

Discussing a More Fundamental Concept Than the Minimal Residual Method to Solve Linear System in a Krylov Subspace

Chein-Shan Liu¹

¹ Department of Civil Engineering, National Taiwan University, Taipei, Taiwan

Correspondence: Chein-Shan Liu, Department of Civil Engineering, National Taiwan University, Taipei, Taiwan.
E-mail: liucs@ntu.edu.tw

Received: October 23, 2013 Accepted: November 10, 2013 Online Published: November 14, 2013

doi:10.5539/jmr.v5n4p58 URL: <http://dx.doi.org/10.5539/jmr.v5n4p58>

Abstract

A more fundamental concept than the minimal residual method is proposed in this paper to solve an n -dimensional linear equations system $\mathbf{Ax} = \mathbf{b}$ in an m -dimensional Krylov subspace. We maximize the orthogonal projection of \mathbf{b} onto $\mathbf{y} = \mathbf{Ax}$. Then, we can prove that the maximal projection solution (MP) is better than that obtained by the least squares solution (LS) with $\|\mathbf{b} - \mathbf{Ax}_{\text{MP}}\| < \|\mathbf{b} - \mathbf{Ax}_{\text{LS}}\|$. Examples are discussed which confirm the above finding.

Keywords: linear equations system, Krylov subspace, maximal projection solution, least squares solution

1. Introduction

In this paper we derive a better Krylov subspace solution method by maximizing the orthogonal projection, instead of that obtained by the method of minimal residual, to solve the following linear equations system:

$$\mathbf{Ax} = \mathbf{b}, \quad (1)$$

where $\mathbf{x} \in \mathbb{R}^n$ is an unknown vector, to be determined from a given non-singular coefficient matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, i.e. $\text{Rank}(\mathbf{A}) = n$, and the input $\mathbf{b} \in \mathbb{R}^n$.

Given an initial guess \mathbf{x}_0 , from Equation (1) we have an initial residual

$$\mathbf{r}_0 = \mathbf{b} - \mathbf{Ax}_0.$$

Upon letting

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

Equation (1) is equivalent to

$$\mathbf{Az} = \mathbf{r}_0, \quad (2)$$

which can be used to search a descent direction \mathbf{z} after giving an initial residual \mathbf{r}_0 . Liu (2013a, 2013b, 2014a) has proposed new methods by minimizing the following merit function:

$$\min \left\{ a_0 = \frac{\|\mathbf{r}_0\|^2 \|\mathbf{Az}\|^2}{[\mathbf{r}_0 \cdot (\mathbf{Az})]^2} \right\}, \quad (3)$$

to obtain a faster descent direction \mathbf{z} in the iterative solution of Equation (1).

In the numerical solution of linear equations system the Krylov subspace method is one of the most important classes of numerical methods (Dongarra, 2000; Saad, 1981; Freund & Nachtigal, 1991; van den Eshof & Sleijpen, 2004; Liu, 2013c). The iterative algorithms that are applied to solve large scale linear systems are mostly the preconditioned Krylov subspace methods (Simoncini & Szyld, 2007).

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 Equation (2):

$$\mathcal{K}_m := \text{Span}\{\mathbf{r}_0, \mathbf{Ar}_0, \dots, \mathbf{A}^{m-1}\mathbf{r}_0\}. \quad (4)$$

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

$$\min\{\|\mathbf{r}_0 - \mathbf{Az}\|^2 = \|\mathbf{b} - \mathbf{Ax}\|^2\}. \quad (5)$$

Throughout this paper the 2-norm of a vector \mathbf{x} is denoted by $\|\mathbf{x}\|$.

Recently, Liu (2014b) has developed a new theory to find the double optimal solution of Equation (1), simultaneously based on the two minimizations in Equations (3) and (5). Here, we only use a similar merit function as that in Equation (3) and employ a scaling invariant property of the proposed merit function to derive a *maximal projection solution* (MP) in the Krylov subspace. More importantly, we can prove that the MP is better than the least squares solution (LS) with an exact estimation equation of the difference between MP and LS provided.

The remaining parts of this paper are arranged as follows. In Section 2 we start from an m -dimensional Krylov subspace to express the solution with coefficients to be optimized in Section 3, where a new merit function is proposed for finding the optimal expansion coefficients. We can derive a closed-form MP of Equation (1). The comparisons between the MP and the LS are performed in Section 4, and an important improvement is verified. The examples of linear problems and discussions are given in Section 5 to display some advantages of the present methodology to find an approximate solution of Equation (1).

2. A Krylov Subspace Method

For the linear equations system (1), by using the Cayley-Hamilton theorem we can expand \mathbf{A}^{-1} by

$$\mathbf{A}^{-1} = \frac{a_1}{a_0} \mathbf{I}_n + \frac{a_2}{a_0} \mathbf{A} + \frac{a_3}{a_0} \mathbf{A}^2 + \dots + \frac{a_{n-1}}{a_0} \mathbf{A}^{n-2} + \frac{1}{a_0} \mathbf{A}^{n-1},$$

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

$$\mathbf{x} = \mathbf{A}^{-1} \mathbf{b} = \left[\frac{a_1}{a_0} \mathbf{I}_n + \frac{a_2}{a_0} \mathbf{A} + \frac{a_3}{a_0} \mathbf{A}^2 + \dots + \frac{1}{a_0} \mathbf{A}^{n-1} \right] \mathbf{b}, \quad (6)$$

where the coefficients a_0, a_1, \dots, a_{n-1} appear in the characteristic equation for \mathbf{A} : $\lambda^n + a_{n-1}\lambda^{n-1} + \dots + a_2\lambda^2 + a_1\lambda - a_0 = 0$. Here, we assume that $a_0 = -\det(\mathbf{A}) \neq 0$. In practice, the above formula to find the solution of \mathbf{x} is quite difficult to be realized, since the coefficients a_j , $j = 0, 1, \dots, n-1$ are hard to find, and the computations of the higher order powers of \mathbf{A} are very expensive, when n is a quite large positive integer.

The idea of projection method, including the GMRES, is searching a solution \mathbf{x} in a smaller subspace than the original space \mathbb{R}^n with dimension $m \ll n$. In doing so, the higher order expansion terms in Equation (6) are truncated, and we can find the lower order expansion coefficients through a suitably designed optimization in a Krylov subspace. A basic ingredient of the Krylov subspace method is the construction of an orthonormal set of linearly independent bases. We describe how to set up the bases \mathbf{u}_k , $k = 1, \dots, m$ by the Krylov subspace method. Suppose that we have an m -dimensional Krylov subspace generated by the coefficient matrix \mathbf{A} from the right-hand side vector \mathbf{b} in Equation (1):

$$\mathcal{K}_m := \text{Span}\{\mathbf{b}, \mathbf{Ab}, \dots, \mathbf{A}^{m-1}\mathbf{b}\}. \quad (7)$$

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

$$\mathbf{U} := [\mathbf{u}_1, \dots, \mathbf{u}_m], \quad (8)$$

which is an $n \times m$ Arnoldi matrix with its j th column being the vector \mathbf{u}_j . Because $\mathbf{u}_1, \dots, \mathbf{u}_m$ are linearly independent and $m < n$, \mathbf{U} has a full column rank, that is, $\text{Rank}(\mathbf{U}) = m$. The expansion of \mathbf{x} in the Krylov subspace is denoted by $\mathbf{x} \in \mathcal{K}_m$. Then, we can prove the following result, where we minimize $\|\mathbf{r}\|$ as shown in Figure 1(a).

Theorem 1 For $\mathbf{x} \in \mathcal{K}_m$, and $\mathbf{b} \neq \mathbf{0} \in \mathbb{R}^n$ being a given vector, the best \mathbf{x} and $\mathbf{y} = \mathbf{Ax}$ which satisfy

$$\min_{\mathbf{y}} \{\|\mathbf{r}\|^2 = \|\mathbf{b} - \mathbf{Ax}\|^2 = \|\mathbf{b} - \mathbf{y}\|^2\} \quad (9)$$

are given by

$$\mathbf{x} = \mathbf{Xb}, \quad (10)$$

$$\mathbf{y} = \mathbf{Eb}, \quad (11)$$

where $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{J}^T$, and $\mathbf{E} = \mathbf{A}\mathbf{X}$.

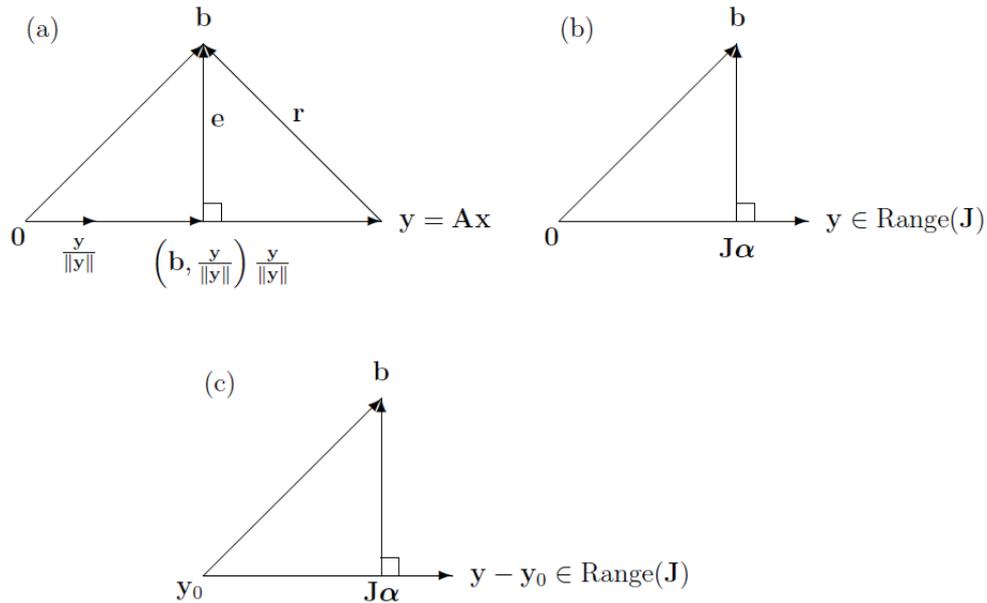


Figure 1. (a) The approximation of \mathbf{b} by \mathbf{y} , and (b) the relation of \mathbf{b} , \mathbf{y} and $\mathbf{J}\alpha$ used in Lemma 1, and (c) the relation of $\mathbf{b} - \mathbf{y}_0$, $\mathbf{y} - \mathbf{y}_0$ and $\mathbf{J}\alpha$ used in Lemma 2

Proof. Because of $\mathbf{x} \in \mathcal{K}_m$ we can write

$$\mathbf{x} = \mathbf{U}\alpha, \tag{12}$$

where $\alpha := (\alpha_1, \dots, \alpha_m)^T$ consists of the expansion coefficients, and the superscript τ denotes the transpose.

Let \mathbf{J} be an $n \times m$ matrix:

$$\mathbf{J} := [\mathbf{A}\mathbf{u}_1, \dots, \mathbf{A}\mathbf{u}_m] = \mathbf{A}\mathbf{U}. \tag{13}$$

By the assumption of the full ranks of \mathbf{A} and \mathbf{U} , \mathbf{J} has a full rank with $\text{Rank}(\mathbf{J}) = m$. Then, $\mathbf{y} = \mathbf{A}\mathbf{x}$ can be written as

$$\mathbf{y} = \mathbf{J}\alpha. \tag{14}$$

Expanding the square residual we have

$$\|\mathbf{b} - \mathbf{y}\|^2 = \|\mathbf{b}\|^2 - 2\mathbf{b} \cdot \mathbf{y} + \|\mathbf{y}\|^2, \tag{15}$$

where

$$\mathbf{b} \cdot \mathbf{y} = \mathbf{b}^T \mathbf{J}\alpha, \tag{16}$$

$$\|\mathbf{y}\|^2 = \alpha^T \mathbf{C}\alpha, \tag{17}$$

$$\mathbf{C} := \mathbf{J}^T \mathbf{J}. \tag{18}$$

A dot between two vectors signifies the inner product of these two vectors. Because \mathbf{J} has a full rank, \mathbf{C} is an $m \times m$ positive definite matrix, whose inversion is denoted by $\mathbf{D} = \mathbf{C}^{-1}$.

Inserting Equations (16) and (17) into Equation (15), taking the differential with respect to α and setting it to be zero, we can find

$$\alpha = \mathbf{D}\mathbf{J}^T \mathbf{b}. \tag{19}$$

Feeding it into Equation (12) we can obtain Equation (10), while Equation (11) is obtained from Equation (10) by multiplying \mathbf{A} on both sides. \square

3. Maximizing the Orthogonal Projection

In this section we will propose a new merit function to improve the solution in Theorem 1.

3.1 An orthogonal Projection of \mathbf{b} onto \mathbf{y}

Let

$$\mathbf{y} := \mathbf{A}\mathbf{x}, \quad (20)$$

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

We consider finding the best approximation of \mathbf{y} to \mathbf{b} . The orthogonal projection of \mathbf{b} to \mathbf{y} is regarded as the approximation of \mathbf{b} by \mathbf{y} as shown in Figure 1(a), whose error vector is written as

$$\mathbf{e} := \mathbf{b} - \left(\mathbf{b}, \frac{\mathbf{y}}{\|\mathbf{y}\|} \right) \frac{\mathbf{y}}{\|\mathbf{y}\|}, \quad (21)$$

where the parenthesis denotes the inner product. The best approximation can be found with \mathbf{y} minimizing

$$\|\mathbf{e}\|^2 = \|\mathbf{b}\|^2 - \frac{(\mathbf{b} \cdot \mathbf{y})^2}{\|\mathbf{y}\|^2}, \quad (22)$$

or maximizing the square norm of the orthogonal projection of \mathbf{b} to \mathbf{y} , i.e.,

$$\max_{\mathbf{y}} \left\{ \frac{(\mathbf{b} \cdot \mathbf{y})^2}{\|\mathbf{y}\|^2} \right\}. \quad (23)$$

Due to this reason the solution of the above equation will be named the *maximal projection* solution (MP), to distinct it from the well-known *least squares* solution (LS).

3.2 A Main Result

The maximum in Equation (23) is equivalent to minimize the following merit function:

$$\min_{\mathbf{y}} \left\{ f := \frac{\|\mathbf{y}\|^2}{(\mathbf{b} \cdot \mathbf{y})^2} \right\}. \quad (24)$$

However, it is a quite difficult optimization problem, and how to solve it is given below.

Theorem 2 For $\mathbf{x} \in \mathcal{K}_m$, and $\mathbf{b} \neq \mathbf{0} \in \mathbb{R}^n$ being a given vector, the best \mathbf{x} and $\mathbf{y} = \mathbf{A}\mathbf{x}$ which satisfy

$$\min_{\mathbf{y}} \left\{ f = \frac{\|\mathbf{y}\|^2}{(\mathbf{b} \cdot \mathbf{y})^2} \right\} \quad (25)$$

are given by

$$\mathbf{x} = \mathbf{X}\mathbf{b} + \alpha_0\mathbf{b} - \alpha_0\mathbf{X}\mathbf{A}\mathbf{b}, \quad (26)$$

$$\mathbf{y} = \mathbf{E}\mathbf{b} + \alpha_0\mathbf{A}\mathbf{b} - \alpha_0\mathbf{E}\mathbf{A}\mathbf{b}, \quad (27)$$

where

$$\begin{aligned} \mathbf{X} &= \mathbf{U}\mathbf{D}\mathbf{J}^T, \\ \mathbf{E} &= \mathbf{A}\mathbf{X} = \mathbf{J}\mathbf{D}\mathbf{J}^T, \\ \alpha_0 &= \frac{\mathbf{b}^T\mathbf{A}\mathbf{b} - \mathbf{b}^T\mathbf{E}\mathbf{A}\mathbf{b}}{\mathbf{b}^T\mathbf{A}^T\mathbf{A}\mathbf{b} - \mathbf{b}^T\mathbf{A}^T\mathbf{E}\mathbf{A}\mathbf{b}}. \end{aligned} \quad (28)$$

Moreover, we have the following implication and identity:

$$\min_{\mathbf{y}} \left\{ \frac{\|\mathbf{y}\|^2}{(\mathbf{b} \cdot \mathbf{y})^2} \right\} \Rightarrow \min_{\mathbf{y}} \{\|\mathbf{r}\|^2 = \|\mathbf{b} - \mathbf{y}\|^2 = \|\mathbf{b} - \mathbf{A}\mathbf{x}\|^2\}, \quad (29)$$

$$\|\mathbf{e}\|^2 = \|\mathbf{r}\|^2. \quad (30)$$

Proof. The proof of this theorem is quite lengthy and we divide it into four parts. (A) Because of $\mathbf{x} \in \mathcal{K}_m$ we can expand \mathbf{x} by

$$\mathbf{x} = \alpha_0 \mathbf{b} + \mathbf{U}\alpha, \quad (31)$$

where α_0 is a scaling factor to be determined below, and $\alpha := (\alpha_1, \dots, \alpha_m)^T \in \mathbb{R}^m$ is the collection of m expansion coefficients. Here, we intentionally divide the coefficient preceded $\mathbf{u}_1 = \mathbf{b}/\|\mathbf{b}\|$ into two parts $\alpha_0\|\mathbf{b}\|$ and α_1 . Due to Equation (31), $\mathbf{y} = \mathbf{A}\mathbf{x}$ reads as

$$\mathbf{y} = \mathbf{y}_0 + \mathbf{J}\alpha, \quad (32)$$

where \mathbf{J} was defined by Equation (13) and

$$\mathbf{y}_0 = \alpha_0 \mathbf{A}\mathbf{b}. \quad (33)$$

With the help of Equation (32), the terms $\mathbf{b} \cdot \mathbf{y}$ and $\|\mathbf{y}\|^2$ in Equation (25) can be written as

$$\mathbf{b} \cdot \mathbf{y} = \mathbf{b} \cdot \mathbf{y}_0 + \mathbf{b}^T \mathbf{J}\alpha, \quad (34)$$

$$\|\mathbf{y}\|^2 = \|\mathbf{y}_0\|^2 + 2\mathbf{y}_0^T \mathbf{J}\alpha + \alpha^T \mathbf{J}^T \mathbf{J}\alpha. \quad (35)$$

From the necessary condition for the minimization of f we have

$$\nabla_{\alpha} \frac{\|\mathbf{y}\|^2}{(\mathbf{b} \cdot \mathbf{y})^2} = \mathbf{0} \Rightarrow (\mathbf{b} \cdot \mathbf{y})^2 \nabla_{\alpha} \|\mathbf{y}\|^2 - 2\mathbf{b} \cdot \mathbf{y} \|\mathbf{y}\|^2 \nabla_{\alpha} (\mathbf{b} \cdot \mathbf{y}) = \mathbf{0}, \quad (36)$$

in which ∇_{α} denotes the gradient with respect to α . Thus, we can derive

$$\mathbf{b} \cdot \mathbf{y}\mathbf{y}_2 - 2\|\mathbf{y}\|^2 \mathbf{y}_1 = \mathbf{0}, \quad (37)$$

where

$$\mathbf{y}_1 := \nabla_{\alpha} (\mathbf{b} \cdot \mathbf{y}) = \mathbf{J}^T \mathbf{b}, \quad (38)$$

$$\mathbf{y}_2 := \nabla_{\alpha} \|\mathbf{y}\|^2 = 2\mathbf{J}^T \mathbf{y}_0 + 2\mathbf{J}^T \mathbf{J}\alpha. \quad (39)$$

(B) With the help of Equation (18), Equations (35) and (39) can be written as

$$\|\mathbf{y}\|^2 = \|\mathbf{y}_0\|^2 + 2\mathbf{y}_0^T \mathbf{J}\alpha + \alpha^T \mathbf{C}\alpha, \quad (40)$$

$$\mathbf{y}_2 = 2\mathbf{J}^T \mathbf{y}_0 + 2\mathbf{C}\alpha. \quad (41)$$

From Equation (37) we can observe that \mathbf{y}_2 is proportional to \mathbf{y}_1 , which is supposed to be

$$\mathbf{y}_2 = \frac{2\|\mathbf{y}\|^2}{\mathbf{b} \cdot \mathbf{y}} \mathbf{y}_1 = 2\lambda \mathbf{y}_1, \quad (42)$$

where 2λ is a multiplier to be determined, abiding to the *principle of simplicity*. From the second equality, by cancelling the common term $2\mathbf{y}_1$ on both sides, we have

$$\|\mathbf{y}\|^2 = \lambda \mathbf{b} \cdot \mathbf{y}. \quad (43)$$

Then, by Equations (38), (41) and (42) we have

$$\alpha = \lambda \mathbf{D}\mathbf{J}^T \mathbf{b} - \mathbf{D}\mathbf{J}^T \mathbf{y}_0, \quad (44)$$

where

$$\mathbf{D} := \mathbf{C}^{-1} = (\mathbf{J}^T \mathbf{J})^{-1}. \quad (45)$$

Inserting Equation (44) into Equations (34) and (40) we have

$$\mathbf{b} \cdot \mathbf{y} = \mathbf{b} \cdot \mathbf{y}_0 + \lambda \mathbf{b}^T \mathbf{E}\mathbf{b} - \mathbf{b}^T \mathbf{E}\mathbf{y}_0, \quad (46)$$

$$\|\mathbf{y}\|^2 = \lambda^2 \mathbf{b}^T \mathbf{E}\mathbf{b} + \|\mathbf{y}_0\|^2 - \mathbf{y}_0^T \mathbf{E}\mathbf{y}_0, \quad (47)$$

where

$$\mathbf{E} := \mathbf{J}\mathbf{D}\mathbf{J}^T \quad (48)$$

is an $n \times n$ positive semi-definite matrix.

Now, from Equations (43), (46) and (47) we can derive a linear equation:

$$\|\mathbf{y}_0\|^2 - \mathbf{y}_0^T \mathbf{E} \mathbf{y}_0 = \lambda [\mathbf{b} \cdot \mathbf{y}_0 - \mathbf{b}^T \mathbf{E} \mathbf{y}_0], \quad (49)$$

such that λ is derived as follows:

$$\lambda = \frac{\|\mathbf{y}_0\|^2 - \mathbf{y}_0^T \mathbf{E} \mathbf{y}_0}{\mathbf{b} \cdot \mathbf{y}_0 - \mathbf{b}^T \mathbf{E} \mathbf{y}_0}. \quad (50)$$

Inserting Equation (50) into Equation (44) and using Equation (33), α is given by

$$\alpha = \alpha_0 \left[\frac{\mathbf{b}^T \mathbf{A}^T \mathbf{A} \mathbf{b} - \mathbf{b}^T \mathbf{A}^T \mathbf{E} \mathbf{A} \mathbf{b}}{\mathbf{b}^T \mathbf{A} \mathbf{b} - \mathbf{b}^T \mathbf{E} \mathbf{A} \mathbf{b}} \mathbf{D} \mathbf{J}^T \mathbf{b} - \mathbf{D} \mathbf{J}^T \mathbf{A} \mathbf{b} \right]. \quad (51)$$

Then, using Equations (32), (33) and (48), we can derive

$$\mathbf{y} = \alpha_0 \mathbf{A} \mathbf{b} + \alpha_0 \left[\frac{\mathbf{b}^T \mathbf{A}^T \mathbf{A} \mathbf{b} - \mathbf{b}^T \mathbf{A}^T \mathbf{E} \mathbf{A} \mathbf{b}}{\mathbf{b}^T \mathbf{A} \mathbf{b} - \mathbf{b}^T \mathbf{E} \mathbf{A} \mathbf{b}} \mathbf{E} \mathbf{b} - \mathbf{E} \mathbf{A} \mathbf{b} \right]. \quad (52)$$

(C) From Equation (25) it can be seen that if \mathbf{y} is a solution, $c\mathbf{y}$, $c \neq 0$ is also a solution. It means that the solution of Equation (25) is scaling invariant. So we can select a suitable scaling factor α_0 to be

$$\alpha_0 = \frac{\mathbf{b}^T \mathbf{A} \mathbf{b} - \mathbf{b}^T \mathbf{E} \mathbf{A} \mathbf{b}}{\mathbf{b}^T \mathbf{A}^T \mathbf{A} \mathbf{b} - \mathbf{b}^T \mathbf{A}^T \mathbf{E} \mathbf{A} \mathbf{b}}, \quad (53)$$

such that \mathbf{y} is simplified to

$$\mathbf{y} = \mathbf{E} \mathbf{b} + \alpha_0 \mathbf{A} \mathbf{b} - \alpha_0 \mathbf{E} \mathbf{A} \mathbf{b}. \quad (54)$$

Defining

$$\mathbf{X} = \mathbf{U} \mathbf{D} \mathbf{J}^T, \quad (55)$$

inserting Equation (51) into Equation (31) and using Equation (53), we can derive Equation (26).

Inserting Equation (33) into Equation (50) we can obtain

$$\lambda = \alpha_0 \frac{\mathbf{b}^T \mathbf{A}^T \mathbf{A} \mathbf{b} - \mathbf{b}^T \mathbf{A}^T \mathbf{E} \mathbf{A} \mathbf{b}}{\mathbf{b}^T \mathbf{A} \mathbf{b} - \mathbf{b}^T \mathbf{E} \mathbf{A} \mathbf{b}}, \quad (56)$$

which, by using Equation (53), leads to

$$\lambda = 1; \quad (57)$$

hence, Equation (43) is simplified to

$$\|\mathbf{y}\|^2 = \mathbf{b} \cdot \mathbf{y}. \quad (58)$$

Under the *principle of simplicity*, we have chosen the scaling factor α_0 to be given by Equation (53), which renders a simple value of $\lambda = 1$, and more importantly a simple relation between $\|\mathbf{y}\|^2$ and $\mathbf{b} \cdot \mathbf{y}$ in Equation (58). The proof of some important properties given below requires this equation.

(D) The minimized function in Equation (9) can be written as

$$\|\mathbf{r}\|^2 = \|\mathbf{b}\|^2 - 2\mathbf{b} \cdot \mathbf{y} + \|\mathbf{y}\|^2. \quad (59)$$

Then by using Equation (58) we have

$$\|\mathbf{r}\|^2 = \|\mathbf{b}\|^2 - \mathbf{b} \cdot \mathbf{y}. \quad (60)$$

When $\mathbf{b} \cdot \mathbf{y}$ is maximized as shown by Equation (23), where we can take $\|\mathbf{y}\| = 1$ by using the scale invariance of \mathbf{y} , it implies that the residual $\|\mathbf{r}\|^2 = \|\mathbf{b} - \mathbf{A} \mathbf{x}\|^2$ is minimized by viewing the above equation. Hence, Equation (29) is proven. Inserting Equation (58) into Equations (22) and (60) we have

$$\|\mathbf{e}\|^2 = \|\mathbf{b}\|^2 - \|\mathbf{y}\|^2 = \|\mathbf{r}\|^2, \quad (61)$$

as shown in Equation (30). This ends the proof of Theorem 2. \square

Remark 1 In Equation (10), \mathbf{Xb} is known a least squares solution of Equation (1) in the Krylov subspace $\mathbf{x} \in \mathcal{K}_m$, which minimizes the residual $\|\mathbf{b} - \mathbf{Ax}\|$. Obviously, Equation (10) is a special case of Equation (26) with $\alpha_0 = 0$. Theorem 2 indicates that the minimization in Equation (25) is a more fundamental concept than the usual minimization of the residual $\|\mathbf{b} - \mathbf{Ax}\|^2$ as shown in Equation (29).

4. Comparing the Maximal Projection and Least Squares Solutions

In this section we compare the two optimal solutions of Equation (1) derived in Theorems 1 and 2, and prove some important results.

Lemma 1 For $\text{Rank}(\mathbf{J}) = m$, α in Theorem 1 appeared in $\mathbf{y} = \mathbf{J}\alpha$ of Equation (14) is a least squares solution of the following least squares problem:

$$\min_{\alpha \in \mathbb{R}^m} \|\mathbf{b} - \mathbf{J}\alpha\|. \quad (62)$$

Proof. α in Equation (14) satisfies Equation (62) and is a least squares solution of the following overdetermined linear system:

$$\mathbf{J}\alpha = \mathbf{b}. \quad (63)$$

In the sense of Penrose, we have

$$\alpha = \mathbf{J}^\dagger \mathbf{b} = \mathbf{D}\mathbf{J}^T \mathbf{b}, \quad (64)$$

where \mathbf{J}^\dagger is the Penrose pseudo-inverse of \mathbf{J} (Trefethen & Bau III, 1997), and $\mathbf{J}\alpha$ is a projection of \mathbf{b} to the nearest point in the space of $\text{Range}(\mathbf{J})$ as shown in Figure 1(b). \square

Lemma 2 For $\text{Rank}(\mathbf{J}) = m$, α in $\mathbf{y} - \mathbf{y}_0 = \mathbf{J}\alpha$ of Equation (32) is a least squares solution of the following least squares problem:

$$\min_{\alpha \in \mathbb{R}^m} \|\mathbf{b} - \mathbf{y}_0 - \mathbf{J}\alpha\|. \quad (65)$$

Proof. By using Equations (44) and (57) we have

$$\alpha = \mathbf{D}\mathbf{J}^T (\mathbf{b} - \mathbf{y}_0). \quad (66)$$

The above α satisfies Equation (65) and is a least squares solution of the following overdetermined linear system:

$$\mathbf{J}\alpha = \mathbf{b} - \mathbf{y}_0. \quad (67)$$

Hence, we have

$$\alpha = \mathbf{J}^\dagger (\mathbf{b} - \mathbf{y}_0) = \mathbf{D}\mathbf{J}^T (\mathbf{b} - \mathbf{y}_0), \quad (68)$$

where \mathbf{J}^\dagger is the Penrose pseudo-inverse of \mathbf{J} , and $\mathbf{J}\alpha$ is a projection of $\mathbf{b} - \mathbf{y}_0$ to the nearest point in the space of $\text{Range}(\mathbf{J})$ as shown in Figure 1(c). \square

Remark 2 From Equations (32), (68) and (48) we have $\mathbf{y} - \mathbf{y}_0 = \mathbf{J}\alpha = \mathbf{E}(\mathbf{b} - \mathbf{y}_0)$, where the orthogonal projector \mathbf{E} plays a role to project $\mathbf{b} - \mathbf{y}_0$ onto the space of $\text{Range}(\mathbf{J}) \subset \text{Range}(\mathbf{A})$. On the other hand, by setting $\alpha_0 = 0$ in Equation (26) and using Equations (12) and (64) we can write

$$\mathbf{x}_{LS} = \mathbf{U}\mathbf{J}^\dagger \mathbf{b}, \quad (69)$$

which is a least squares solution in the Krylov subspace for the minimization in Equation (9). Because of $\alpha_0 = 0$, the point \mathbf{y}_0 in Figure 1(c) moves to the zero point as shown in Figure 1(b). In general, α_0 not necessarily be a small number, and thus the above solution will be less accurate than that obtained from Equation (26), which can be recast to

$$\mathbf{x}_{MP} = \mathbf{U}\mathbf{J}^\dagger \mathbf{b} + \alpha_0 (\mathbf{b} - \mathbf{X}\mathbf{A}\mathbf{b}). \quad (70)$$

Then, we can claim that the *maximal projection* solution in Equation (70) is better than the *least squares* solution obtained from the minimum of residual in Equation (9), which is recast to Equation (69).

Below we prove two main results about the residuals of the optimal solutions derived from Theorems 1 and 2; however, before that we need the following lemma.

Lemma 3 Both \mathbf{E} and $\mathbf{I}_n - \mathbf{E}$ are projection operators, which render

$$\mathbf{x}^T \mathbf{E} \mathbf{x} > 0, \quad \forall \mathbf{x} \neq \mathbf{0} \in \mathbb{R}^n / \text{Null}(\mathbf{J}^T), \quad (71)$$

$$\mathbf{x}^T (\mathbf{I}_n - \mathbf{E}) \mathbf{x} > 0, \quad \forall \mathbf{x} \neq \mathbf{0} \in \mathbb{R}^n. \quad (72)$$

Proof. From Equations (48) and (45) it follows that

$$\mathbf{E}^2 = \mathbf{E}, \quad (73)$$

which means that \mathbf{E} and $\mathbf{I}_n - \mathbf{E}$ are projection operators. By using the following identity:

$$\mathbf{x}^T \mathbf{E} \mathbf{x} = \mathbf{x}^T \mathbf{E}^2 \mathbf{x} = (\mathbf{E} \mathbf{x})^T (\mathbf{E} \mathbf{x}) > 0, \quad \forall \mathbf{x} \neq \mathbf{0} \in \mathbb{R}^n / \text{Null}(\mathbf{J}^T), \quad (74)$$

Equation (71) is proven. Similarly, we can prove Equation (72). \square

In Theorem 2, we have given a qualitative implication between the two minimizations in Equations (9) and (25); however, a quantitative description is still absent. Now we can offer the following crucial results.

Theorem 3 For $\mathbf{x} \in \mathcal{K}_m$, and $\mathbf{b} \neq \mathbf{0} \in \mathbb{R}^n$ being a given vector, the best \mathbf{x}_{MP} and $\mathbf{y}_{MP} = \mathbf{A} \mathbf{x}_{MP}$ which minimize the merit function in Equation (25) has the following residual:

$$\|\mathbf{b} - \mathbf{y}_{MP}\|^2 = \mathbf{b}^T (\mathbf{I}_n - \mathbf{E}) \mathbf{b} - \alpha_0^2 \mathbf{b}^T \mathbf{A}^T (\mathbf{I}_n - \mathbf{E}) \mathbf{A} \mathbf{b}, \quad (75)$$

$$\|\mathbf{b} - \mathbf{y}_{MP}\|^2 < \mathbf{b}^T (\mathbf{I}_n - \mathbf{E}) \mathbf{b}. \quad (76)$$

Proof. Inserting Equation (58) into Equation (59) and using the notation \mathbf{y}_{MP} for \mathbf{y} we have

$$\|\mathbf{b} - \mathbf{y}_{MP}\|^2 = \|\mathbf{b}\|^2 - \|\mathbf{y}_{MP}\|^2. \quad (77)$$

Using Equation (47) for \mathbf{y}_{MP} and taking Equations (57) and (33) into account, we have

$$\|\mathbf{y}_{MP}\|^2 = \mathbf{b}^T \mathbf{E} \mathbf{b} + \alpha_0^2 \mathbf{b}^T \mathbf{A}^T (\mathbf{I}_n - \mathbf{E}) \mathbf{A} \mathbf{b}. \quad (78)$$

Then, Equation (75) follows from the above two equations. In view of Lemma 3, the following term

$$\mathbf{b}^T \mathbf{A}^T (\mathbf{I}_n - \mathbf{E}) \mathbf{A} \mathbf{b} > 0 \quad (79)$$

is positive. Hence, the inequality in Equation (76) follows from Equation (75) by using $\alpha_0^2 > 0$. \square

As a consequence we have

Theorem 4 For $\mathbf{x} \in \mathcal{K}_m$, and $\mathbf{b} \neq \mathbf{0} \in \mathbb{R}^n$ being a given vector, the best \mathbf{x}_{LS} and $\mathbf{y}_{LS} = \mathbf{A} \mathbf{x}_{LS}$ which minimize the merit function in Equation (9) has the following residual:

$$\|\mathbf{b} - \mathbf{y}_{LS}\|^2 = \mathbf{b}^T (\mathbf{I}_n - \mathbf{E}) \mathbf{b}. \quad (80)$$

Proof. Taking $\alpha_0 = 0$ in Equation (75) ends the proof. \square

Theorem 5 For $\mathbf{x} \in \mathcal{K}_m$, and $\mathbf{b} \neq \mathbf{0} \in \mathbb{R}^n$ being a given vector, the best vectors \mathbf{y}_{MP} and \mathbf{y}_{LS} which minimize, respectively, the merit functions in Equations (25) and (9) have the following relations about the residuals:

$$\|\mathbf{b} - \mathbf{y}_{MP}\|^2 = \|\mathbf{b} - \mathbf{y}_{LS}\|^2 - \frac{[\mathbf{b}^T (\mathbf{I}_n - \mathbf{E}) \mathbf{A} \mathbf{b}]^2}{\mathbf{b}^T \mathbf{A}^T (\mathbf{I}_n - \mathbf{E}) \mathbf{A} \mathbf{b}}, \quad (81)$$

$$\|\mathbf{b} - \mathbf{y}_{MP}\|^2 < \|\mathbf{b} - \mathbf{y}_{LS}\|^2. \quad (82)$$

Proof. Inserting Equation (53) for α_0 into Equation (75) we have

$$\|\mathbf{b} - \mathbf{y}_{MP}\|^2 = \mathbf{b}^T (\mathbf{I}_n - \mathbf{E}) \mathbf{b} - \frac{[\mathbf{b}^T (\mathbf{I}_n - \mathbf{E}) \mathbf{A} \mathbf{b}]^2}{\mathbf{b}^T \mathbf{A}^T (\mathbf{I}_n - \mathbf{E}) \mathbf{A} \mathbf{b}}. \quad (83)$$

Subtracting it by Equation (80) we can prove Equation (81). The inequality in Equation (82) is obtained by using Equations (81) and (79). \square

5. Results and Discussions

In order to compare the performance of the newly developed maximal projection (MP) solution and that obtained by the least squares (LS) solution, we test two linear problems, one direct problem and one inverse problem.

5.1 Example 1

Finding an n -order polynomial function $p(x) = a_0 + a_1x + \dots + a_nx^n$ to best match a continuous function $f(x)$ in the interval of $x \in [0, 1]$:

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

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

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

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

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

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

The Hilbert matrix is a notorious example of highly ill-conditioned matrices. Equation (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. The ill-posedness of Equation (1) with the above coefficient matrix \mathbf{A} increases very fast with n . Todd (1954) has proven that the asymptotic of condition number of the Hilbert matrix is

$$O\left(\frac{(1 + \sqrt{2})^{4n+4}}{\sqrt{n}}\right).$$

We consider an exact solution with $x_j = 1$, $j = 1, \dots, n$ and b_i is given by

$$b_i = \sum_{j=1}^n \frac{1}{i + j - 1}.$$

Then, we solve this problem by using the LS and MP solutions under $n = 300$. In Figures 2(a) we show the values of α_0 of the MP solution with respect to m in the range of $6 \leq m \leq 13$. The values of α_0 are in the range of $[0.3, 2.5]$. Then the residuals and the maximum errors of x_i with respect to m are compared in Figures 2(b) and 2(c). It can be seen that the MP solutions are slightly better than that of the LS solutions. When we take $m = 12$, the solution is the best one with the maximum error being 8.96×10^{-4} and the residual being 4.18×10^{-9} .

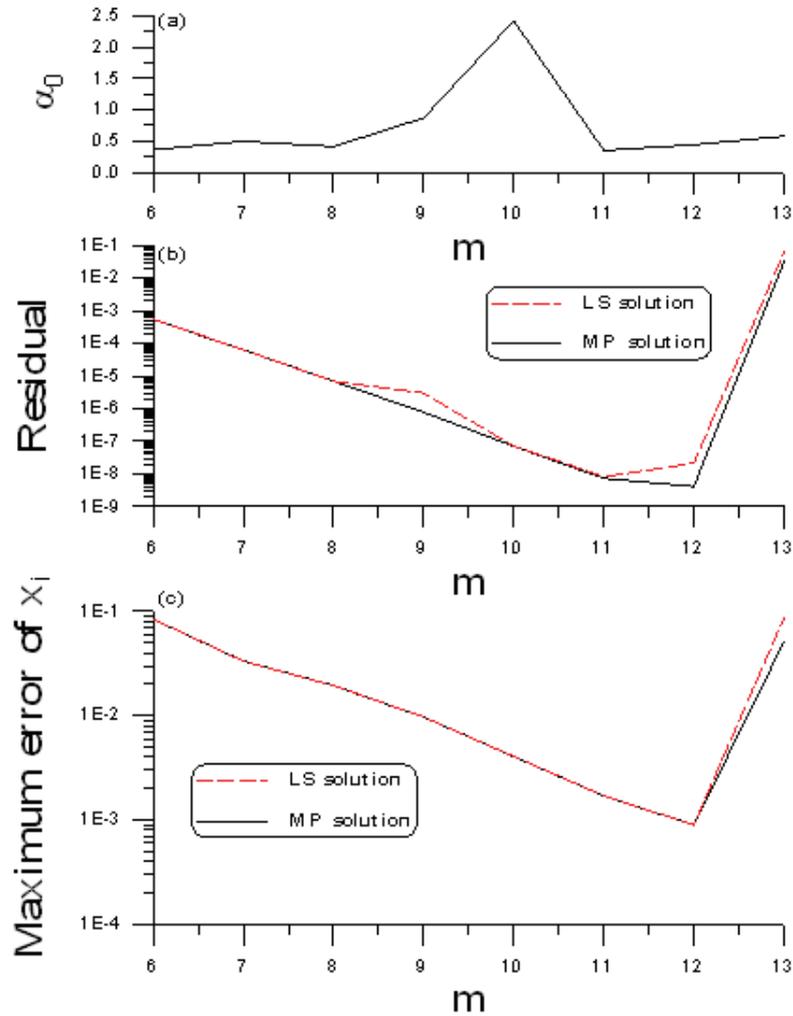


Figure 2. For the Hilbert linear problem with $n = 300$, (a) α_0 of MP solution, comparing (b) residual errors and (c) maximum errors obtained by MP and LS solutions

5.2 Example 2

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:

$$u_t(x, t) = \kappa u_{xx}(x, t), \quad 0 < t < T, \quad 0 < x < \ell,$$

$$u(0, t) = u_0(t), \quad u(\ell, t) = u_\ell(t),$$

we solve u under a final time condition:

$$u(x, T) = u^T(x).$$

The fundamental solution of Equation (84) is by

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

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

The method of fundamental solutions (MFS) has a serious drawback that the resulting linear equations system is always highly ill-conditioned, when the number of source points is increased, or when the distances of source points are increased.

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

$$u(\mathbf{z}) = \sum_{j=1}^n c_j U(\mathbf{z}, \mathbf{s}_j), \quad \mathbf{s}_j = (\eta_j, \tau_j) \in \Omega^c, \tag{84}$$

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

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

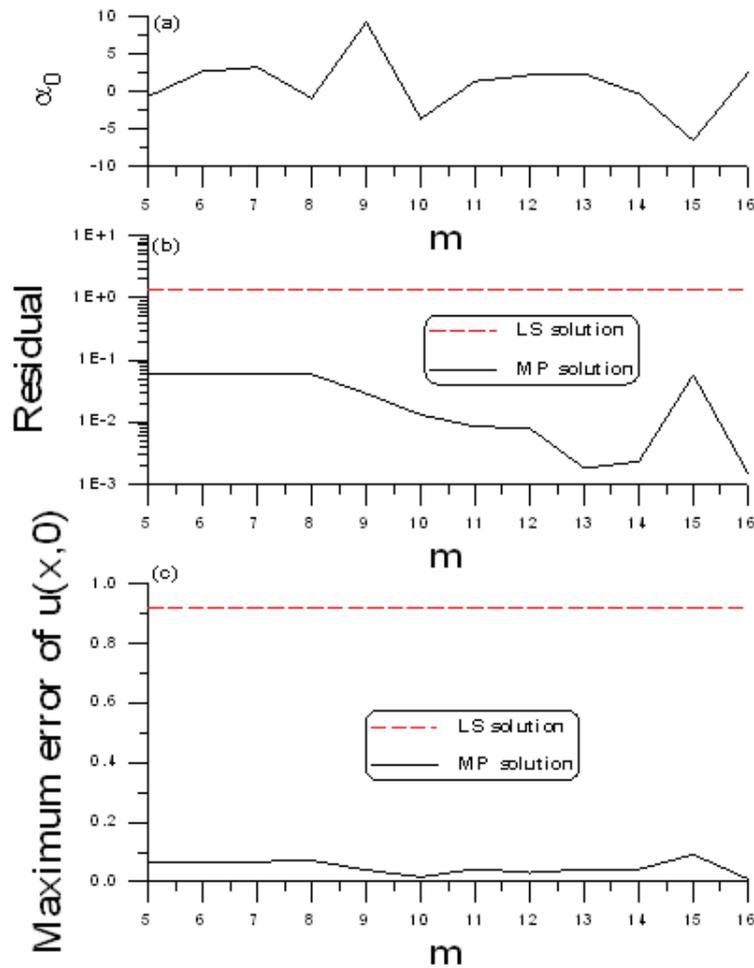


Figure 3. For backward heat conduction problem under a noise, (a) α_0 of MP solution, comparing (b) residual errors and (c) maximum errors obtained by MP and LS solutions

After imposing the boundary conditions and the final time condition to Equation (84) we can obtain a linear equations system:

$$\mathbf{Ax} = \mathbf{b}, \tag{85}$$

where

$$A_{ij} = U(\mathbf{z}_i, \mathbf{s}_j), \quad \mathbf{x} = (c_1, \dots, c_n)^T, \\ \mathbf{b} = (u_\ell(t_i), i = 1, \dots, m_1; u^T(x_j), j = 1, \dots, m_2; u_0(t_k), k = m_1, \dots, 1)^T,$$

and $n = 2m_1 + m_2$.

Since the BHCP is highly ill-posed, the ill-condition of the coefficient matrix \mathbf{A} in Equation (85) is serious. To overcome the ill-posedness of Equation (85) we can use the MP to solve this problem. Here we compare the

optimal 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 $O(1)$. We impose a relative random noise with an intensity $\sigma = 10\%$ being imposed on the final time data, which is used to test the stability of MP solution. Under the following parameters $m_1 = 11$ and $m_2 = 6$, and hence $n = 28$, we first plot the values of α_0 of the MP solution with respect to m in the range of $5 \leq m \leq 16$. The values of α_0 are in the range of $[-6.5, 9.4]$. The residual error $\|\mathbf{b} - \mathbf{Ax}\|$ with respect to m is plotted in Figure 3(b), while the maximum error of $u(x, 0)$ is plotted in Figure 3(c), of which $m = 16$ is the best one. With $m = 16$ we can obtain very accurate solution with the maximum error being 9.37×10^{-3} . The solutions obtained by the LS method are very bad, which show that the LS solution is not applicable to the inverse problem. Because α_0 is quite large, neglecting which in Equation (10) causes a large error as shown in Figures 3(b) and 3(c) by dashed lines.

5.3 Discussions

Because the least squares methods are popularly used in the mathematical, physical and engineering science (Blais, 2010), the results presented in this paper are quite significant and promising that a more fundamental and better solution than the least squares solution exists. It can be seen that Theorem 5 guarantees that the MP solution is better than the LS solution as shown in example 1 for a direct problem. For the inverse problem of example 2 the superiority of MP solution is fully exposed, of which the LP solution is thoroughly failure, but the MP solution is still workable, giving solutions with higher accuracy and higher robustness against a large noise up to 10%. More studies are required in order to test the performance of the newly developed *maximal projection* solution in the Krylov subspace for other systems. Because the theorems were proven without needing of the restriction of m , the *maximal projection* solution is always better than the LS solution, independent to the Krylov subspace and its dimension m . The new methodology presented here may shed a new light on numerical methods which are based on the least squares method.

Acknowledgements

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, and the 2011 Taiwan Research Front Award from Thomson Reuters. It is also acknowledged that the author has been promoted as being a Lifetime Distinguished Professor of National Taiwan University since 2013.

References

- Blais, J. A. R. (2010). Least squares for practitioners. *Math. Probl. Eng.*, 2010. <http://dx.doi.org/10.1155/2010/508092>
- Dongarra, J., & Sullivan, F. (2000). Guest editors' introduction to the top 10 algorithms. *Comput. Sci. Eng.*, 2, 22-23. <http://dx.doi.org/10.1109/MCISE.2000.814652>
- Freund, R. W., & Nachtigal, N. M. (1991). QMR: a quasi-minimal residual method for non-Hermitian linear systems. *Numer. Math.*, 60, 315-339. <http://dx.doi.org/10.1007/BF01385726>
- Liu, C. S. (2013a). An optimal tri-vector iterative algorithm for solving ill-posed linear inverse problems. *Inv. Prob. Sci. Eng.*, 21, 650-681. <http://dx.doi.org/10.1080/17415977.2012.717077>
- Liu, C. S. (2013b). A dynamical Tikhonov regularization for solving ill-posed linear algebraic systems. *Acta Appl. Math.*, 123, 285-307. <http://dx.doi.org/10.1007/s10440-012-9766-3>
- Liu, C. S. (2013c). An optimal multi-vector iterative algorithm in a Krylov subspace for solving the ill-posed linear inverse problems. *CMC Comput. Mater. Contin.*, 33, 175-198.
- Liu, C. S. (2014a). A globally optimal tri-vector method to solve an ill-posed linear system. *J. Comp. Appl. Math.*, 260, 18-35. <http://dx.doi.org/10.1016/j.cam.2013.09.017>
- Liu, C. S. (2014b). A doubly optimized solution of linear equations system expressed in an affine Krylov subspace. *J. Comp. Appl. Math.*, 260, 375-394. <http://dx.doi.org/10.1016/j.cam.2013.10.013>
- Saad, Y. (1981). Krylov subspace methods for solving large unsymmetric linear systems. *Math. Comput.*, 37, 105-126. <http://dx.doi.org/10.1090/S0025-5718-1981-0616364-6>

- Saad, Y. (2003). *Iterative Methods for Sparse Linear Systems* (2nd ed.). Pennsylvania: SIAM.
- Saad, Y., & Schultz, M. H. (1986). GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Stat. Comput.*, 7, 856-869. <http://dx.doi.org/10.1137/0907058>
- Simoncini, V., & Szyld, D. B. (2007). Recent computational developments in Krylov subspace methods for linear systems. *Numer. Linear Algebra Appl.*, 14, 1-59. <http://dx.doi.org/10.1002/nla.499>
- Todd, J. (1954). The condition of finite segments of the Hilbert matrix. In The Solution of Systems of Linear Equations and the Determination of Eigenvalues. In O. Taussky (Ed.), *Nat. Bur. of Standards Appl. Math. Series*, 39, 109-116.
- Trefethen, L. N., & Bau, III, D. (1997). *Numerical Linear Algebra*. Pennsylvania: SIAM.
- van Den Eshof, J., & Sleijpen, G. L. G. (2004). Inexact Krylov subspace methods for linear systems. *SIAM J. Matrix Anal. Appl.*, 26, 125-153. <http://dx.doi.org/10.1137/S0895479802403459>

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).