and DORA against large noise. These methods are of use in the ill-posed problems of structural health-monitoring.

Keywords: Ill-posed linear equations system, Double optimal solution, Affine Krylov subspace, Double optimal iterative algorithm, Double optimal regularization algorithm.

1 Introduction

A double optimal solution of a linear equations system has been derived in an affine Krylov subspace by Liu (2014a). The Krylov subspace methods are among the most widely used iterative algorithms for solving systems of linear equations [Dongarra and Sullivan (2000); Freund and Nachtigal (1991); Liu (2013a); Saad (1981); van Den Eshof and Sleijpen (2004)]. The iterative algorithms that are applied to solve large scale linear systems are likely to be the preconditioned Krylov subspace

[1] Department of Civil Engineering, National Taiwan University, Taipei, Taiwan. E-mail: liucs@ntu.edu.tw

[2] Center for Aerospace Research & Education, University of California, Irvine.

methods [Simoncini and Szyld (2007)]. Since the pioneering works of Hestenes (1952) and Lanczos (1952), the Krylov subspace methods have been further studied, like the minimum residual algorithm [Paige and Saunders (1975)], the generalized minimal residual method (GMRES) [Saad (1981); Saad and Schultz (1986)], the quasi-minimal residual method [Freund and Nachtigal (1991)], the biconjugate gradient method [Fletcher (1976)], the conjugate gradient squared method [Sonneveld (1989)], and the biconjugate gradient stabilized method [van der Vorst (1992)]. There are a lot of discussions on the Krylov subspace methods in Simoncini and Szyld (2007), Saad and van der Vorst (2000), Saad (2003), and van der Vorst (2003). The iterative method GMRES and several implementations for the GMRES were assessed for solving ill-posed linear systems by Matinfar, Zareamoghaddam, Eslami and Saeidy (2012). On the other hand, the Arnoldi's full orthogonalization method (FOM) is also an effective and useful algorithm to solve a system of linear equations [Saad (2003)].

Based on two minimization techniques being realized in an affine Krylov subspace, Liu (2014a) has recently developed a new theory to find a double optimal solution of the following linear equations system:

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^n$ is an unknown vector, to be determined from a given non-singular coefficient matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, and the input vector $\mathbf{b} \in \mathbb{R}^n$. For the existence of solution $\mathbf{x}$ we suppose that $\text{rank}(\mathbf{A}) = n$.

Sometimes the above equation is obtained via an n-dimensional discretization of a bounded linear operator equation under a noisy input. We only look for a generalized solution $\mathbf{x} = \mathbf{A}^\dagger\mathbf{b}$, where $\mathbf{A}^\dagger$ is a pseudo-inverse of $\mathbf{A}$ in the Penrose sense. When $\mathbf{A}$ is severely ill-posed and the data are disturbanced by random noise, the numerical solution of Eq. (1) might deviate from the exact one. If we only know the perturbed input data $\mathbf{b}^\delta \in \mathbb{R}^n$ with $\|\mathbf{b}-\mathbf{b}^\delta\| \le \delta$, and if the problem is ill-posed, i.e., the range$(\mathbf{A})$ is not closed or equivalently $\mathbf{A}^\dagger$ is unbounded, we have to solve Eq. (1) by a regularization method [Daubechies and Defrise (2004)].

Given an initial guess $\mathbf{x}_0$, from Eq. (1) we have an initial residual:

$$\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0. \tag{2}$$

Upon letting

$$\mathbf{z} = \mathbf{x} - \mathbf{x}_0, \tag{3}$$

Eq. (1) is equivalent to

$$\mathbf{A}\mathbf{z} = \mathbf{r}_0, \tag{4}$$

which can be used to search the Newton descent direction $\mathbf{z}$ after giving an initial residual $\mathbf{r}_0$ [Liu (2012a)]. Therefore, Eq. (4) may be called the *residual equation*.

Liu (2012b, 2013b, 2013c) has proposed the following merit function:

$$\min_{\mathbf{z}} \left\{ a_0 = \frac{\|\mathbf{r}_0\|^2 \|\mathbf{A}\mathbf{z}\|^2}{[\mathbf{r}_0 \cdot (\mathbf{A}\mathbf{z})]^2} \right\}, \tag{5}$$

and minimized it to obtain a fast descent direction $\mathbf{z}$ in the iterative solution of Eq. (1) in a two- or three-dimensional subspace.

Suppose that we have an m-dimensional Krylov subspace generated by the coefficient matrix $\mathbf{A}$ from the right-hand side vector $\mathbf{r}_0$ in Eq. (4):

$$\mathscr{K}_m := \text{span}\{\mathbf{r}_0, \mathbf{A}\mathbf{r}_0, \ldots, \mathbf{A}^{m-1}\mathbf{r}_0\}. \tag{6}$$

Let $\mathscr{L}_m = \mathbf{A}\mathscr{K}_m$. The idea of GMRES is using the Galerkin method to search the solution $\mathbf{z} \in \mathscr{K}_m$, such that the residual $\mathbf{r}_0 - \mathbf{A}\mathbf{z}$ is perpendicular to $\mathscr{L}_m$ [Saad and Schultz (1986)]. It can be shown that the solution $\mathbf{z} \in \mathscr{K}_m$ minimizes the residual [Saad (2003)]:

$$\min_{\mathbf{z}} \{\|\mathbf{r}_0 - \mathbf{A}\mathbf{z}\|^2 = \|\mathbf{b} - \mathbf{A}\mathbf{x}\|^2\}. \tag{7}$$

The Arnoldi process is used to normalize and orthogonalize the Krylov vectors $\mathbf{A}^j\mathbf{r}_0,\ j = 0, \ldots, m-1$, such that the resultant vectors $\mathbf{u}_i,\ i = 1, \ldots, m$ satisfy $\mathbf{u}_i \cdot \mathbf{u}_j = \delta_{ij},\ i, j = 1, \ldots, m$, where δ_{ij} is the Kronecker delta symbol.

The FOM used to solve Eq. (1) can be summarized as follows [Saad (2003)].

(i) Select m and give an initial $\mathbf{x}_0$.

(ii) For $k = 0, 1, \ldots$, we repeat the following computations:

$$\begin{aligned}
&\mathbf{r}_k = \mathbf{b} - \mathbf{A}\mathbf{x}_k, \\
&\text{Arnoldi procedure to set up } \mathbf{u}_j^k,\ j = 1, \ldots, m,\ (\text{from } \mathbf{u}_1^k = \mathbf{r}_k / \|\mathbf{r}_k\|), \\
&\mathbf{U}_k = [\mathbf{u}_1^k, \ldots, \mathbf{u}_m^k], \\
&\mathbf{V}_k = \mathbf{A}\mathbf{U}_k, \\
&\text{Solve } (\mathbf{U}_k^{\mathrm{T}}\mathbf{V}_k)\alpha_k = \mathbf{U}_k^{\mathrm{T}}\mathbf{r}_k, \text{ obtaining } \alpha_k, \\
&\mathbf{z}_k = \mathbf{U}_k\alpha_k, \\
&\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{z}_k.
\end{aligned} \tag{8}$$

If $\mathbf{x}_{k+1}$ converges according to a given stopping criterion $\|\mathbf{r}_{k+1}\| < \varepsilon$, then stop; otherwise, go to step (ii). $\mathbf{U}_k$ and $\mathbf{V}_k$ are both $n \times m$ matrices. In above, the superscript T signifies the transpose.

The GMRES used to solve Eq. (1) can be summarized as follows [Saad (2003)].
(i) Select m and give an initial $\mathbf{x}_0$.
(ii) For $k = 0, 1, \ldots$, we repeat the following computations:

$$\begin{aligned}
&\mathbf{r}_k = \mathbf{b} - \mathbf{A}\mathbf{x}_k, \\
&\text{Arnoldi procedure to set up } \mathbf{u}_j^k,\ j = 1, \ldots, m,\ (\text{from } \mathbf{u}_1^k = \mathbf{r}_k / \|\mathbf{r}_k\|), \\
&\mathbf{U}_k = [\mathbf{u}_1^k, \ldots, \mathbf{u}_m^k], \\
&\text{Solve } (\bar{\mathbf{H}}_k^{\mathrm{T}} \bar{\mathbf{H}}_k) \alpha_k = \|\mathbf{r}_k\| \bar{\mathbf{H}}_k^{\mathrm{T}} \mathbf{e}_1, \text{ obtaining } \alpha_k, \\
&\mathbf{z}_k = \mathbf{U}_k \alpha_k, \\
&\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{z}_k.
\end{aligned} \tag{9}$$

If $\mathbf{x}_{k+1}$ converges according to a given stopping criterion $\|\mathbf{r}_{k+1}\| < \varepsilon$, then stop; otherwise, go to step (ii). $\mathbf{U}_k$ is an $n \times m$ Krylov matrix, while $\bar{\mathbf{H}}_k$ is an augmented Heissenberg upper triangular matrix with $(m+1) \times m$, and $\mathbf{e}_1$ is the first column of $\mathbf{I}_{m+1}$.

So far, there are only a few works in Liu (2013d, 2014b, 2014c, 2015) that the numerical methods to solve Eq. (1) are based on the two minimizations in Eqs. (5) and (7). As a continuation of these works, we will employ an affine Krylov subspace method to derive a closed-form double optimal solution $\mathbf{z}$ of the residual Eq. (4), which is used in the iterative algorithm for solving the ill-posed linear system (1) by $\mathbf{x} = \mathbf{x}_0 + \mathbf{z}$.

The remaining parts of this paper are arranged as follows. In Section 2 we start from an affine m-dimensional Krylov subspace to expand the solution of the residual Eq. (4) with $m+1$ coefficients to be obtained in Section 3, where two merit functions are proposed for the determination of the $m+1$ expansion coefficients. We can derive a closed-form double optimal solution of the residual Eq. (4). The resulting algorithm, namely the double optimal iterative algorithm (DOIA), based on the idea of double optimal solution is developed in Section 4, which is proven to be absolutely convergent with the square residual norm being reduced by a positive quantity $\|\mathbf{A}\mathbf{x}_k - \mathbf{A}\mathbf{x}_0\|^2$ at each iteration step. In order to solve the ill-posed linear problem under a large noise, we derive a double optimal regularization algorithm (DORA) in Section 5. The examples of linear inverse problems solved by the FOM, GMRES, DOIA and DORA are compared in Section 6, of which some advantages of the DOIA and DORA to solve Eq. (1) under a large noise are displayed. Finally, we conclude this study in Section 7.

2 An affine Krylov subspace method

For Eq. (4), by using the Cayley-Hamilton theorem we can expand $\mathbf{A}^{-1}$ by

$$\mathbf{A}^{-1} = \frac{c_1}{c_0}\mathbf{I}_n + \frac{c_2}{c_0}\mathbf{A} + \frac{c_3}{c_0}\mathbf{A}^2 + \ldots + \frac{c_{n-1}}{c_0}\mathbf{A}^{n-2} + \frac{1}{c_0}\mathbf{A}^{n-1}, \tag{10}$$

and hence, the solution $\mathbf{z}$ is given by

$$\mathbf{z} = \mathbf{A}^{-1}\mathbf{r}_0 = \left[\frac{c_1}{c_0}\mathbf{I}_n + \frac{c_2}{c_0}\mathbf{A} + \frac{c_3}{c_0}\mathbf{A}^2 + \ldots + \frac{c_{n-1}}{c_0}\mathbf{A}^{n-2} + \frac{1}{c_0}\mathbf{A}^{n-1}\right]\mathbf{r}_0, \tag{11}$$

where the coefficients $c_0, c_1, \ldots, c_{n-1}$ are those that appear in the characteristic equation for $\mathbf{A}$: $\lambda^n + c_{n-1}\lambda^{n-1} + \ldots + c_2\lambda^2 + c_1\lambda - c_0 = 0$. Here, $c_0 = -\det(\mathbf{A}) \neq 0$ due to the fact that $\mathrm{rank}(\mathbf{A}) = n$. In practice, the above process to find the exact solution of $\mathbf{z}$ is quite difficult, since the coefficients $c_j,\ j = 0, 1, \ldots, n-1$ are hard to find when the problem dimension n is very large.

Instead of the m-dimensional Krylov subspace in Eq. (6), we consider an affine Krylov subspace generated by the following processes. First we introduce an m-dimensional Krylov subspace generated by the coefficient matrix $\mathbf{A}$ from $\mathscr{K}_m$:

$$\mathscr{L}_m := \mathrm{span}\{\mathbf{A}\mathbf{r}_0, \ldots, \mathbf{A}^m\mathbf{r}_0\} = \mathbf{A}\mathscr{K}_m. \tag{12}$$

Then, the Arnoldi process is used to normalize and orthogonalize the Krylov vectors $\mathbf{A}^j\mathbf{r}_0,\ j = 1, \ldots, m$, such that the resultant vectors $\mathbf{u}_i,\ i = 1, \ldots, m$ satisfy $\mathbf{u}_i \cdot \mathbf{u}_j = \delta_{ij},\ i, j = 1, \ldots, m$.

While in the FOM, $\mathbf{z}$ is searched such that the square residual error of $\mathbf{r}_0 - \mathbf{A}\mathbf{z}$ in Eq. (7) is minimized, in the GMRES, $\mathbf{z}$ is searched such that the residual vector $\mathbf{r}_0 - \mathbf{A}\mathbf{z}$ is orthogonal to $\mathscr{L}_m$ [Saad and Schultz (1986)]. In this paper we seek a different and better $\mathbf{z}$, than those in Eqs. (8) and (9), with a *more fundamental method* by expanding the solution $\mathbf{z}$ of Eq. (4) in the following affine Krylov subspace:

$$\mathscr{K}'_m = \mathrm{span}\{\mathbf{r}_0, \mathscr{L}_m\} = \mathrm{span}\{\mathbf{r}_0, \mathbf{A}\mathscr{K}_m\}, \tag{13}$$

that is,

$$\mathbf{z} = \alpha_0\mathbf{r}_0 + \sum_{k=1}^{m} \alpha_k\mathbf{u}_k \in \mathscr{K}'_m. \tag{14}$$

It is motivated by Eq. (11), and is to be determined as a *double optimal combination* of $\mathbf{r}_0$ and the m-vector $\mathbf{u}_k,\ k = 1, \ldots, m$ in an affine Krylov subspace, of which the coefficients α_0 and α_k are determined in Section 3.2. For finding the solution $\mathbf{z}$ in a smaller subspace we suppose that $m \ll n$.

Let

$$\mathbf{U} := [\mathbf{u}_1, \ldots, \mathbf{u}_m] \tag{15}$$

be an $n \times m$ matrix with its jth column being the vector $\mathbf{u}_j$, which is specified below Eq. (12). The dimension m is selected such that $\mathbf{u}_1, \ldots, \mathbf{u}_m$ are linearly independent vectors, which renders $\text{rank}(\mathbf{U}) = m$, and $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_m$. Now, Eq. (14) can be written as

$$\mathbf{z} = \mathbf{z}_0 + \mathbf{U}\alpha, \tag{16}$$

where

$$\mathbf{z}_0 = \alpha_0 \mathbf{r}_0, \tag{17}$$

$$\alpha := (\alpha_1, \ldots, \alpha_m)^{\mathrm{T}}. \tag{18}$$

Below we will introduce two merit functions, whose minimizations determine the coefficients (α_0, α) uniquely.

3 A double optimal descent direction

3.1 Two merit functions

Let

$$\mathbf{y} := \mathbf{A}\mathbf{z}, \tag{19}$$

and we attempt to establish a merit function, such that its minimization leads to the best fit of $\mathbf{y}$ to $\mathbf{r}_0$, because $\mathbf{A}\mathbf{z} = \mathbf{r}_0$ is the residual equation we want to solve.

The orthogonal projection of $\mathbf{r}_0$ to $\mathbf{y}$ is regarded as an approximation of $\mathbf{r}_0$ by $\mathbf{y}$ with the following error vector:

$$\mathbf{e} := \mathbf{r}_0 - \left(\mathbf{r}_0, \frac{\mathbf{y}}{\|\mathbf{y}\|}\right) \frac{\mathbf{y}}{\|\mathbf{y}\|}, \tag{20}$$

where the parenthesis denotes the inner product. The best approximation can be found by $\mathbf{y}$ minimizing the square norm of $\mathbf{e}$:

$$\min_{\mathbf{y}} \left\{ \|\mathbf{e}\|^2 = \|\mathbf{r}_0\|^2 - \frac{(\mathbf{r}_0 \cdot \mathbf{y})^2}{\|\mathbf{y}\|^2} \right\}, \tag{21}$$

or maximizing the square orthogonal projection of $\mathbf{r}_0$ to $\mathbf{y}$:

$$\max \left(\mathbf{r}_0, \frac{\mathbf{y}}{\|\mathbf{y}\|}\right)^2 = \max_{\mathbf{y}} \left\{ \frac{(\mathbf{r}_0 \cdot \mathbf{y})^2}{\|\mathbf{y}\|^2} \right\}. \tag{22}$$

Let us define the following merit function:

$$f := \frac{\|\mathbf{y}\|^2}{(\mathbf{r}_0 \cdot \mathbf{y})^2}, \tag{23}$$

which is similar to a_0 in Eq. (5) by noting that $\mathbf{y} = \mathbf{A}\mathbf{z}$. Let $\mathbf{J}$ be an $n \times m$ matrix:

$$\mathbf{J} := \mathbf{A}\mathbf{U}, \tag{24}$$

where $\mathbf{U}$ is defined by Eq. (15). Due to the fact that $\text{rank}(\mathbf{A}) = n$ and $\text{rank}(\mathbf{U}) = m$ one has $\text{rank}(\mathbf{J}) = m$. Then, Eq. (19) can be written as

$$\mathbf{y} = \mathbf{y}_0 + \mathbf{J}\alpha, \tag{25}$$

with the aid of Eq. (16), where

$$\mathbf{y}_0 := \mathbf{A}\mathbf{z}_0 = \alpha_0 \mathbf{A}\mathbf{r}_0. \tag{26}$$

Inserting Eq. (25) for $\mathbf{y}$ into Eq. (23), we encounter the following minimization problem:

$$\min_{\alpha=(\alpha_1,\ldots,\alpha_m)^{\mathrm{T}}} \left\{ f = \frac{\|\mathbf{y}\|^2}{(\mathbf{r}_0 \cdot \mathbf{y})^2} = \frac{\|\mathbf{y}_0 + \mathbf{J}\alpha\|^2}{(\mathbf{r}_0 \cdot \mathbf{y}_0 + \mathbf{r}_0 \cdot \mathbf{J}\alpha)^2} \right\}. \tag{27}$$

The optimization probelms in Eqs. (21), (22) and (27) are mathematically equivalent.

The minimization problem in Eq. (27) is used to find α; however, for $\mathbf{y}$ there still is an unknown scalar α_0 in $\mathbf{y}_0 = \alpha_0 \mathbf{A}\mathbf{r}_0$. So we can further consider the minimization problem of the square residual:

$$\min_{\alpha_0} \{\|\mathbf{r}\|^2 = \|\mathbf{b} - \mathbf{A}\mathbf{x}\|^2\}. \tag{28}$$

By using Eqs. (2) and (3) we have

$$\mathbf{b} - \mathbf{A}\mathbf{x} = \mathbf{b} - \mathbf{A}\mathbf{x}_0 - \mathbf{A}\mathbf{z} = \mathbf{r}_0 - \mathbf{A}\mathbf{z}, \tag{29}$$

such that we have the second merit function to be minimized:

$$\min_{\alpha_0} \|\mathbf{r}_0 - \mathbf{A}\mathbf{z}\|^2. \tag{30}$$

3.2 Main result

In the above we have introduced two merit functions to determine the expansion coefficients $\alpha_j,\ j=0,1,\ldots,m$. We must emphasize that the two merit functions in Eqs. (27) and (30) are different, from which, Liu (2014a) has proposed the method to solve these two optimization problems for Eq. (1). For making this paper reasonably self-content, we repeat some results in Liu (2014a) for the residual Eq. (4), instead of the original Eq. (1). As a consequence, we can prove the following main theorem.

Theorem 1: *For $\mathbf{z}\in\mathscr{K}'_m$, the double optimal solution of the residual Eq. (4) derived from the minimizations in Eqs. (27) and (30) is given by*

$$\mathbf{z}=\mathbf{X}\mathbf{r}_0+\alpha_0(\mathbf{r}_0-\mathbf{X}\mathbf{A}\mathbf{r}_0), \tag{31}$$

where

$$\begin{aligned}&\mathbf{C}=\mathbf{J}^{\mathrm{T}}\mathbf{J},\ \mathbf{D}=(\mathbf{J}^{\mathrm{T}}\mathbf{J})^{-1},\ \mathbf{X}=\mathbf{U}\mathbf{D}\mathbf{J}^{\mathrm{T}},\ \mathbf{E}=\mathbf{A}\mathbf{X},\\ &\alpha_0=\frac{\mathbf{r}_0^{\mathrm{T}}(\mathbf{I}_n-\mathbf{E})\mathbf{A}\mathbf{r}_0}{\mathbf{r}_0^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}(\mathbf{I}_n-\mathbf{E})\mathbf{A}\mathbf{r}_0}.\end{aligned} \tag{32}$$

The proof of Theorem 1 is quite complicated and delicate. Before embarking on the proof of Theorem 1, we need to prove the following two lemmas.

Lemma 1: *For $\mathbf{z}\in\mathscr{K}'_m$, the double optimal solution of Eq. (4) derived from the minimizations in Eqs. (27) and (30) is given by*

$$\mathbf{z}=\alpha_0(\mathbf{r}_0+\lambda_0\mathbf{X}\mathbf{r}_0-\mathbf{X}\mathbf{A}\mathbf{r}_0), \tag{33}$$

where $\mathbf{C}$*,* $\mathbf{D}$*,* $\mathbf{X}$*, and* $\mathbf{E}$ *were defined in Theorem 1, and others are given by*

$$\begin{aligned}&\lambda_0=\frac{\mathbf{r}_0^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}(\mathbf{I}_n-\mathbf{E})\mathbf{A}\mathbf{r}_0}{\mathbf{r}_0^{\mathrm{T}}(\mathbf{I}_n-\mathbf{E})\mathbf{A}\mathbf{r}_0},\\ &\mathbf{w}=\lambda_0\mathbf{E}\mathbf{r}_0+\mathbf{A}\mathbf{r}_0-\mathbf{E}\mathbf{A}\mathbf{r}_0,\\ &\alpha_0=\frac{\mathbf{w}\cdot\mathbf{r}_0}{\|\mathbf{w}\|^2}.\end{aligned} \tag{34}$$

Proof: First with the help of Eq. (25), the terms $\mathbf{r}_0\cdot\mathbf{y}$ and $\|\mathbf{y}\|^2$ in Eq. (27) can be written as

$$\mathbf{r}_0\cdot\mathbf{y}=\mathbf{r}_0\cdot\mathbf{y}_0+\mathbf{r}_0^{\mathrm{T}}\mathbf{J}\alpha, \tag{35}$$

$$\|\mathbf{y}\|^2 = \|\mathbf{y}_0\|^2 + 2\mathbf{y}_0^{\mathrm{T}}\mathbf{J}\alpha + \alpha^{\mathrm{T}}\mathbf{J}^{\mathrm{T}}\mathbf{J}\alpha. \tag{36}$$

For the minimization of f we have a necessary condition:

$$\nabla\alpha \frac{\|\mathbf{y}\|^2}{(\mathbf{r}_0 \cdot \mathbf{y})^2} = \mathbf{0} \Rightarrow (\mathbf{r}_0 \cdot \mathbf{y})^2 \nabla\alpha \|\mathbf{y}\|^2 - 2\mathbf{r}_0 \cdot \mathbf{y}\|\mathbf{y}\|^2 \nabla\alpha(\mathbf{r}_0 \cdot \mathbf{y}) = \mathbf{0}, \tag{37}$$

in which $\nabla\alpha$ denotes the gradient with respect to α. Thus, we can derive the following equation to solve α:

$$\mathbf{r}_0 \cdot \mathbf{y}\mathbf{y}_2 - 2\|\mathbf{y}\|^2\mathbf{y}_1 = \mathbf{0}, \tag{38}$$

where

$$\mathbf{y}_1 := \nabla\alpha(\mathbf{r}_0 \cdot \mathbf{y}) = \mathbf{J}^{\mathrm{T}}\mathbf{r}_0, \tag{39}$$

$$\mathbf{y}_2 := \nabla\alpha\|\mathbf{y}\|^2 = 2\mathbf{J}^{\mathrm{T}}\mathbf{y}_0 + 2\mathbf{J}^{\mathrm{T}}\mathbf{J}\alpha. \tag{40}$$

By letting

$$\mathbf{C} := \mathbf{J}^{\mathrm{T}}\mathbf{J}, \tag{41}$$

which is an $m \times m$ positive definite matrix because of $\mathrm{rank}(\mathbf{J}) = m$, Eqs. (36) and (40) can be written as

$$\|\mathbf{y}\|^2 = \|\mathbf{y}_0\|^2 + 2\mathbf{y}_0^{\mathrm{T}}\mathbf{J}\alpha + \alpha^{\mathrm{T}}\mathbf{C}\alpha, \tag{42}$$

$$\mathbf{y}_2 = 2\mathbf{J}^{\mathrm{T}}\mathbf{y}_0 + 2\mathbf{C}\alpha. \tag{43}$$

From Eq. (38) we can observe that $\mathbf{y}_2$ is proportional to $\mathbf{y}_1$, written as

$$\mathbf{y}_2 = \frac{2\|\mathbf{y}\|^2}{\mathbf{r}_0 \cdot \mathbf{y}}\mathbf{y}_1 = 2\lambda\mathbf{y}_1, \tag{44}$$

where 2λ is a multiplier to be determined. By cancelling $2\mathbf{y}_1$ on both sides from the second equality, we have

$$\|\mathbf{y}\|^2 = \lambda\mathbf{r}_0 \cdot \mathbf{y}. \tag{45}$$

Then, from Eqs. (39), (43) and (44) it follows that

$$\alpha = \lambda\mathbf{D}\mathbf{J}^{\mathrm{T}}\mathbf{r}_0 - \mathbf{D}\mathbf{J}^{\mathrm{T}}\mathbf{y}_0, \tag{46}$$

where

$$\mathbf{D} := \mathbf{C}^{-1} = (\mathbf{J}^{\mathrm{T}}\mathbf{J})^{-1} \tag{47}$$

is an $m \times m$ positive definite matrix. Inserting Eq. (46) into Eqs. (35) and (42) we have

$$\mathbf{r}_0 \cdot \mathbf{y} = \mathbf{r}_0 \cdot \mathbf{y}_0 + \lambda \mathbf{r}_0^{\mathrm{T}} \mathbf{E} \mathbf{r}_0 - \mathbf{r}_0^{\mathrm{T}} \mathbf{E} \mathbf{y}_0, \tag{48}$$

$$\|\mathbf{y}\|^2 = \lambda^2 \mathbf{r}_0^{\mathrm{T}} \mathbf{E} \mathbf{r}_0 + \|\mathbf{y}_0\|^2 - \mathbf{y}_0^{\mathrm{T}} \mathbf{E} \mathbf{y}_0, \tag{49}$$

where

$$\mathbf{E} := \mathbf{J}\mathbf{D}\mathbf{J}^{\mathrm{T}} \tag{50}$$

is an $n \times n$ positive semi-definite matrix. By Eq. (47) it is easy to check

$$\mathbf{E}^2 = \mathbf{J}\mathbf{D}\mathbf{J}^{\mathrm{T}}\mathbf{J}\mathbf{D}\mathbf{J}^{\mathrm{T}} = \mathbf{J}\mathbf{D}\mathbf{D}^{-1}\mathbf{D}\mathbf{J}^{\mathrm{T}} = \mathbf{J}\mathbf{D}\mathbf{J}^{\mathrm{T}} = \mathbf{E},$$

such that $\mathbf{E}$ is a projection operator, satisfying

$$\mathbf{E}^2 = \mathbf{E}. \tag{51}$$

Now, from Eqs. (45), (48) and (49) we can derive a linear equation:

$$\|\mathbf{y}_0\|^2 - \mathbf{y}_0^{\mathrm{T}} \mathbf{E} \mathbf{y}_0 = \lambda (\mathbf{r}_0 \cdot \mathbf{y}_0 - \mathbf{r}_0^{\mathrm{T}} \mathbf{E} \mathbf{y}_0), \tag{52}$$

such that λ is given by

$$\lambda = \frac{\|\mathbf{y}_0\|^2 - \mathbf{y}_0^{\mathrm{T}} \mathbf{E} \mathbf{y}_0}{\mathbf{r}_0 \cdot \mathbf{y}_0 - \mathbf{r}_0^{\mathrm{T}} \mathbf{E} \mathbf{y}_0}. \tag{53}$$

Inserting it into Eq. (46), the solution of α is obtained:

$$\alpha = \frac{\|\mathbf{y}_0\|^2 - \mathbf{y}_0^{\mathrm{T}} \mathbf{E} \mathbf{y}_0}{\mathbf{r}_0 \cdot \mathbf{y}_0 - \mathbf{r}_0^{\mathrm{T}} \mathbf{E} \mathbf{y}_0} \mathbf{D}\mathbf{J}^{\mathrm{T}} \mathbf{r}_0 - \mathbf{D}\mathbf{J}^{\mathrm{T}} \mathbf{y}_0. \tag{54}$$

Since $\mathbf{y}_0 = \alpha_0 \mathbf{A}\mathbf{r}_0$ still includes an unknown scalar α_0, we need another equation to determine α_0, and hence, α.

By inserting the above α into Eq. (16) we can obtain

$$\mathbf{z} = \mathbf{z}_0 + \lambda \mathbf{X}\mathbf{r}_0 - \mathbf{X}\mathbf{y}_0 = \alpha_0 (\mathbf{r}_0 + \lambda_0 \mathbf{X}\mathbf{r}_0 - \mathbf{X}\mathbf{A}\mathbf{r}_0), \tag{55}$$

where

$$\mathbf{X} := \mathbf{U}\mathbf{D}\mathbf{J}^{\mathrm{T}}, \tag{56}$$

$$\lambda_0 := \frac{\mathbf{r}_0^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} (\mathbf{I}_n - \mathbf{E}) \mathbf{A}\mathbf{r}_0}{\mathbf{r}_0^{\mathrm{T}} (\mathbf{I}_n - \mathbf{E}) \mathbf{A}\mathbf{r}_0}. \tag{57}$$

Multiplying Eq. (56) by $\mathbf{A}$ and using Eq. (24), then comparing the resultant with Eq. (50), it immediately follows that

$$\mathbf{E} = \mathbf{AX}. \tag{58}$$

Upon letting

$$\mathbf{v} := \mathbf{r}_0 + \lambda_0 \mathbf{X}\mathbf{r}_0 - \mathbf{X}\mathbf{A}\mathbf{r}_0, \tag{59}$$

$\mathbf{z}$ in Eq. (55) can be expressed as

$$\mathbf{z} = \alpha_0 \mathbf{v}, \tag{60}$$

where α_0 can be determined by minimizing the square residual error in Eq. (30). Inserting Eq. (60) into Eq. (30) we have

$$\|\mathbf{r}_0 - \mathbf{A}\mathbf{z}\|^2 = \|\mathbf{r}_0 - \alpha_0 \mathbf{A}\mathbf{v}\|^2 = \alpha_0^2 \|\mathbf{w}\|^2 - 2\alpha_0 \mathbf{w} \cdot \mathbf{r}_0 + \|\mathbf{r}_0\|^2, \tag{61}$$

where with the aid of Eq. (58) we have

$$\mathbf{w} := \mathbf{A}\mathbf{v} = \mathbf{A}\mathbf{r}_0 + \lambda_0 \mathbf{E}\mathbf{r}_0 - \mathbf{E}\mathbf{A}\mathbf{r}_0. \tag{62}$$

Taking the derivative of Eq. (61) with respect to α_0 and equating it to zero we can obtain

$$\alpha_0 = \frac{\mathbf{w} \cdot \mathbf{r}_0}{\|\mathbf{w}\|^2}. \tag{63}$$

Hence, $\mathbf{z}$ is given by

$$\mathbf{z} = \alpha_0 \mathbf{v} = \frac{\mathbf{w} \cdot \mathbf{r}_0}{\|\mathbf{w}\|^2} \mathbf{v}, \tag{64}$$

of which upon inserting Eq. (59) for $\mathbf{v}$ we can obtain Eq. (33). □

Lemma 2: *In Lemma 1, the two parameters α_0 and λ_0 satisfy the following reciprocal relation*:

$$\alpha_0 \lambda_0 = 1. \tag{65}$$

Proof: From Eq. (62) it follows that

$$\|\mathbf{w}\|^2 = \lambda_0^2 \mathbf{r}_0^{\mathrm{T}} \mathbf{E}\mathbf{r}_0 + \mathbf{r}_0^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \mathbf{A}\mathbf{r}_0 - \mathbf{r}_0^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \mathbf{E}\mathbf{A}\mathbf{r}_0, \tag{66}$$

$$\mathbf{r}_0 \cdot \mathbf{w} = \lambda_0 \mathbf{r}_0^{\mathrm{T}} \mathbf{E}\mathbf{r}_0 + \mathbf{r}_0^{\mathrm{T}} \mathbf{A}\mathbf{r}_0 - \mathbf{r}_0^{\mathrm{T}} \mathbf{E}\mathbf{A}\mathbf{r}_0, \tag{67}$$

where Eq. (51) was used in the first equation. With the aid of Eq. (57), Eq. (67) is further reduced to

$$\mathbf{r}_0 \cdot \mathbf{w} = \lambda_0 \mathbf{r}_0^{\mathrm{T}} \mathbf{E} \mathbf{r}_0 + \frac{1}{\lambda_0} [\mathbf{r}_0^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \mathbf{A} \mathbf{r}_0 - \mathbf{r}_0^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \mathbf{E} \mathbf{A} \mathbf{r}_0]. \tag{68}$$

Now, after inserting Eq. (68) into Eq. (63) we can obtain

$$\alpha_0 \lambda_0 \|\mathbf{w}\|^2 = \lambda_0^2 \mathbf{r}_0^{\mathrm{T}} \mathbf{E} \mathbf{r}_0 + \mathbf{r}_0^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \mathbf{A} \mathbf{r}_0 - \mathbf{r}_0^{\mathrm{T}} \mathbf{A}^{\mathrm{T}} \mathbf{E} \mathbf{A} \mathbf{r}_0. \tag{69}$$

In view of Eq. (66) the right-hand side is just equal to $\|\mathbf{w}\|^2$; hence, Eq. (65) is obtained readily. □

Proof of Theorem 1: According to Eq. (65) in Lemma 2, we can rearrange $\mathbf{z}$ in Eq. (55) to that in Eq. (31). Again, due to Eq. (65) in Lemma 2, α_0 can be derived as that in Eq. (32) by taking the reciprocal of λ_0 in Eq. (57). This ends the proof of Theorem 1. □

Remark 1: At the very beginning we have expanded $\mathbf{z}$ in an affine Krylov subspace as shown in Eq. (14). Why did we not expand $\mathbf{z}$ in a Krylov subspace? If we get rid of the term $\alpha_0 \mathbf{r}_0$ from $\mathbf{z}$ in Eq. (14), and expand $\mathbf{z}$ in a Krylov subspace, the term $\mathbf{y}_0$ in Eq. (26) is zero. As a result, λ in Eq. (53) and α in Eq. (54) cannot be defined. In summary, we cannot optimize the first merit function (27) in a Krylov subspace, and as that done in the above we should optimize the first merit function (27) in an *affine Krylov subspace*.

Remark 2: Indeed, the optimization of $\mathbf{z}$ in the Krylov subspace $\mathscr{K}_m$ has been done in the FOM and GMRES, which is the "best descent vector" in that space as shown in Eq. (7). So it is impossible to find "more best descent vector" in the Krylov subspace $\mathscr{K}_m$. The presented $\mathbf{z}$ in Theorem 1 is the "best descent vector" in the affine Krylov subspace $\mathscr{K}'_m$. Due to two reasons of the double optimal property of $\mathbf{z}$ and $\mathscr{K}'_m$ being larger than $\mathscr{K}_m$, we will prove in Section 4 that the algorithm based on Theorem 1 is better than the FOM and GMRES. Since the poineering work in Saad and Schultz (1986), there are many improvements of the GMRES; however, a qualitatively different improvement is not yet seen in the past literature.

Corollary 1: *In the double optimal solution of Eq. (4), if $m = n$ then $\mathbf{z}$ is the exact solution, given by*

$$\mathbf{z} = \mathbf{A}^{-1} \mathbf{r}_0. \tag{70}$$

Proof: If $m = n$, then $\mathbf{J}$ defined by Eq. (24) is an $n \times n$ non-singular matrix, due to $\text{rank}(\mathbf{J}) = n$. Simultaneously, $\mathbf{E}$ defined by Eq. (50) is an identity matrix, and meanwhile Eq. (54) reduces to

$$\alpha = \mathbf{J}^{-1}\mathbf{r}_0 - \mathbf{J}^{-1}\mathbf{y}_0. \tag{71}$$

Now, by Eqs. (25) and (71) we have

$$\mathbf{y} = \mathbf{y}_0 + \mathbf{J}\alpha = \mathbf{y}_0 + \mathbf{J}[\mathbf{J}^{-1}\mathbf{r}_0 - \mathbf{J}^{-1}\mathbf{y}_0] = \mathbf{r}_0. \tag{72}$$

Under this condition we have obtained the closed-form optimal solution of $\mathbf{z}$, by inserting Eq. (71) into Eq. (16) and using Eqs. (24) and (26):

$$\mathbf{z} = \mathbf{z}_0 + \mathbf{U}\mathbf{J}^{-1}\mathbf{r}_0 - \mathbf{U}\mathbf{J}^{-1}\mathbf{y}_0 = \mathbf{A}^{-1}\mathbf{r}_0 + \mathbf{z}_0 - \mathbf{A}^{-1}\mathbf{y}_0 = \mathbf{A}^{-1}\mathbf{r}_0. \tag{73}$$

This ends the proof. □

Inserting $\mathbf{y}_0 = \alpha_0\mathbf{A}\mathbf{r}_0$ into Eq. (54) and using Eq. (65) in Lemma 2, we can fully determine α by

$$\begin{aligned}\alpha &= \alpha_0 \left[\frac{\mathbf{r}_0^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{E})\mathbf{A}\mathbf{r}_0}{\mathbf{r}_0^{\mathrm{T}}(\mathbf{I}_n - \mathbf{E})\mathbf{A}\mathbf{r}_0}\mathbf{D}\mathbf{J}^{\mathrm{T}}\mathbf{r}_0 - \mathbf{D}\mathbf{J}^{\mathrm{T}}\mathbf{A}\mathbf{r}_0 \right] \\ &= \alpha_0(\lambda_0\mathbf{D}\mathbf{J}^{\mathrm{T}}\mathbf{r}_0 - \mathbf{D}\mathbf{J}^{\mathrm{T}}\mathbf{A}\mathbf{r}_0) \\ &= \mathbf{D}\mathbf{J}^{\mathrm{T}}\mathbf{r}_0 - \alpha_0\mathbf{D}\mathbf{J}^{\mathrm{T}}\mathbf{A}\mathbf{r}_0,\end{aligned} \tag{74}$$

where α_0 is defined by Eq. (32) in Theorem 1. It is interesting that if $\mathbf{r}_0$ is an eigenvector of $\mathbf{A}$, λ_0 defined by Eq. (57) is the corresponding eigenvalue of $\mathbf{A}$. By inserting $\mathbf{A}\mathbf{r}_0 = \lambda\mathbf{r}_0$ into Eq. (57), $\lambda_0 = \lambda$ follows immediately.

Remark 3: Upon comparing the above α with those used in the FOM and GMRES as shown in Eq. (8) and Eq. (9), respectively, we have an extra term $-\alpha_0\mathbf{D}\mathbf{J}^{\mathrm{T}}\mathbf{A}\mathbf{r}_0$. Moreover, for $\mathbf{z}$, in addition to the common term $\mathbf{U}\alpha$ as those used in the FOM and GMRES, we have an extra term $\alpha_0\mathbf{r}_0$ as shown by Eq. (31) in Theorem 1. Thus, for the descent vector $\mathbf{z}$ we have totally two extra terms $\alpha_0(\mathbf{r}_0 - \mathbf{X}\mathbf{A}\mathbf{r}_0)$ than those used in the FOM and GMRES. If we take $\alpha_0 = 0$ the two new extra terms disappear.

3.3 The estimation of residual error

About the residual errors of Eqs. (1) and (4) we can prove the following relations.

Theorem 2: *Under the double optimal solution of* $\mathbf{z} \in \mathscr{K}_m'$ *given in Theorem 1, the residual errors of Eqs. (1) and (4) have the following relations*:

$$\|\mathbf{r}_0 - \mathbf{A}\mathbf{z}\|^2 = \|\mathbf{r}_0\|^2 - \mathbf{r}_0^{\mathrm{T}}\mathbf{E}\mathbf{r}_0 - \frac{[\mathbf{r}_0^{\mathrm{T}}(\mathbf{I}_n - \mathbf{E})\mathbf{A}\mathbf{r}_0]^2}{\mathbf{r}_0^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{E})\mathbf{A}\mathbf{r}_0}, \tag{75}$$

$$\|\mathbf{r}_0 - \mathbf{A}\mathbf{z}\|^2 < \|\mathbf{r}_0\|^2. \tag{76}$$

Proof: Let us investigate the error of the residual equation (4):

$$\|\mathbf{r}_0 - \mathbf{A}\mathbf{z}\|^2 = \|\mathbf{r}_0 - \mathbf{y}\|^2 = \|\mathbf{r}_0\|^2 - 2\mathbf{r}_0 \cdot \mathbf{y} + \|\mathbf{y}\|^2, \tag{77}$$

where

$$\mathbf{y} = \mathbf{A}\mathbf{z} = \alpha_0\mathbf{A}\mathbf{r}_0 - \alpha_0\mathbf{E}\mathbf{A}\mathbf{r}_0 + \mathbf{E}\mathbf{r}_0. \tag{78}$$

is obtained from Eq. (31) by using Eqs. (19) and (58).

From Eq. (78) it follows that

$$\|\mathbf{y}\|^2 = \alpha_0^2(\mathbf{r}_0^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{r}_0 - \mathbf{r}_0^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{E}\mathbf{A}\mathbf{r}_0) + \mathbf{r}_0^{\mathrm{T}}\mathbf{E}\mathbf{r}_0, \tag{79}$$

$$\mathbf{r}_0 \cdot \mathbf{y} = \alpha_0\mathbf{r}_0^{\mathrm{T}}\mathbf{A}\mathbf{r}_0 - \alpha_0\mathbf{r}_0^{\mathrm{T}}\mathbf{E}\mathbf{A}\mathbf{r}_0 + \mathbf{r}_0^{\mathrm{T}}\mathbf{E}\mathbf{r}_0, \tag{80}$$

where Eq. (51) was used in the first equation. Then, inserting the above two equations into Eq. (77) we have

$$\|\mathbf{r}_0 - \mathbf{A}\mathbf{z}\|^2 = \alpha_0^2(\mathbf{r}_0^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{r}_0 - \mathbf{r}_0^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{E}\mathbf{A}\mathbf{r}_0) + 2\alpha_0(\mathbf{r}_0^{\mathrm{T}}\mathbf{E}\mathbf{A}\mathbf{r}_0 - \mathbf{r}_0^{\mathrm{T}}\mathbf{A}\mathbf{r}_0) + \|\mathbf{r}_0\|^2 - \mathbf{r}_0^{\mathrm{T}}\mathbf{E}\mathbf{r}_0. \tag{81}$$

Consequently, inserting Eq. (32) for α_0 into the above equation, yields Eq. (75). Since both $\mathbf{E}$ and $\mathbf{I}_n - \mathbf{E}$ are projection operators, we have

$$\begin{aligned} &\mathbf{r}_0^{\mathrm{T}}\mathbf{E}\mathbf{r}_0 > 0, \\ &\mathbf{r}_0^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{E})\mathbf{A}\mathbf{r}_0 > 0. \end{aligned} \tag{82}$$

Then, according to Eqs. (75) and (82) we can derive Eq. (76). □

3.4 Two merit functions are the same

In Section 3.1, the first merit function is used to adjust the orientation of $\mathbf{y}$ by best approaching to the orientation of $\mathbf{r}_0$, which is however disregarding the length of $\mathbf{y}$. Then, in the second merit function, we also ask the length of $\mathbf{y}$ best approaching to

the length of $\mathbf{r}_0$. Now, we can prove the following theorem.

Theorem 3: *Under the double optimal solution of* $\mathbf{z} \in \mathscr{K}_m^l$ *given in Theorem 1, the minimal values of the two merit functions are the same*, i.e.,

$$\|\mathbf{e}\|^2 = \|\mathbf{r}_0 - \mathbf{A}\mathbf{z}\|^2, \tag{83}$$

Moreover, we have

$$\|\mathbf{r}\|^2 < \|\mathbf{r}_0\|^2. \tag{84}$$

Proof: Inserting Eq. (32) for α_0 into Eqs. (79) and (80) we can obtain

$$\|\mathbf{y}\|^2 = \frac{[\mathbf{r}_0^{\mathrm{T}}(\mathbf{I}_n - \mathbf{E})\mathbf{A}\mathbf{r}_0]^2}{\mathbf{r}_0^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{E})\mathbf{A}\mathbf{r}_0} + \mathbf{r}_0^{\mathrm{T}}\mathbf{E}\mathbf{r}_0, \tag{85}$$

$$\mathbf{r}_0 \cdot \mathbf{y} = \frac{[\mathbf{r}_0^{\mathrm{T}}(\mathbf{I}_n - \mathbf{E})\mathbf{A}\mathbf{r}_0]^2}{\mathbf{r}_0^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{E})\mathbf{A}\mathbf{r}_0} + \mathbf{r}_0^{\mathrm{T}}\mathbf{E}\mathbf{r}_0; \tag{86}$$

consequently, we have

$$\|\mathbf{y}\|^2 = \mathbf{r}_0 \cdot \mathbf{y}. \tag{87}$$

From Eqs. (75) and (85) we can derive

$$\|\mathbf{r}_0 - \mathbf{A}\mathbf{z}\|^2 = \|\mathbf{r}_0\|^2 - \|\mathbf{y}\|^2. \tag{88}$$

Then, inserting Eq. (87) for $\mathbf{r}_0 \cdot \mathbf{y}$ into Eq. (21) we have

$$\|\mathbf{e}\|^2 = \|\mathbf{r}_0\|^2 - \|\mathbf{y}\|^2. \tag{89}$$

Comparing the above two equations we can derive Eq. (83). By using Eq. (29) and the definition of residual vector $\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}$ for Eq. (1), and from Eqs. (83) and (89) we have

$$\|\mathbf{r}\|^2 = \|\mathbf{e}\|^2 = \|\mathbf{r}_0\|^2 - \|\mathbf{y}\|^2. \tag{90}$$

Because of $\|\mathbf{y}\|^2 > 0$, Eq. (84) follows readily. This ends the proof of Eq. (84). □

Remark 4: In Section 3.2 we have solved two different optimization problems to find α_0 and α_i, $i = 1, \ldots, m$, and then the key equation (65) puts them the same value. That is, the two merit functions $\|\mathbf{e}\|^2$ and $\|\mathbf{r}\|^2$ are the same as shown in Eq. (90). More importantly, Eq. (84) guarantees that the residual error is absolutely decreased, while Eq. (75) gives the residual error estimation.

4 A numerical algorithm

Through the above derivations, the present *double optimal iterative algorithm* (DOIA) based on Theorem 1 can be summarized as follows.
(i) Select m and give an initial value of $\mathbf{x}_0$.
(ii) For $k = 0, 1, \ldots$, we repeat the following computations:

$$\begin{aligned}
&\mathbf{r}_k = \mathbf{b} - \mathbf{A}\mathbf{x}_k, \\
&\text{Arnoldi procedure to set up } \mathbf{u}_j^k,\ j = 1, \ldots, m,\ (\text{from } \mathbf{u}_1^k = \mathbf{A}\mathbf{r}_k / \|\mathbf{A}\mathbf{r}_k\|), \\
&\mathbf{U}_k = [\mathbf{u}_1^k, \ldots, \mathbf{u}_m^k], \\
&\mathbf{J}_k = \mathbf{A}\mathbf{U}_k, \\
&\mathbf{C}_k = \mathbf{J}_k^{\mathrm{T}}\mathbf{J}_k, \\
&\mathbf{D}_k = \mathbf{C}_k^{-1}, \\
&\mathbf{X}_k = \mathbf{U}_k\mathbf{D}_k\mathbf{J}_k^{\mathrm{T}}, \\
&\mathbf{E}_k = \mathbf{A}\mathbf{X}_k, \\
&\alpha_0^k = \frac{\mathbf{r}_k^{\mathrm{T}}(\mathbf{I}_n - \mathbf{E}_k)\mathbf{A}\mathbf{r}_k}{\mathbf{r}_k^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{E}_k)\mathbf{A}\mathbf{r}_k}, \\
&\mathbf{z}_k = \mathbf{X}_k\mathbf{r}_k + \alpha_0^k(\mathbf{r}_k - \mathbf{X}_k\mathbf{A}\mathbf{r}_k), \\
&\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{z}_k.
\end{aligned} \tag{91}$$

If $\mathbf{x}_{k+1}$ converges according to a given stopping criterion $\|\mathbf{r}_{k+1}\| < \varepsilon$, then stop; otherwise, go to step (ii).

Corollary 2: *In the algorithm DOIA, Theorem 3 guarantees that the residual is decreasing step-by-step, of which the residual vectors* $\mathbf{r}_{k+1}$ *and* $\mathbf{r}_k$ *have the following monotonically decreasing relations*:

$$\|\mathbf{r}_{k+1}\|^2 = \|\mathbf{r}_k\|^2 - \mathbf{r}_k^{\mathrm{T}}\mathbf{E}_k\mathbf{r}_k - \frac{[\mathbf{r}_k^{\mathrm{T}}(\mathbf{I}_n - \mathbf{E}_k)\mathbf{A}\mathbf{r}_k]^2}{\mathbf{r}_k^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{E}_k)\mathbf{A}\mathbf{r}_k}, \tag{92}$$

$$\|\mathbf{r}_{k+1}\| < \|\mathbf{r}_k\|. \tag{93}$$

Proof: For the DOIA, from Eqs. (90) and (85) by taking $\mathbf{r} = \mathbf{r}_{k+1}$, $\mathbf{r}_0 = \mathbf{r}_k$ and $\mathbf{y} = \mathbf{y}_k$ we have

$$\|\mathbf{r}_{k+1}\|^2 = \|\mathbf{r}_k\|^2 - \|\mathbf{y}_k\|^2, \tag{94}$$

$$\|\mathbf{y}_k\|^2 = \mathbf{r}_k^{\mathrm{T}}\mathbf{E}_k\mathbf{r}_k + \frac{[\mathbf{r}_k^{\mathrm{T}}(\mathbf{I}_n - \mathbf{E}_k)\mathbf{A}\mathbf{r}_k]^2}{\mathbf{r}_k^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{E}_k)\mathbf{A}\mathbf{r}_k}. \tag{95}$$

Inserting Eq. (95) into Eq. (94) we can derive Eq. (92), whereas Eq. (93) follows from Eq. (92) by noting that both $\mathbf{E}_k$ and $\mathbf{I}_n - \mathbf{E}_k$ are projection operators as shown in Eq. (82). □

Corollary 2 is very important, which guarantees that the algorithm DOIA is absolutely convergent step-by-step with a positive quantity $\|\mathbf{y}_k\|^2 > 0$ being decreased at each iteration step. By using Eqs. (3) and (19), $\|\mathbf{y}_k\|^2 = \|\mathbf{A}\mathbf{x}_k - \mathbf{A}\mathbf{x}_0\|^2$.

Corollary 3: *In the algorithm DOIA, the residual vector* $\mathbf{r}_{k+1}$ *is* $\mathbf{A}$*-orthogonal to the descent direction* $\mathbf{z}_k$, i.e.,

$$\mathbf{r}_{k+1} \cdot (\mathbf{A}\mathbf{z}_k) = 0. \tag{96}$$

Proof: From $\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}$ and Eqs. (29) and (19) we have

$$\mathbf{r} = \mathbf{r}_0 - \mathbf{y}. \tag{97}$$

Taking the inner product with $\mathbf{y}$ and using Eq. (87), it follows that

$$\mathbf{r} \cdot \mathbf{y} = \mathbf{r} \cdot (\mathbf{A}\mathbf{z}) = 0. \tag{98}$$

Letting $\mathbf{r} = \mathbf{r}_{k+1}$ and $\mathbf{z} = \mathbf{z}_k$, Eq. (96) is proven. □

The DOIA provides a good approximation of the residual Eq. (4) with a better descent direction $\mathbf{z}_k$ in the affine Krylov subspace. Under this situation we can prove the following corollary.

Corollary 4: *In the algorithm DOIA, the residual vectors* $\mathbf{r}_k$ *and* $\mathbf{r}_{k+1}$ *are nearly orthogonal*, i.e.,

$$\mathbf{r}_k \cdot \mathbf{r}_{k+1} \approx 0. \tag{99}$$

Moreover, the convergence rate is given by

$$\frac{\|\mathbf{r}_k\|}{\|\mathbf{r}_{k+1}\|} = \frac{1}{\sin\theta} > 1, \;\; 0 < \theta < \pi, \tag{100}$$

where θ *is the intersection angle between* $\mathbf{r}_k$ *and* $\mathbf{y}$.

Proof: First, in the DOIA, the residual Eq. (4) is approximately satisfied:

$$\mathbf{A}\mathbf{z}_k - \mathbf{r}_k \approx \mathbf{0}. \tag{101}$$

Taking the inner product with $\mathbf{r}_{k+1}$ and using Eq. (96), Eq. (99) is proven.

Next, from Eqs. (90) and (87) we have

$$\|\mathbf{r}_{k+1}\|^2 = \|\mathbf{r}_k\|^2 - \|\mathbf{r}_k\|\|\mathbf{y}\|\cos\theta, \tag{102}$$

where $0 < \theta < \pi$ is the intersection angle between $\mathbf{r}_k$ and $\mathbf{y}$. Again, with the help of Eq. (87) we also have

$$\|\mathbf{y}\| = \|\mathbf{r}_k\|\cos\theta. \tag{103}$$

Then, Eq. (102) can be further reduced to

$$\|\mathbf{r}_{k+1}\|^2 = \|\mathbf{r}_k\|^2(1-\cos^2\theta) = \|\mathbf{r}_k\|^2\sin^2\theta. \tag{104}$$

Taking the square roots of both sides we can obtain Eq. (100). □

Remark 5: For Eq. (95) in terms of the intersection angle ϕ between $(\mathbf{I}_n - \mathbf{E})\mathbf{r}_k$ and $(\mathbf{I}_n - \mathbf{E})\mathbf{A}\mathbf{r}_k$ we have

$$\|\mathbf{y}_k\|^2 = \mathbf{r}_k^{\mathrm{T}}\mathbf{E}\mathbf{r}_k + \|(\mathbf{I}_n - \mathbf{E})\mathbf{r}_k\|^2\cos^2\phi. \tag{105}$$

If $\phi = 0$, for example $\mathbf{r}_k$ is an eigenvector of $\mathbf{A}$, $\|\mathbf{y}_k\|^2 = \|\mathbf{r}_k\|^2$ can be deduced from Eq. (105) by

$$\|\mathbf{y}_k\|^2 = \mathbf{r}_k^{\mathrm{T}}\mathbf{E}\mathbf{r}_k + \mathbf{r}_k^{\mathrm{T}}(\mathbf{I}_n - \mathbf{E})^2\mathbf{r}_k = \mathbf{r}_k^{\mathrm{T}}\mathbf{E}\mathbf{r}_k + \mathbf{r}_k^{\mathrm{T}}(\mathbf{I}_n - \mathbf{E})\mathbf{r}_k = \|\mathbf{r}_k\|^2.$$

Then, by Eq. (94) the DOIA converges with one step more. On the other hand, if we take $m = n$, then the DOIA also converges with one step. We can see that if a suitable value of m is taken then the DOIA can converge within n steps. Therefore, we can use the following convergence criterion of the DOIA: If

$$\rho_N = \sum_{j=0}^{N}\|\mathbf{y}_j\|^2 \geq \|\mathbf{r}_0\|^2 - \varepsilon_1, \tag{106}$$

then the iterations in Eq. (91) terminate, where $N \leq n$. ε_1 is a given error tolerance.

Theorem 4: The square residual norm obtained by the algorithm which minimizes the merit function (7) in an m-dimensional Krylov subspace is denoted by $\|\mathbf{r}_{k+1}\|_{\mathrm{LS}}^2$. Then, we have

$$\|\mathbf{r}_{k+1}\|_{\mathrm{DOIA}}^2 < \|\mathbf{r}_{k+1}\|_{\mathrm{LS}}^2. \tag{107}$$

Proof: The algorithm which minimizes the merit function (7) in an m-dimensional Krylov subspace is a special case of the present theory with $\alpha_0 = 0$ [Liu (2013d)].

Refer Eqs. (8) and (9) as well as Remark 3 for definite. For this case, by Eq. (81) we have

$$\|\mathbf{r}_{k+1}\|_{\mathrm{LS}}^2 = \|\mathbf{r}_k\|^2 - \mathbf{r}_k^{\mathrm{T}}\mathbf{E}\mathbf{r}_k. \tag{108}$$

On the other hand, by Eqs. (94) and (95) we have

$$\|\mathbf{r}_{k+1}\|_{\mathrm{DOIA}}^2 = \|\mathbf{r}_k\|^2 - \mathbf{r}_k^{\mathrm{T}}\mathbf{E}\mathbf{r}_k - \frac{[\mathbf{r}_k^{\mathrm{T}}(\mathbf{I}_n - \mathbf{E})\mathbf{A}\mathbf{r}_k]^2}{\mathbf{r}_k^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{E})\mathbf{A}\mathbf{r}_k}. \tag{109}$$

Subtracting the above two equations we can derive

$$\|\mathbf{r}_{k+1}\|_{\mathrm{DOIA}}^2 = \|\mathbf{r}_{k+1}\|_{\mathrm{LS}}^2 - \frac{[\mathbf{r}_k^{\mathrm{T}}(\mathbf{I}_n - \mathbf{E})\mathbf{A}\mathbf{r}_k]^2}{\mathbf{r}_k^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{E})\mathbf{A}\mathbf{r}_k}, \tag{110}$$

which by Eq. (82) leads to Eq. (107). □

The algorithm DOIA is better than those algorithms which are based on the minimization in Eq. (7) in an m-dimensional Krylov subspace, including the FOM and GMRES.

5 A double optimal regularization method and algorithm

If $\mathbf{A}$ in Eq. (1) is a severely ill-conditioned matrix and the right-hand side data $\mathbf{b}$ are disturbanced by a large noise, we may encounter the problem that the numerical solution of Eq. (1) might deviate from the exact one to a great extent. Under this situation we have to solve system (1) by a *regularization method*. Hansen (1992) and Hansen and O'Leary (1993) have given an illuminating explanation that the Tikhonov regularization method to cope ill-posed linear problem is taking a trade-off between the size of the regularized solution and the quality to fit the given data by solving the following minimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left[\|\mathbf{b} - \mathbf{A}\mathbf{x}\|^2 + \beta\|\mathbf{x}\|^2\right]. \tag{111}$$

In this regularization theory a parameter β needs to be determined [Tikhonov and Arsenin (1977)].

In order to solve Eqs. (1) and (4) we propose a novel regularization method, instead of the Tikhonov regularization method, by minimizing

$$\min_{\mathbf{z}\in\mathbb{R}^n} \left\{ f = \frac{\|\mathbf{y}\|^2}{(\mathbf{r}_0 \cdot \mathbf{y})^2} + \beta\|\mathbf{z}\|^2 \right\} \tag{112}$$

to find the double optimal regularization solution of $\mathbf{z}$, where $\mathbf{y} = \mathbf{A}\mathbf{z}$. In Theorem 4 we have proved that the above minimization of the first term is better than that of the minimization of $\|\mathbf{r}_0 - \mathbf{A}\mathbf{z}\|^2$.

We can derive the following result as an approximate solution of Eq. (112).

Theorem 5: *Under the double optimal solution of* $\mathbf{z} \in \mathscr{K}'_m$ *given in Theorem 1, an approximate solution of Eq. (112), denoted by* $\mathbf{Z}$*, is given by*

$$\gamma = \frac{1}{(\beta \|\mathbf{z}\|^2 \|\mathbf{A}\mathbf{z}\|^2)^{1/4}}, \tag{113}$$

$$\mathbf{Z} = \gamma \mathbf{z}. \tag{114}$$

Proof: The first term in f in Eq. (112) is scaling invariant, which means that if $\mathbf{y}$ is a solution, then $\gamma\mathbf{y}$, $\gamma \neq 0$, is also a solution. Let $\mathbf{z}$ be a solution given in Theorem 1, which is a double optimal solution of Eq. (112) with $\beta = 0$. We suppose that an approximate solution of Eq. (112) with $\beta > 0$ is given by $\mathbf{Z} = \gamma\mathbf{z}$, where γ is to be determined.

By using Eq. (87), f can be written as

$$f = \frac{1}{\|\mathbf{y}\|^2} + \beta \|\mathbf{Z}\|^2, \tag{115}$$

which upon using $\mathbf{Z} = \gamma\mathbf{z}$ and $\mathbf{y} = \mathbf{A}\mathbf{Z} = \gamma\mathbf{A}\mathbf{z}$ becomes

$$f = \frac{1}{\gamma^2 \|\mathbf{A}\mathbf{z}\|^2} + \beta\gamma^2 \|\mathbf{z}\|^2. \tag{116}$$

Taking the differential of f with respect to γ and setting the resultant equal to zero we can derive Eq. (113). □

The numerical algorithm based on Theorem 5 is labeled a *double optimal regularization algorithm* (DORA), which can be summarized as follows.

(i) Select β, m and give an initial value of $\mathbf{x}_0$.

(ii) For $k = 0, 1, \ldots$, we repeat the following computations:

$$
\begin{aligned}
&\mathbf{r}_k = \mathbf{b} - \mathbf{A}\mathbf{x}_k, \\
&\text{Arnoldi procedure to set up } \mathbf{u}_j^k,\ j = 1, \ldots, m,\ (\text{from } \mathbf{u}_1^k = \mathbf{A}\mathbf{r}_k / \|\mathbf{A}\mathbf{r}_k\|), \\
&\mathbf{U}_k = [\mathbf{u}_1^k, \ldots, \mathbf{u}_m^k], \\
&\mathbf{J}_k = \mathbf{A}\mathbf{U}_k, \\
&\mathbf{C}_k = \mathbf{J}_k^{\mathrm{T}}\mathbf{J}_k, \\
&\mathbf{D}_k = \mathbf{C}_k^{-1}, \\
&\mathbf{X}_k = \mathbf{U}_k\mathbf{D}_k\mathbf{J}_k^{\mathrm{T}}, \\
&\mathbf{E}_k = \mathbf{A}\mathbf{X}_k, \\
&\alpha_0^k = \frac{\mathbf{r}_k^{\mathrm{T}}(\mathbf{I}_n - \mathbf{E}_k)\mathbf{A}\mathbf{r}_k}{\mathbf{r}_k^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}(\mathbf{I}_n - \mathbf{E}_k)\mathbf{A}\mathbf{r}_k}, \\
&\mathbf{z}_k = \mathbf{X}_k\mathbf{r}_k + \alpha_0^k(\mathbf{r}_k - \mathbf{X}_k\mathbf{A}\mathbf{r}_k), \\
&\gamma_k = \frac{1}{(\beta\|\mathbf{z}_k\|^2\|\mathbf{A}\mathbf{z}_k\|^2)^{1/4}}, \\
&\mathbf{x}_{k+1} = \mathbf{x}_k + \gamma_k\mathbf{z}_k.
\end{aligned}
\tag{117}
$$

If $\mathbf{x}_{k+1}$ converges according to a given stopping criterion $\|\mathbf{r}_{k+1}\| < \varepsilon$, then stop; otherwise, go to step (ii). It is better to select the value of regularization parameter β in a range such that the value of γ_k is O(1). We can view γ_k as a dynamical relaxation factor.

6 Numerical examples

In order to evaluate the performance of the newly developed algorithms DOIA and DORA, we test some linear problems and compare the numerical results with that obtained by the FOM and GMRES.

6.1 Example 1

First we consider Eq. (1) with the following cyclic coefficient matrix:

$$
\mathbf{A} = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 \\
2 & 3 & 4 & 5 & 6 & 1 \\
3 & 4 & 5 & 6 & 1 & 2 \\
4 & 5 & 6 & 1 & 2 & 3 \\
5 & 6 & 1 & 2 & 3 & 4 \\
6 & 1 & 2 & 3 & 4 & 5
\end{bmatrix},
\tag{118}
$$

where the right-hand side of Eq. (1) is supposed to be $b_i = i^2,\ i = 1,\ldots,6$. We use the QR method to find the exact solutions of $x_i,\ i = 1,\ldots,6$, which are plotted in Fig. 1 by solid black line.

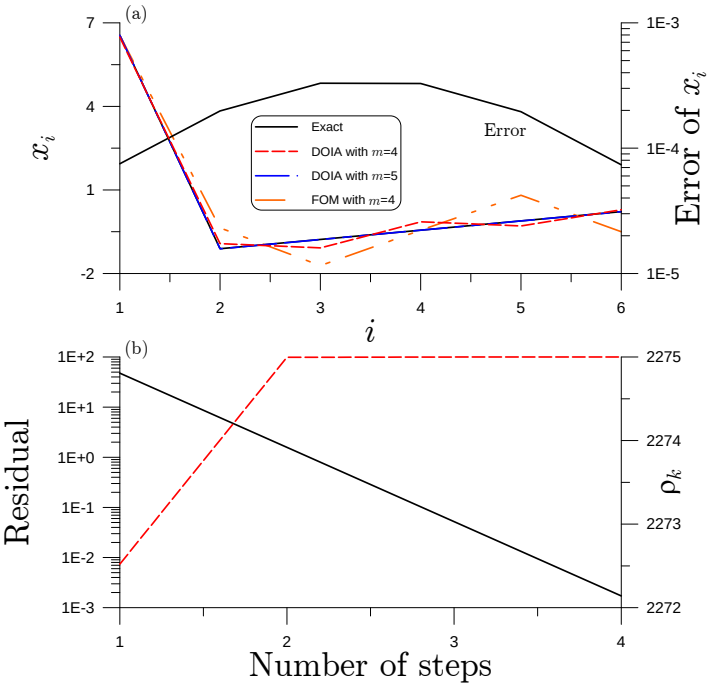

Figure 1: For example 1 solved by optimal solutions of DOIA and FOM, (a) comparing with exact solution and numerical error, and (b) residual and ρ_k.

In order to evaluate the performance of the DOIA, we use this example to demonstrate the idea of double optimal solution. When we take $\mathbf{x}_0 = \mathbf{0}$ without iterating the solution of $\mathbf{z}$ is just the solution $\mathbf{x}$ of Eq. (1). We take resrectively $m = 4$ and $m = 5$ in the DOIA double optimal solutions, which as shown in Fig. 1 are close to the exact one. Basically, when $m = 5$ the double optimal solution is equal to the exact one. We can also see that the double optimal solution with $m = 4$ is already close to the exact one. We apply the same idea to the FOM solution with $m = 4$, which is less accurate than that obtained by the DOIA with $m = 4$.

Now we apply the DOIA under the convergence criterion in Eq. (106) to solve this problem, where we take $m = 4$ and $\varepsilon_1 = 10^{-8}$. With four steps $N = 4$ we can obtain numerical solution whose error is plotted in Fig. 1(a) by solid black line, which is quite accurate with the maximum error being smaller than 3.3×10^{-4}. The residual

is plotted in Fig. 1(b) by solid black line. The value of square initial residual is $\|\mathbf{r}_0\|^2 = 2275$ and the term ρ_k fast tends to 2275 at the first two steps, and then at $N = 4$ the value of ρ_N is over $\|\mathbf{r}_0\|^2 - \varepsilon_1$.

6.2 Example 2

When we apply a central difference scheme to the following two-point boundary value problem:

$$\begin{aligned} &-u''(x) = f(x), \ \ 0 < x < 1, \\ &u(0) = 1, \ \ u(1) = 2, \end{aligned} \tag{119}$$

we can derive a linear equations system:

$$\mathbf{Au} = \begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & \cdot & \cdot & \cdot \\ & & & & -1 & 2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} (\Delta x)^2 f(\Delta x) + a \\ (\Delta x)^2 f(2\Delta x) \\ \vdots \\ (\Delta x)^2 f((n-1)\Delta x) \\ (\Delta x)^2 f(n\Delta x) + b \end{bmatrix}, \tag{120}$$

where $\Delta x = 1/(n+1)$ is the spatial length, and $u_i = u(i\Delta x), \ i = 1, \ldots, n$, are unknown values of $u(x)$ at the grid points $x_i = i\Delta x$. $u_0 = a$ and $u_{n+1} = b$ are the given boundary conditions. The above matrix $\mathbf{A}$ is known as a central difference matrix, which is a typical sparse matrix. In the applications of FOM, GMRES and DOIA we take $n = 99$ and $m = 10$, of which the condition number of $\mathbf{A}$ is about 4052.

The numerical solutions obtained by the FOM, GMRES and DOIA are compared with the exact solution:

$$u(x) = 1 + x + \frac{1}{\pi^2} \sin \pi x. \tag{121}$$

The residuals and numerical errors are compared in Figs. 2(a) and 2(b), whose maximum errors are the same value 8.32×10^{-6}. Under the convergence criterion with $\varepsilon = 10^{-10}$, the FOM converges with 414 steps, the GMRES converges with 379 steps, and the DOIA is convergent with 322 steps. As shown in Fig. 2(a) the DOIA is convergent faster than the GMRES and FOM; however, these three algorithms give the numerical solution having the same accuracy. The orthogonalities specified in Eqs. (96) and (99) for the DOIA are plotted in Fig. 2(c).

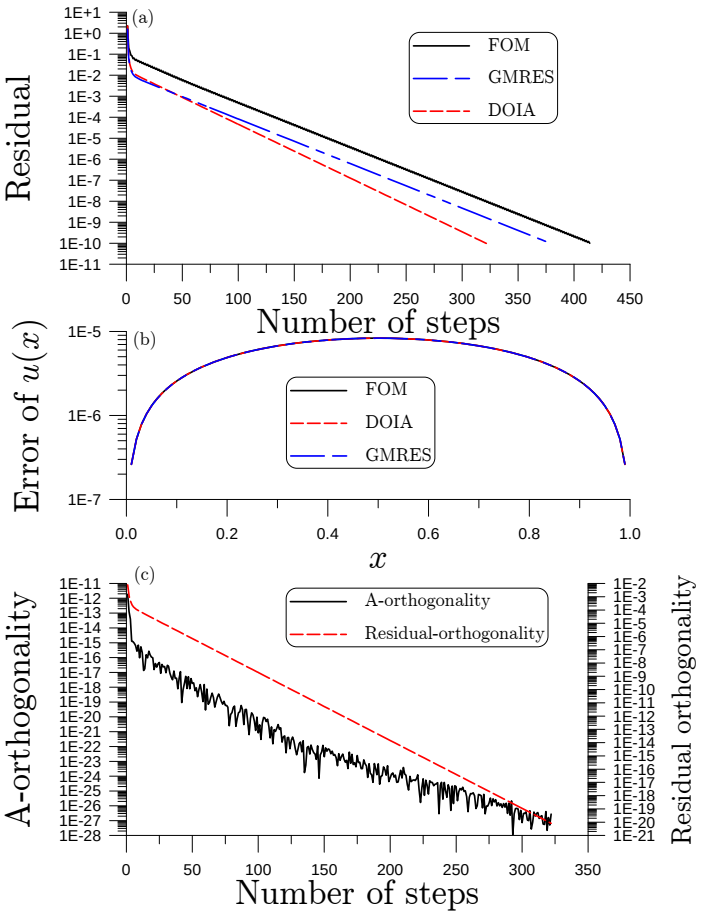

Figure 2: For example 2 solved by the FOM, GMRES and DOIA, comparing (a) residuals, (b) numerical errors, and (c) proving orthogonalities of DOIA.

6.3 Example 3

Finding an n-order polynomial function $p(x) = a_0 + a_1 x + \ldots + a_n x^n$ to best match a continuous function $f(x)$ in the interval of $x \in [0,1]$:

$$\min_{\deg(p)\le n} \int_0^1 [f(x) - p(x)]^2 dx, \tag{122}$$

leads to a problem governed by Eq. (1), where $\mathbf{A}$ is the $(n+1)\times(n+1)$ Hilbert matrix defined by

$$A_{ij}=\frac{1}{i+j-1}, \tag{123}$$

$\mathbf{x}$ is composed of the $n+1$ coefficients $a_0,a_1,\ldots,a_n$ appeared in $p(x)$, and

$$\mathbf{b}=\begin{bmatrix} \int_0^1 f(x)dx \\ \int_0^1 xf(x)dx \\ \vdots \\ \int_0^1 x^n f(x)dx \end{bmatrix} \tag{124}$$

is uniquely determined by the function $f(x)$.

The Hilbert matrix is a notorious example of highly ill-conditioned matrices. Eq. (1) with the matrix $\mathbf{A}$ having a large condition number usually displays that an arbitrarily small perturbation of data on the right-hand side may lead to an arbitrarily large perturbation to the solution on the left-hand side.

In this example we consider a highly ill-conditioned linear system (1) with $\mathbf{A}$ given by Eq. (123). The ill-posedness of Eq. (1) increases fast with n. We consider an exact solution with $x_j=1,\ j=1,\ldots,n$ and b_i is given by

$$b_i=\sum_{j=1}^{n}\frac{1}{i+j-1}+\sigma R(i), \tag{125}$$

where $R(i)$ are random numbers between $[-1,1]$.

First, a noise with intensity $\sigma=10^{-6}$ is added on the right-hand side data. For $n=300$ we take $m=5$ for both FOM and DOIA. Under $\varepsilon=10^{-3}$, both FOM and DOIA are convergent with 3 steps. The maximum error obtained by the FOM is 0.037, while that obtained by the DOIA is 0.0144. In Fig. 3 we show the numerical results, of which the accuracy of DOIA is good, although the ill-posedness of the linear Hilbert problem $n=300$ is highly increased.

Then we raise the noise to $\sigma=10^{-3}$, of which we find that both the GMRES and DOIA under the above convergence $\varepsilon=10^{-3}$ and $m=5$ lead to failure solutions. So we take $\varepsilon=10^{-1}$ for the GMRES, DOIA and DORA, where $\beta=0.00015$ is used in the DORA. The GMRES runs two steps as shown in Fig. 4(a) by dashed-dotted line, and the maximum error as shown in Fig. 4(b) by dashed-dotted line is 0.5178. The DOIA also runs with two iterations as shown in Fig. 4(a) by dashed line, and the maximum error as shown in Fig. 4(b) by dashed line is reduced to

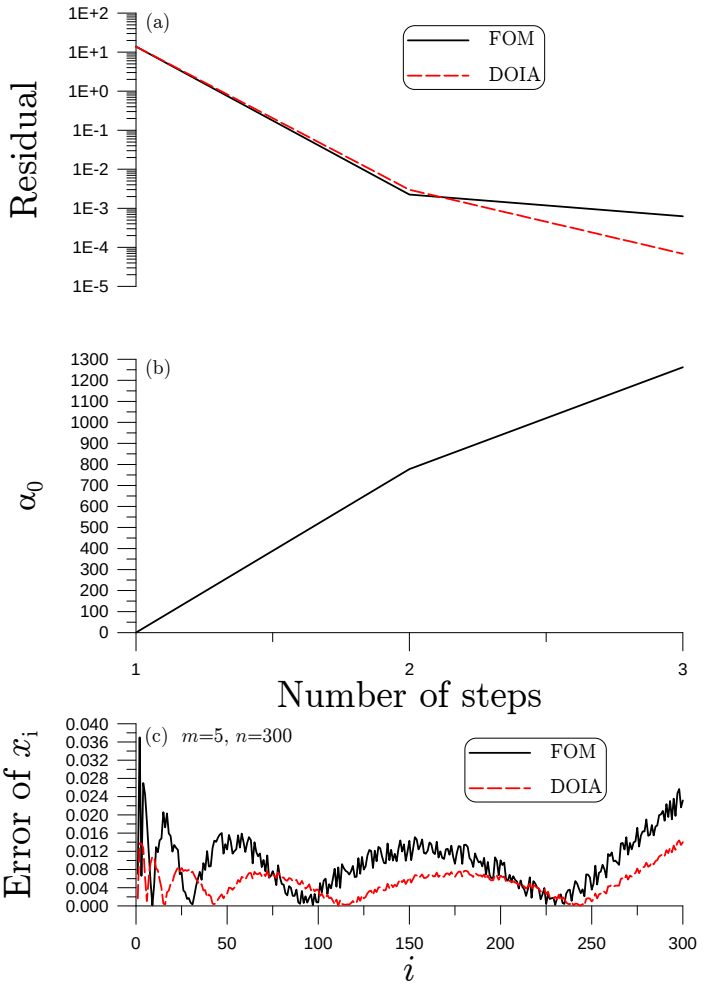

Figure 3: For example 3 solved by the FOM and DOIA, (a) residuals, (b) showing α_0, and (c) numerical errors.

0.1417. It is interesting that although the DORA runs 49 iterations as shown in Fig. 4(a) by solid line, the maximum error as shown in Fig. 4(b) by solid line is largely reduced to 0.0599. For this highly noised case the DOIA is better than the GMRES, while the DORA is better than the DOIA. It can be seen that the improvement of regularization by the DORA is obvious.

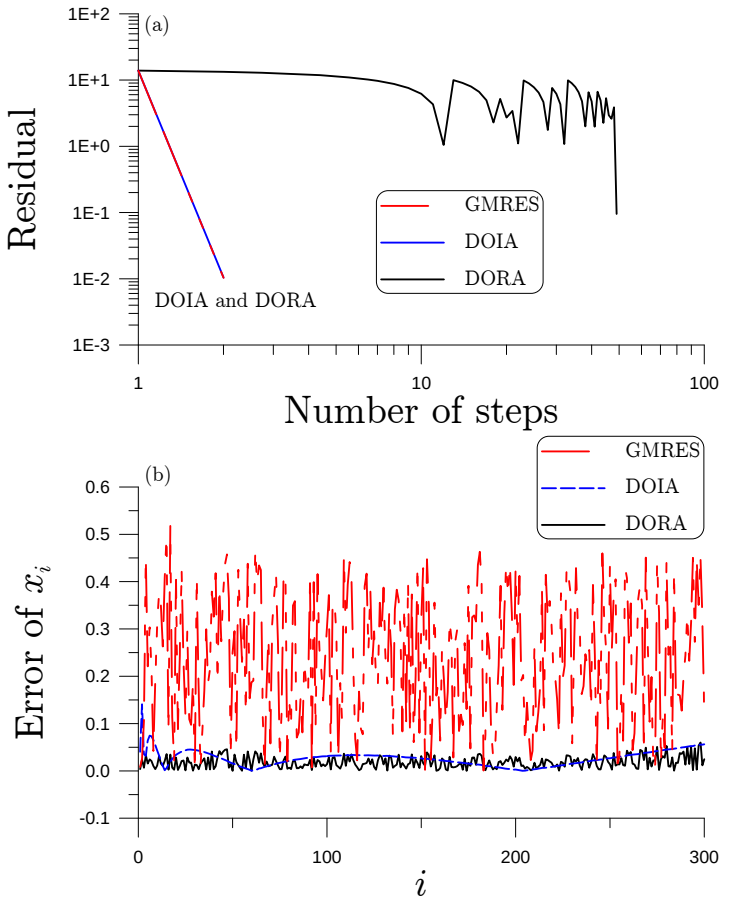

Figure 4: For example 3 under a large noise solved by the GMRES, DOIA and DORA, (a) residuals, and (b) numerical errors.

6.4 Example 4

When the backward heat conduction problem (BHCP) is considered in a spatial interval of $0 < x < \ell$ by subjecting to the boundary conditions at two ends of a slab:

$$\begin{aligned} &u_t(x,t) = u_{xx}(x,t), \quad 0 < t < T, \quad 0 < x < \ell, \\ &u(0,t) = u_0(t), \quad u(\ell,t) = u_\ell(t), \end{aligned} \tag{126}$$

we solve u under a final time condition:

$$u(x,T) = u^T(x). \tag{127}$$

The fundamental solution of Eq. (126) is by

$$K(x,t)=\frac{H(t)}{2\sqrt{\pi t}}\exp\left(\frac{-x^2}{4t}\right), \tag{128}$$

where $H(t)$ is the Heaviside function.

In the MFS the solution of u at the field point $\mathbf{p}=(x,t)$ can be expressed as a linear combination of the fundamental solutions $U(\mathbf{p},\mathbf{s}_j)$:

$$u(\mathbf{p})=\sum_{j=1}^{n} c_j U(\mathbf{p},\mathbf{s}_j),\ \ \mathbf{s}_j=(\eta_j,\tau_j)\in\Omega^c, \tag{129}$$

where n is the number of source points, c_j are unknown coefficients, and $\mathbf{s}_j$ are source points being located in the complement Ω^c of $\Omega=[0,\ell]\times[0,T]$. For the heat conduction equation we have the basis functions

$$U(\mathbf{p},\mathbf{s}_j)=K(x-\eta_j,t-\tau_j). \tag{130}$$

It is known that the location of source points in the MFS has a great influence on the accuracy and stability. In a practical application of MFS to solve the BHCP, the source points are uniformly located on two vertical straight lines parallel to the t-axis, not over the final time, which was adopted by Hon and Li (2009) and Liu (2011), showing a large improvement than the line location of source points below the initial time. After imposing the boundary conditions and the final time condition to Eq. (129) we can obtain a linear equations system (1) with

$$\begin{aligned} &A_{ij}=U(\mathbf{p}_i,\mathbf{s}_j),\ \ \mathbf{x}=(c_1,\cdots,c_n)^{\mathrm{T}},\\ &\mathbf{b}=(u_\ell(t_i),\ i=1,\ldots,m_1;u^T(x_j),\ j=1,\ldots,m_2;u_0(t_k),\ k=m_1,\ldots,1)^{\mathrm{T}}, \end{aligned} \tag{131}$$

and $n=2m_1+m_2$.

Since the BHCP is highly ill-posed, the ill-condition of the coefficient matrix $\mathbf{A}$ in Eq. (1) is serious. To overcome the ill-posedness of Eq. (1) we can use the DOIA and DORA to solve this problem. Here we compare the numerical solution with an exact solution:

$$u(x,t)=\cos(\pi x)\exp(-\pi^2 t).$$

For the case with $T=1$ the value of final time data is in the order of 10^{-4}, which is small by comparing with the value of the initial temperature $f(x)=u_0(x)=\cos(\pi x)$ to be retrieved, which is $O(1)$. First we impose a relative random noise with an intensity $\sigma=10\%$ on the final time data. Under the following parameters $m_1=15$,

$m_2 = 8$, $m = 16$, and $\varepsilon = 10^{-2}$, we solve this problem by the FOM, GMRES and DOIA. With two iterations the FOM is convergent as shown Fig. 5(a); however, the numerical error as shown in Fig. 5(b) is quite large, with the maximum error being 0.264. It means that the FOM is failure for this inverse problem. The DOIA is convergent with five steps, and its maximum error is 0.014, while the GMRES is convergent with 16 steps and with the maximum error being 0.148. It can be seen that the present DOIA converges very fast and is very robust against noise, and we can provide a very accurate numerical result of the BHCP by using the DOIA.

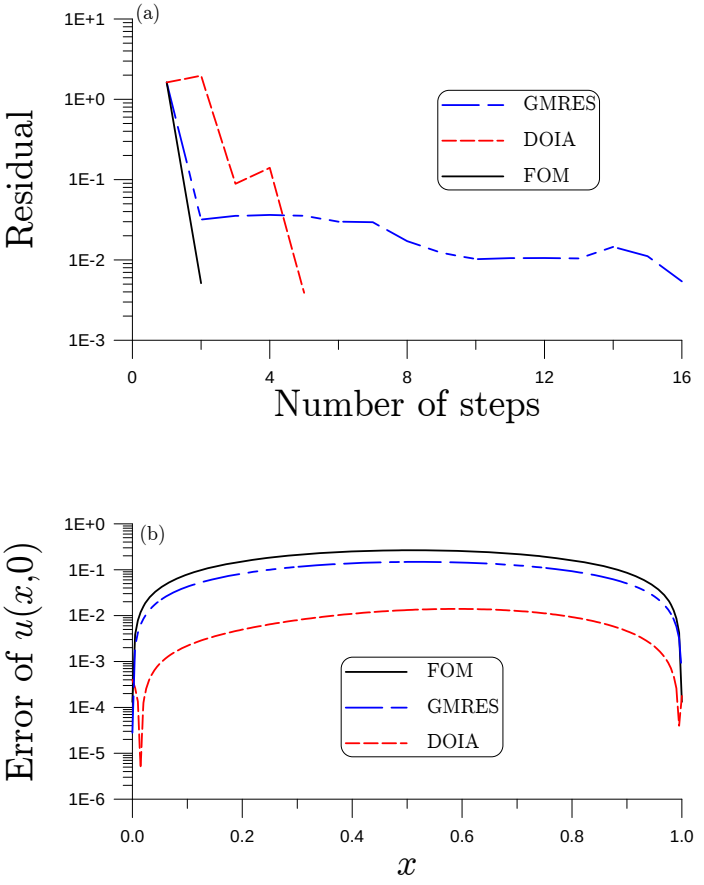

Figure 5: For example 4 solved by the FOM, GMRES and DOIA, comparing (a) residuals, and (b) numerical errors.

Next, we come to a very highly ill-posed case with $T = 5$ and $\sigma = 100\%$. When the final time data are in the order of 10^{-21}, we attempt to recover the initial tempera-

ture $f(x) = \cos(\pi x)$ which is in the order of 10^0. Under the following parameters $m_1 = 10$, $m_2 = 8$, $m = 16$, and $\varepsilon = 10^{-4}$ and $\varepsilon_1 = 10^{-8}$, we solve this problem by the DOIA and DORA, where $\beta = 0.4$ is used in the DORA. We let DOIA and DORA run 100 steps, because they do not converge under the above convegence criterion $\varepsilon = 10^{-4}$ as shown in Fig. 6(a). Due to a large value of $\beta = 0.4$, the residual curve of DORA is quite different from that of the DOIA. The numerical results of $u(x,0)$ are compared with the exact one $f(x) = \cos(\pi x)$ in Fig. 6(b), whose maximum error of the DOIA is about 0.2786, while that of the DORA is about 0.183. It can be seen that the improvement is obvious.

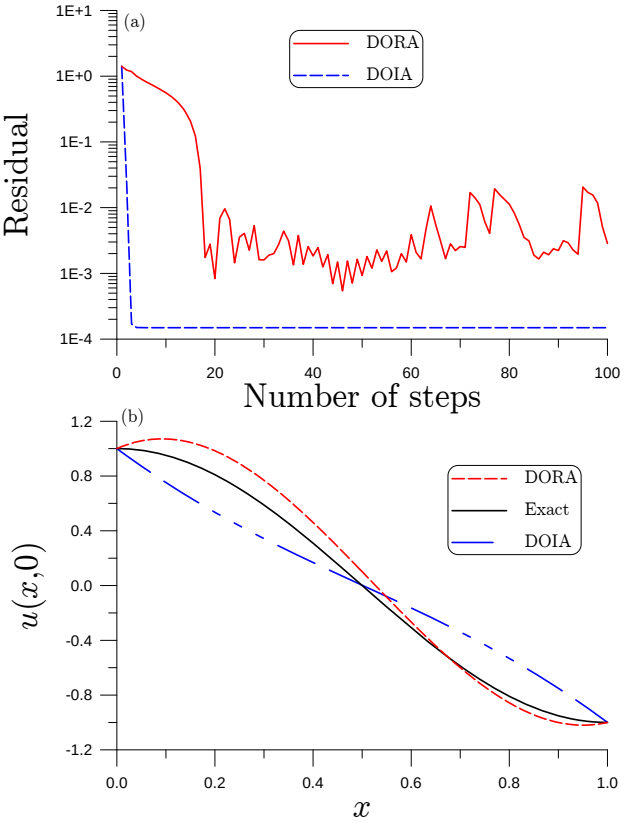

Figure 6: For example 4 under large final time and large noise solved by the DOIA and DORA, comparing (a) residuals, (b) numerical solutions and exact solution, and (c) numerical errors.

6.5 Example 5

Let us consider the inverse Cauchy problem for the Laplace equation:

$$\Delta u = u_{rr} + \frac{1}{r}u_r + \frac{1}{r^2}u_{\theta\theta} = 0, \tag{132}$$

$$u(\rho,\theta) = h(\theta),\ \ 0 \le \theta \le \pi, \tag{133}$$

$$u_n(\rho,\theta) = g(\theta),\ \ 0 \le \theta \le \pi, \tag{134}$$

where $h(\theta)$ and $g(\theta)$ are given functions. The inverse Cauchy problem is specified as follows: Seeking an unknown boundary function $f(\theta)$ on the part $\Gamma_2 := \{(r,\theta)|r = \rho(\theta),\ \pi < \theta < 2\pi\}$ of the boundary under Eqs. (132)-(134) with the overspecified data being given on $\Gamma_1 := \{(r,\theta)|r = \rho(\theta),\ 0 \le \theta \le \pi\}$. It is well known that the inverse Cauchy problem is a highly ill-posed problem. In the past, almost all of the checks to the ill-posedness of the inverse Cauchy problem, the illustrating examples have led to that the inverse Cauchy problem is actually severely ill-posed. Belgacem (2007) has provided an answer to the ill-posedness degree of the inverse Cauchy problem by using the theory of kernel operators. The foundation of his proof is the Steklov-Poincaré approach introduced in Belgacem and El Fekih (2005).

The method of fundamental solutions (MFS) can be used to solve the Laplace equation, of which the solution of u at the field point $\mathbf{p} = (r\cos\theta, r\sin\theta)$ can be expressed as a linear combination of fundamental solutions $U(\mathbf{p},\mathbf{s}_j)$:

$$u(\mathbf{p}) = \sum_{j=1}^{n} c_j U(\mathbf{p},\mathbf{s}_j),\ \ \mathbf{s}_j \in \Omega^c. \tag{135}$$

For the Laplace equation (132) we have the fundamental solutions:

$$U(\mathbf{p},\mathbf{s}_j) = \ln r_j,\ \ r_j = \|\mathbf{p}-\mathbf{s}_j\|. \tag{136}$$

In the practical application of MFS, by imposing the boundary conditions (133) and (134) at N points on Eq. (135) we can obtain a linear equations system (1) with

$$\mathbf{p}_i = (p_i^1, p_i^2) = (\rho(\theta_i)\cos\theta_i, \rho(\theta_i)\sin\theta_i),$$
$$\mathbf{s}_j = (s_j^1, s_j^2) = (R(\theta_j)\cos\theta_j, R(\theta_j)\sin\theta_j),$$
$$A_{ij} = \ln\|\mathbf{p}_i - \mathbf{s}_j\|,\ \ \text{if } i \text{ is odd},$$
$$A_{ij} = \frac{\eta(\theta_i)}{\|\mathbf{p}_i-\mathbf{s}_j\|^2}\left(\rho(\theta_i) - s_j^1\cos\theta_i - s_j^2\sin\theta_i - \frac{\rho'(\theta_i)}{\rho(\theta_i)}[s_j^1\sin\theta_i - s_j^2\cos\theta_i]\right), \text{if } i \text{ is even},$$
$$\mathbf{x} = (c_1,\ldots,c_n)^{\mathrm{T}},\ \ \mathbf{b} = (h(\theta_1), g(\theta_1),\ldots,h(\theta_N), g(\theta_N))^{\mathrm{T}},$$

(137)

in which $n = 2N$, and

$$\eta(\theta) = \frac{\rho(\theta)}{\sqrt{\rho^2(\theta) + [\rho'(\theta)]^2}}. \tag{138}$$

The above $R(\theta) = \rho(\theta) + D$ with an offset D can be used to locate the source points along a contour with a radius $R(\theta)$.

For the purpose of comparison we consider the following exact solution:

$$u(x,y) = \cos x \cosh y + \sin x \sinh y, \tag{139}$$

defined in a domain with a complex amoeba-like irregular shape as a boundary:

$$\rho(\theta) = \exp(\sin\theta)\sin^2(2\theta) + \exp(\cos\theta)\cos^2(2\theta). \tag{140}$$

We solve this problem by the DOIA and DORA with $n = 2N = 40$ and $m = 10$, where the noise being imposed on the measured data h and g is quite large with $\sigma = 0.3$, and $\beta = 0.0003$ is used in the DORA. Through 10 iterations for both the DOIA and DORA the residuals are shown in Fig. 7(a). The numerical solutions and exact solution are compared in Fig. 7(b). It can be seen that the DOIA and DORA can accurately recover the unknown boundary condition. As shown in Fig. 7(c), when the DOIA has the maximum error 0.381, the DORA has the maximum error 0.253. We also apply the GMRES with $m = 10$ to solve this problem; however, as shown in Fig. 7 it is failure, whose maximum error is 2.7.

Next we consider a large noise with $\sigma = 40\%$. The value of β used in the DORA is changed to $\beta = 0.000095$. Through 100 iterations for both the DOIA and DORA the residuals are shown in Fig. 8(a). The numerical solutions and exact solution are compared in Fig. 8(b). As shown in Fig. 8(c), when the DOIA has the maximum error 0.5417, the DORA has the maximum error 0.2737. The improvement of the accuracy by using the DORA than the DOIA is obvious.

Accordingly, we can observe that the algorithms DOIA and DORA can deal with the Cauchy problem of the Laplace equation in a domain with a complex amoeba-like irregular shape even under very large noises up to $\sigma = 30\%$ and $\sigma = 40\%$, and yield much better results than that obtained by the GMRES.

6.6 Example 6

One famous mesh-less numerical method used in the data interpolation for two-dimensional function $u(x,y)$ is the radial basis function (RBF) method, which expands u by

$$u(x,y) = \sum_{k=1}^{n} a_k \phi_k, \tag{141}$$

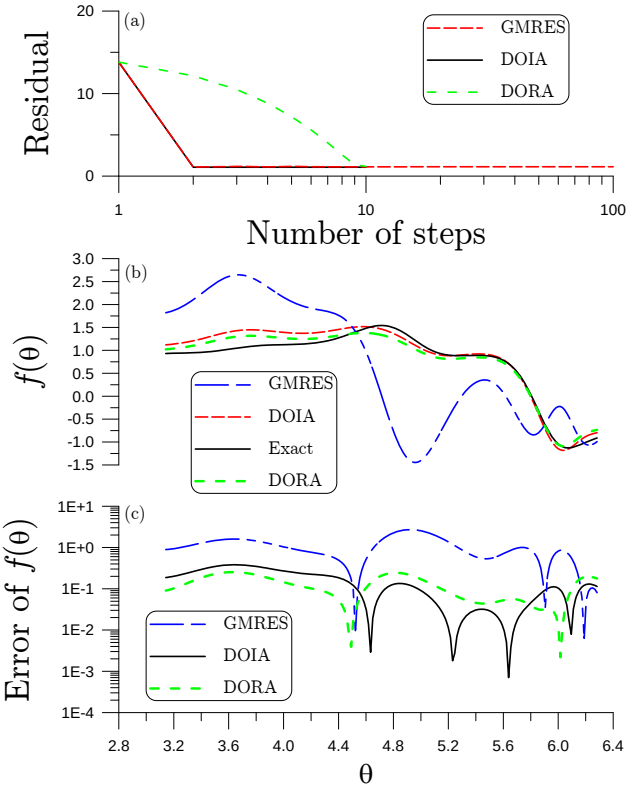

Figure 7: For example 5 solved by the GMRES, DOIA and DORA, comparing (a) residuals, (b) numerical solutions and exact solution, and (c) numerical errors.

where a_k are the expansion coefficients to be determined and ϕ_k is a set of RBFs, for example,

$$\begin{aligned}
&\phi_k = (r_k^2 + c^2)^{N-3/2},\ N = 1,2,\ldots,\\
&\phi_k = r_k^{2N} \ln r_k,\ N = 1,2,\ldots,\\
&\phi_k = \exp\left(-\frac{r_k^2}{a^2}\right),\\
&\phi_k = (r_k^2 + c^2)^{N-3/2} \exp\left(-\frac{r_k^2}{a^2}\right),\ N = 1,2,\ldots,
\end{aligned} \tag{142}$$

where the radius function r_k is given by $r_k = \sqrt{(x-x_k)^2 + (y-y_k)^2}$, while (x_k, y_k), $k = 1,\ldots,n$ are called source points. The constants a and c are shape parameters. In

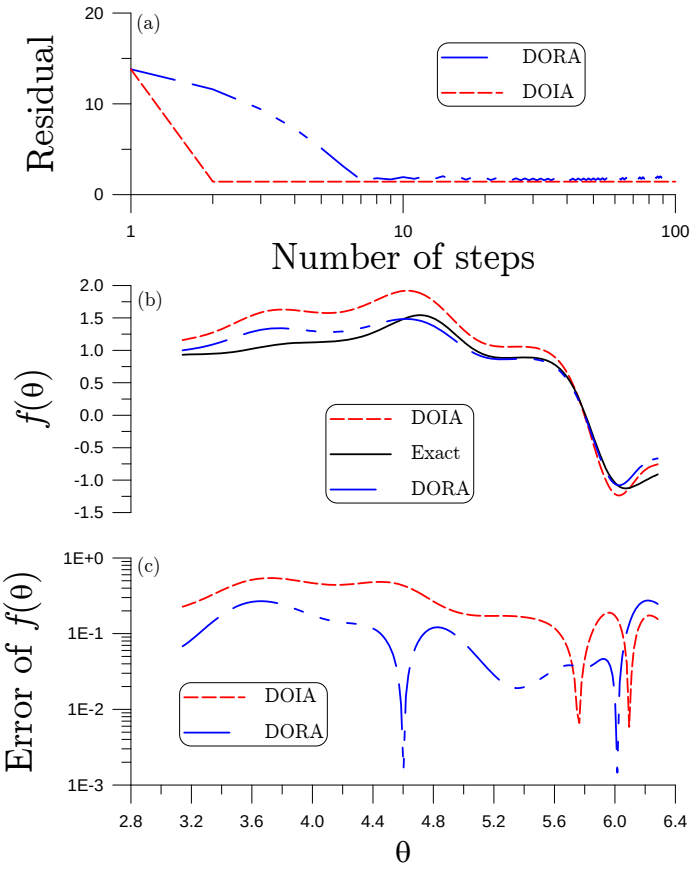

Figure 8: For example 5 under large noise solved by the DOIA and DORA, comparing (a) residuals, (b) numerical solutions and exact solution, and (c) numerical errors.

the below we take the first set of ϕ_k as trial functions, with $N = 2$ which is known as a multi-quadric RBF [Golberg and Chen (1996); Cheng, Golberg and Kansa (2003)]. There are some discussions about the optimal shape factor c used in the MQ-RBF [Huang, Lee and Cheng (2007); Huang, Yen and Cheng (2010); Bayona, Moscoso and Kindelan (2011); and Cheng (2012)].

Let $\Omega := \{(x,y)|\sqrt{x^2+y^2} \leq \rho(\theta)\}$ be the domain of data interpolation, where $\rho(\theta)$ is the boundary shape function. By collocating n points in Ω to match the given data and using Eq. (141) we can derive a linear system (1) with the coefficient matrix

given by

$$A_{ij} = \sqrt{r_{ij}^2 + c^2}, \tag{143}$$

where $r_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$, and (x_i, y_i), $i = 1, \ldots, n$ are interpolated points. Usually the shape factor c is a fixed constant, whose value is sensitive to the problem we attempt to solve.

We solve a quite difficult and well known interpolation problem of Franke function [Franke (1982)]:

$$\begin{aligned} u(x,y) = & \frac{3}{4}\exp\left(-\frac{(9x-2)^2+(9y-2)^2}{4}\right) + \frac{3}{4}\exp\left(-\frac{(9x+1)^2}{49} - \frac{(9y+1)^2}{10}\right) \\ & + \frac{1}{2}\exp\left(-\frac{(9x-7)^2+(9y-3)^2}{4}\right) - \frac{1}{5}\exp[-(9x-4)^2 - (9y-7)^2] \end{aligned} \tag{144}$$

on the unit square. We apply the DOIA with $m = 5$ and $c = 0.3$ to solve this problem under the convergence criterion $\varepsilon = 0.1$, where we take $\Delta x = \Delta y = 1/9$ to be a uniform spacing of distribution of source points as that used in Fasshauer (2002). The residual is shown in Fig. 9(a), which is convergent with 22 iterations, and the numerical error is shown in Fig. 9(b). The maximum error 0.0576 obtained by the DOIA is better than 0.1556 obtained by Fasshauer (2002).

7 Conclusions

In an affine m-dimensional Krylov subspace we have derived a closed-form double optimal solution of the n-dimensional linear residual equation (4), which was obtained by optimizing two merit functions in Eqs. (27) and (30). The main properties were analyzed, and a key equation was proven to link these two optimizations. Based on the double optimal solution, the iterative algorithm DOIA was developed, which has an **A**-orthogonal property, and is proven to be absolutely convergent step-by-step with the square residual error being reduced by a positive quantity $\|\mathbf{A}\mathbf{z}_k\|^2$ at each iteration step. We have proved that the residual error obtained by the algorithm DOIA is smaller than that obtained by other algorithms which are based on the minimization of the square residual error $\|\mathbf{b} - \mathbf{A}\mathbf{x}\|^2$, including the FOM and GMRES. We developed as well a simple double optimal regularization algorithm (DORA) to tackle the ill-posed linear problem under a large noise, and as compared with the GMRES and DOIA the regularization effect obtained by the DORA is obvious. Because the computational costs of DOIA and DORA are very inexpensive with the need of only inverting an $m \times m$ matrix one time at each iterative

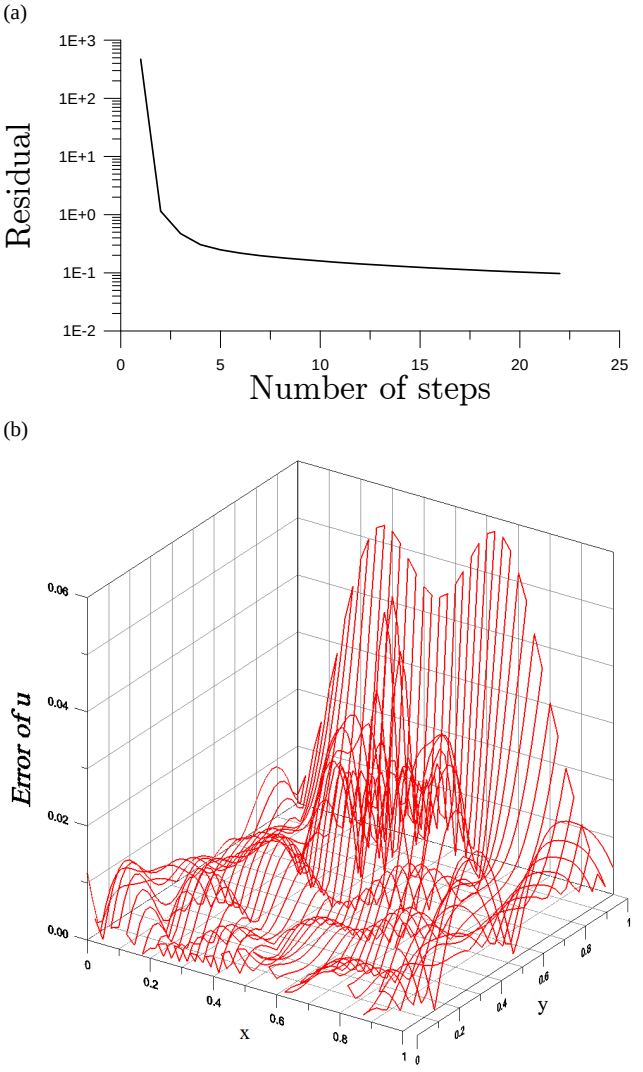

Figure 9: For the data interpolation of Franke function, showing (a) residual, and (b) numerical error computed by the DOIA.

step, they are very useful to solve a large scale system of linear equations with an ill-conditioned coefficient matrix. Even when we imposed a large noise on the ill-posed linear inverse problem, the DOIA and DORA were robust against large noise, but both the FOM and GMRES were not workable.

Acknowledgement: Highly appreciated are the project NSC-102-2221-E-002-125-MY3 and the 2011 Outstanding Research Award from the National Science Council of Taiwan to the first author.The work of the second author is supported by a UCI/ARL colloborative research grant.

References

Bayona, V.; Moscoso, M.; Kindelan, M. (2011): Optimal constant shape parameter for multiquadric based RBF-FD method. *J. Comput. Phys.*, vol. 230, pp. 7384-7399.

Ben Belgacem, F. (2007): Why is the Cauchy problem severely ill-posed? *Inverse Problems*, vol. 23, pp. 823-836.

Ben Belgacem, F.; El Fekih, H. (2005): On Cauchy's problem: I. A variational Steklov-Poincaré theory. *Inverse Problems*, vol. 21, pp. 1915-1936.

Cheng, A. H. D. (2012): Multiquadric and its shape parameter – A numerical investigation of error estimate, condition number, and round-off error by arbitrary precision computation. *Eng. Anal. Bound. Elem.*, vol. 36, pp. 220-239.

Cheng, A. H. D.; Golberg, M. A.; Kansa, E. J.; Zammito, G. (2003): Exponential convergence and H-c multiquadric collocation method for partial differential equations. *Numer. Meth. Par. Diff. Eqs.*, vol. 19, pp. 571-594.

Daubechies, I.; Defrise, M.; De Mol, C. (2004): An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commun. Pure Appl. Math.*, vol. 57, pp. 1413-1457.

Dongarra, J.; Sullivan, F. (2000): Guest editors' introduction to the top 10 algorithms. *Comput. Sci. Eng.*, vol. 2, pp. 22-23.

Fasshauer, G. E. (2002): Newton iteration with multiquadrics for the solution of nonlinear PDEs. *Comput. Math. Appl.*, vol. 43, pp. 423-438.

Fletcher, R. (1976): Conjugate gradient methods for indefinite systems. Lecture Notes in Math., vol. 506, pp. 73-89, Springer-Verlag, Berlin.

Franke, R. (1982): Scattered data interpolation: tests of some method. *Math. Comput.*, vol. 38, pp. 181-200.

Freund, R. W.; Nachtigal, N. M. (1991): QMR: a quasi-minimal residual method for non-Hermitian linear systems. *Numer. Math.*, vol. 60, pp. 315-339.

Golberg, M. A.; Chen, C. S.; Karur, S. R. (1996): Improved multiquadric approximation for partial differential equations. *Eng. Anal. Bound. Elem.*, vol. 18, pp. 9-17.

Hansen, P. C. (1992): Analysis of discrete ill-posed problems by means of the

L-curve. *SIAM Rev.*, vol. 34, pp. 561-580.

Hansen, P. C.; O'Leary, D. P. (1993): The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM J. Sci. Comput.*, vol. 14, pp. 1487-1503.

Hestenes, M. R.; Stiefel, E. L. (1952): Methods of conjugate gradients for solving linear systems. *J. Res. Nat. Bur. Stand.*, vol. 49, pp. 409-436.

Hon, Y. C.; Li, M. (2009): A discrepancy principle for the source points location in using the MFS for solving the BHCP. *Int. J. Comput. Meth.*, vol. 6, pp. 181-197.

Huang, C. S.; Lee, C. F.; Cheng, A. H. D. (2007): Error estimate, optimal shape factor, and high precision computation of multiquadric collocation method. *Eng. Anal. Bound. Elem.*, vol. 31, pp. 614-623.

Huang, C. S.; Yen, H. D.; Cheng, A. H. D. (2010): On the increasingly flat radial basis function and optimal shape parameter for the solution of elliptic PDEs. *Eng. Anal. Bound. Elem.*, vol. 34, pp. 802-809.

Lanczos, C. (1952): Solution of systems of linear equations by minimized iterations. *J. Res. Nat. Bur. Stand.*, vol. 49, pp. 33-53.

Liu, C.-S. (2011): The method of fundamental solutions for solving the backward heat conduction problem with conditioning by a new post-conditioner. *Numer. Heat Transf. B: Fund.*, vol. 60, pp. 57-72.

Liu, C.-S. (2012a): The concept of best vector used to solve ill-posed linear inverse problems. *CMES: Computer Modeling in Engineering & Sciences*, vol. 83, pp. 499-525.

Liu, C.-S. (2012b): A globally optimal iterative algorithm to solve an ill-posed linear system, *CMES: Computer Modeling in Engineering & Sciences*, vol. 84, pp. 383-403.

Liu, C.-S. (2013a): An optimal multi-vector iterative algorithm in a Krylov subspace for solving the ill-posed linear inverse problems. *CMC: Computers, Materials & Continua*, vol. 33, pp. 175-198.

Liu, C.-S. (2013b): An optimal tri-vector iterative algorithm for solving ill-posed linear inverse problems. *Inv. Prob. Sci. Eng.*, vol. 21, pp. 650-681.

Liu, C.-S. (2013c): A dynamical Tikhonov regularization for solving ill-posed linear algebraic systems. *Acta Appl. Math.*, vol. 123, pp. 285-307.

Liu, C.-S. (2013d): Discussing a more fundamental concept than the minimal residual method for solving linear system in a Krylov subspace. *J. Math. Research*, vol. 5, pp. 58-70.

Liu, C.-S. (2014a): A doubly optimized solution of linear equations system expressed in an affine Krylov subspace. *J. Comput. Appl. Math.*, vol. 260, pp.

375-394.

Liu, C.-S. (2014b): A maximal projection solution of ill-posed linear system in a column subspace, better than the least squares solution. *Comput. Math. Appl.*, vol. 67, pp. 1998-2014.

Liu, C.-S. (2014c): Optimal algorithms in a Krylov subspace for solving linear inverse problems by MFS. *Eng. Anal. Bound. Elem.*, vol. 44, pp. 64-75.

Liu, C.-S. (2015): A double optimal descent algorithm for iteratively solving ill-posed linear inverse problems. *Inv. Prob. Sci. Eng.*, vol. 23, pp. 38-66.

Matinfar, M.; Zareamoghaddam, H.; Eslami, M.; Saeidy, M. (2012): GMRES implementations and residual smoothing techniques for solving ill-posed linear systems. *Comput. Math. Appl.*, vol. 63, pp. 1-13.

Paige, C. C.; Saunders, M. A. (1975): Solution of sparse indefinite systems of linear equations. *SIAM J. Numer. Ana.*, vol. 12, pp. 617-629.

Saad, Y. (1981): Krylov subspace methods for solving large unsymmetric linear systems. *Math. Comput.*, vol. 37, pp. 105-126.

Saad, Y. (2003): Iterative Methods for Sparse Linear Systems. 2nd Ed., SIAM, Pennsylvania.

Saad, Y.; Schultz, M. H. (1986): GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Stat. Comput.*, vol. 7, pp. 856-869.

Saad, Y.; van der Vorst, H. A. (2000): Iterative solution of linear systems in the 20th century. *J. Comput. Appl. Math.*, vol. 123, pp. 1-33.

Simoncini, V.; Szyld, D. B. (2007): Recent computational developments in Krylov subspace methods for linear systems. *Numer. Linear Algebra Appl.*, vol. 14, pp. 1-59.

Sonneveld, P. (1989): CGS: a fast Lanczos-type solver for nonsymmetric linear systems. *SIAM J. Sci. Stat. Comput.*, vol. 10, pp. 36-52.

Tikhonov, A. N.; Arsenin, V. Y. (1977): Solutions of Ill-Posed Problems. John-Wiley & Sons, New York.

van Den Eshof, J.; Sleijpen, G. L. G. (2004): Inexact Krylov subspace methods for linear systems. *SIAM J. Matrix Anal. Appl.*, vol. 26, pp. 125-153.

van der Vorst, H. A. (1992): Bi-CGSTAB: a fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems. *SIAM J. Sci. Stat. Comput.*, vol. 13, pp. 631-644.

van der Vorst, H. A. (2003): Iterative Krylov Methods for Large Linear Systems. Cambridge University Press, New York.