Motivations and realizations of Krylov subspace methods for large sparse linear systems

Zhong-Zhi Bai *

Qian Weichang College, Shanghai University, Shanghai 200436, PR China
State Key Laboratory of Scientific/Engineering Computing, Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, P.O. Box 2719, Beijing 100190, PR China

A R T I C L E   I N F O

A B S T R A C T

We briefly introduce typical and important direct and iterative methods for solving systems of linear equations, concretely describe their fundamental characteristics in viewpoints of both theory and applications, and clearly clarify the substantial differences among these methods. In particular, the motivations of searching the solution of a linear system in a Krylov subspace are described and the algorithmic realizations of the *generalized minimal residual* (**GMRES**) method are shown, and several classes of state-of-the-art algebraic pre-conditioners are briefly reviewed. All this is useful for correctly, deeply and completely understanding the application scopes, theoretical properties and numerical behaviors of these methods, and is also helpful in designing new methods for solving systems of linear equations.

Contents

1. Introduction

The system of linear equations

$$Ax = b, \quad \text{with } A \in \mathbb{R}^{n \times n} \text{ nonsingular and } x, b \in \mathbb{R}^n, \tag{1}$$

can be solved efficiently by either a direct or an iterative method [1–5]. Roughly speaking, the direct methods are based on the lower-upper triangular and the orthogonal-triangular factorizations, or in brief, the LU (or the Gaussian elimination) and

* Correspondence to: State Key Laboratory of Scientific/Engineering Computing, Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, P.O. Box 2719, Beijing 100190, PR China.
    E-mail address: bzz@lsec.cc.ac.cn.

the QR factorizations [5], and the iterative methods are based on the matrix splittings [1,2,5–7] and the Krylov subspaces [4,5,8,9]. These two classes of methods are principally different but computationally dependent. More specifically, an iterative method can be used to refine an approximate solution computed by a direct method, and a direct method can be employed to precondition an iterative method. In particular, in the Krylov subspace iteration methods the orthogonal basis of the subspace is often computed through a stable variant of the QR factorization [10]. Therefore, a technical and skillful combination of direct and iterative methods can produce fast, stable and accurate linear solvers for the linear system (1).

While numerical stability and computational complexity are main issues in theoretically analyzing the direct methods, and balanced scaling, proper pivoting and effective ordering are essential strategies in their practical implementations [3,11], a challenge in theoretical study of the Krylov subspace iteration methods is convergence analysis of the iteration sequences, and a major difficulty in practical usage of these methods is algorithmic construction of high-quality preconditioner and algebraic analysis of the corresponding preconditioned matrix [8,9,12].

In this paper, we will briefly review several typical and important direct and iterative methods that are economical and effective for solving the linear system (1). The motivations of searching the solution x_* of the linear system (1) in a Krylov subspace are described in detail, and the algorithmic realizations of GMRES [13] are shown deliberately. Also, we discuss convergence properties of the Krylov subspace iteration methods such as GMRES when the matrix A is symmetric and when it is nonsymmetric but diagonalizable, and review several classes of effective preconditioners such as those based on *incomplete LU* (**ILU**), *incomplete QR* (**IQR**), sparse approximate inverses, matrix splitting iterations, and algebraic multilevel and multigrid iteration techniques. All this could be useful for correctly, deeply and completely understanding the application scopes, theoretical properties and numerical behaviors of these methods, and should be helpful in designing new methods for solving systems of linear equations.

2. The direct methods

In the Gaussian elimination method, we successively operate the Gauss transforms one a time on the expanded matrix $[A \mid b]$, and finally obtain the target matrix $[I \mid x_*]$, that is, $[A \mid b] \rightarrow [I \mid x_*]$ in symbolic, where I is the identity matrix. This method can solve only one system a time, requiring approximately the storage $n^2 + n$ (n^2 for the coefficient matrix A and n for the right-hand side b) and the operations $\frac{2}{3}n^3$. The methodology of the LU factorization is a little bit different from the Gaussian elimination, which first factorizes the coefficient matrix A into the product of a lower-triangular matrix L and an upper-triangular matrix U, i.e., $A = LU$, and then computes the exact solution x_* of the linear system (1) through a forward elimination and a backward substitution. The LU factorization method can solve many linear systems having the same coefficient matrix A but different right-hand sides b a time.

For the QR factorization method, we first factorize the coefficient matrix A into a product of an orthogonal matrix Q and an upper-triangular matrix R, obtaining $A = QR$, and then compute the exact solution x_* of the linear system (1) through a backward substitution due to $Rx = Q^T b$. The QR factorization requires approximately the storage $n^2 + n$ and the operations $2n^3$. Here and in the sequel, we use $(\cdot)^T$ to indicate the transpose of either a vector or a matrix.

Besides the differences in storage and operation mentioned above, it has been proved that the LU factorization exists only for strictly diagonal dominant matrices and symmetric positive definite matrices, but the QR factorization may exist for any matrix even for a rectangular one. Hence, the QR factorization can be employed to solve the linear least-squares problems via, e.g., the seminormal equation $R^T Rx = A^T b$; see [14]. Moreover, if $A \in \mathbb{R}^{n \times n}$ is sparse, then both L and U may be also sparse, but Q and R could be dense. Hence, each of these two factorizations has its pros and cons.

As we have known, Givens rotation, Householder reflection and Gram–Schmidt orthogonalization are three classical and typical tools for computing a QR factorization for a given matrix. Below we review the classical Gram–Schmidt orthogonalization process and its stabilized modification, in which the latter is the elementary ingredient of the Krylov subspace iteration methods.

Let

$$A = [a_1, a_2, \ldots, a_n], \qquad Q = [q_1, q_2, \ldots, q_n]$$

and

$$R = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ & r_{22} & & \vdots \\ & & \ddots & \vdots \\ & & & r_{nn} \end{bmatrix},$$

where a_i and q_i are the ith columns of the matrices A and Q, respectively. Then $A = QR$ or

$$[a_1, a_2, \ldots, a_n] = [q_1, q_2, \ldots, q_n] \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ & r_{22} & & \vdots \\ & & \ddots & \vdots \\ & & & r_{nn} \end{bmatrix}$$