High-Performance Algorithms for Large-Scale Singular Value Problems and Big Data Applications

Lingfei Wu and Andreas Stathopoulos (Advisor)

Department of Computer Science
College of William and Mary
Williamsburg, Virginia, 23187
Email: lfwu, andreas@cs.wm.edu

I. Introduction

As "big data" has increasing influence on our daily life and research activities, it poses significant challenges on various research areas. Some applications often demand a fast solution of large, sparse singular value problems; In other applications, extracting knowledge from large-scale data requires many techniques such as statistical calculations, data mining, and high performance computing. In this dissertation, our objective is to develop efficient numerical methods and practical data mining techniques to cope with very large-scale problems on extremely large parallel machines.

An ubiquitous computational kernel in science and engineering is the singular value decomposition (SVD) of a matrix. The problem of computing the SVD can be formulated as an equivalent Hermitian eigenvalue problem. Many applications require a few of the largest singular values of a large, sparse matrix A and the associated left and right singular vectors (all together we call them singular triplets). These applications are from diverse areas, such as pattern recognition, social network analysis, image processing, textual database searching, and control theory. A smaller, but increasingly important set of applications require a few smallest singular triplets. Examples include least squares problems, determination of matrix rank, low rank approximation of matrix inverse [1], and computation of the pseudospectrum of an operator. The main focus of this research is to develop efficient and robust iterative methods for solving difficult large, sparse singular value problems.

In the era of big data, data mining techniques and high-performance computing are two key technologies of the big data analytics for extracting knowledge from vary large volumes of a wide variety of data. Furthermore, it is important to perform real-time analysis for providing immediate feedback to the control system in many safety critical applications. To accelerate processing of the huge data sets, high-performance computing, grid computing and in-memory analytics are essential to achieving the goal. In our research, we explore data mining techniques and high-performance computing to help understand the dynamics of fusion plasma in fusion experiments or numerical simulations by leveraging extremely large parallel machines.

The research we are pursuing in this dissertation has led to three important contributions:

- We propose a preconditioned two-stage SVD method that significantly advances the current state-of-the-art in singular value problem solving.
- We develop a high quality SVD software, PRIMME_SVDS, based on the state-of-the-art package PRIMME (PReconditioned Iterative MultiMethod Eigensolver). PRIMME_SVDS supports accurate

computation of both largest and smallest singular triplets, for both square and rectangular matrices, either with preconditioning or without, on a massively parallel machine.

- We propose a high-performance region outlier detection method for finding blob-filaments in real fusion experiments or numerical simulations by exploring an HPC cluster. To the best of our knowledge, this is the first work to achieve real-time blob-filaments detection in a few milliseconds.

II. A State-Of-The-Art Preconditioned SVD Solver Software

A. A Preconditioned Two-Stage SVD Method

We consider the problem of finding a small number of extreme singular values and corresponding left and right singular vectors of a large sparse matrix $A \in \Re^{m \times n}$ $(m \geq n)$,

$$Av_i = \sigma_i u_i, \qquad i = 1, \ldots, k, \quad k \ll n \tag{1}$$

The computation of the smallest singular triplets presents challenges both to the speed of convergence and to the accuracy of iterative methods. In this research, we compute extreme singular triplets of a large, sparse matrix. Especially, we focus on the most difficult problem of finding the smallest singular triplets.

There are two approaches to compute the singular triplets $\{\sigma_i, u_i, v_i\}$ by using a Hermitian eigensolver. Using MATLAB notation, the first approach seeks eigenpairs of the augmented matrix $B = [0\ A^T; A\ 0]$. Seeking eigenpairs of B computes the smallest singular values accurately. However, convergence of eigenvalue is very slow since it is a highly interior eigenvalue problem. The second approach computes eigenpairs of the normal equations matrix $C = A^T A$. This approach has been theoretically and practically proven to have better convergence compared to all other methods [2], but it suffers accuracy problem.

In [2], we proposed that accuracy and efficiency can be achieved through a hybrid, two-stage meta-method. In the first stage, the proposed method takes advantage of faster convergence on C up to the user required accuracy or up to the accuracy achievable by the normal equations. If further accuracy is required, the method switches to a second stage where it utilizes the eigenvectors and eigenvalues from C as initial guesses to a Jacobi-Davidson method on B, which has been enhanced by a refined projection method. The appropriate choices for tolerances, transitions, selection of target shifts, and initial guesses are handled automatically by the method. Our extensive numerical experiments show that our method can be considerably more efficient than existing state-of-the-art methods when computing a few of the smallest singular triplets, even without a preconditioner. With a good preconditioner, the proposed method can be much more efficient and more robust than current best methods. See reference [2] for more details.

B. A High-Performance SVD Solver Software

Given the above research activities in SVD algorithms, it is surprising that there is a lack of good quality software for computing the partial SVD, especially with preconditioning. Without preconditioning, SVDPACK and PROPACK implement variants of (block) Lanczos methods. In addition, PROPACK implements an implicitly restarted Lanczos bidiagonalization (LBD) method. However, SVDPACK can only compute largest singular triplets while PROPACK has to leverage shift-and-invert techniques to search for smallest. SLEPc offers some limited functionality for computing the partial SVD problem of a large, sparse rectangular matrix using various eigensolvers working on B or C. It also implements a parallel LBD method but focuses mainly on largest singular values. With the growing size and difficulty of real-world problems, there is a clear need for a high quality SVD solver software that allows for additional flexibility, implements state-of-the-art methods, and allows for preconditioning.

In this work we address this need by developing a high quality SVD software, PRIMME_SVDS, based on the state-of-the-art package PRIMME. This research presents our development in PRIMME in order to provide state-of-the-art robust, high performance SVD solver supporting accurate computation of both largest and smallest singular triplets, for both square and rectangular matrices, with either preconditioning or without. Our numerical experiments show that PRIMME_SVDS can be much more efficient than all other SVD solver software when computing a few of the largest or smallest singular triplets. In addition, we demonstrate the good scalability of PRIMME_SVDS for solving large-scale problems in various real applications under different parameter settings [3], [4].

C. Experiments and Results

For our performance test we consider seeking small number of smallest and largest SVD problems from various real applications including ill-conditioned least-squares problem, DNA electrophoresis, and graph partitioning and clustering. To test the weak scaling performance, we used laplacian matrices due to easy manipulation of matrix size. We perform our experiments on the NERSC's Edison and a cluster SciClone at college of William and Mary.

Figures in the result I of the poster reveal the comparison of different methods on SciClone. When seeking smallest singular values, PRIMME_SVDS is substantially faster than JD, Krylov-Shur, and TRLAN by a factor of 3, 10, and 60 respectively. For largest singular values, PRIMME_SVDS is still faster than other methods for seeking a few. When seeking many largest singular triplets, we expect Krylov method will be a winner due to its better global convergence.

Figures in the result II show the scalability performance of PRIMME_SVDS on Edison when seeking a small number of extreme singular triplets. PRIMME_SVDS can achieve near-ideal speedup until 256 processes on largest rectangular matrix in the Florida Sparse Matrix Collection. When reaching 512 processes, the communication cost becomes close to computation cost due to too small local matrix dimension in each process. For difficult smallest SVD problems, a good preconditioner is a necessity for solving real large application. The good speedup largely relies on the efficiency of a preconditioner and heavy sparsity of a matrix. Our results illustrate the same good scalability performance when seeking largest singular triplets without preconditioning. Finally, we also demonstrate its good performance under weak scaling when adjusting the nodes number of a 3D laplacian matrix from 80 to 200.

III. A Fast and Scalable Outlier Detection Method

To extract knowledge from the massive amounts of available data, data mining techniques are frequently used. Many traditional data mining techniques attempt to find patterns occurring frequently in the data, but in this work, we explore outlier detection approaches to discover patterns happening infrequently. Outlier detection is an important task in many safety critical environments since the outlier indicates abnormal running conditions from which significant performance degradation may well result. An outlier in these applications demands to be detected in real-time and a suitable feedback is provided to alarm the control system. Moreover, the size of ever increasing amounts of data sets dictates the needs for fast and scalable outlier detection methods. In this research, we apply the outlier detection techniques to effectively tackle the fusion blob detection problem on extremely large parallel machines. The blob-filaments are detected as outliers by constantly monitoring specific features of the experimental or simulation data and comparing the real-time data with these features.

Magnetic fusion could provide an inexhaustible, clean, and safe solution to the global energy needs. The success of magnetically-confined fusion reactors demands steady-state plasma confinement which is challenged by the blob-filaments driven by the edge turbulence. Real-time analysis can be used to monitor the progress of fusion experiments and prevent catastrophic events. However, terabytes of data are generated over short time periods in fusion experiments. Timely access to and analyzing this amount of data demands properly responding to extreme scale computing and big data challenges. In this paper, we apply outlier detection techniques to effectively tackle the fusion blob detection problem on extremely large parallel machines. We present a real-time region outlier detection algorithm to efficiently find blobs in fusion experiments and simulations. In addition, we propose an efficient scheme to track the movement of region outliers over time. We have implemented our algorithms with hybrid MPI/OpenMP and demonstrated the accuracy and efficiency of the proposed blob detection and tracking methods with a set of data from the XGC1 fusion simulation code. Our tests illustrate that we can achieve linear time speedup and complete blob detection in two or three milliseconds using Edison, a Cray XC30 system at NERSC [5].

IV. Conclusion and Future Work

We highlight our high-performance SVD solver as following:

- Among the fastest and most robust production level software for computing a small number of singular triplets
- Exploiting preconditioning for very large-scale problems
- Computes efficiently smallest singular triplets in full accuracy
- Demonstrates good scalability under both strong and weak scaling even for highly sparse matrices
- Free software at: https://github.com/primme/primme

We are planing to fully test our PRIMME_SVDS solver on different applications and conduct a comprehensive study on performance comparison between PRIMME_SVDS and other available SVD sovler software.

We summarize our work on blob-filaments detection below:

- For the first time, we propose a real-time blob-filaments detection approach
- The key idea is a two-phase region outlier detection scheme
- Demonstrates linear time speedup and completes blob detection in two or three milliseconds on Edison.

We are currently working on integrating our blob detection algorithm into the ICEE system for consuming fusion plasma data streams where the blob detection function is used in a central data analysis component and the resulting detection results are monitored and controlled from portable devices, such as an iPad. We plan to test the proposed method in both numerical simulations and real fusion experiments.

Acknowledgment

This work is supported by NSF under grants No. CCF 1218349 and ACI SI2-SSE 1440700, by DOE under a grant No. DE-FC02-12ER41890 and by DOE under Contract No. DE-AC02-05CH11231.

References

[1] L. Wu, A. Stathopoulos, J. Laeuchli, V. Kalantzis, and E. Gallopoulos, “Estimating the trace of the matrix inverse by interpolating from the diagonal of an approximate inverse,” *submitted to the Journal of Computational Physics*, preprint: http://arxiv.org/abs/1507.07227, July 2015.

[2] L. Wu and A. Stathopoulos, “A preconditioned hybrid SVD method for computing accurately singular triplets of large matrices,” *accepted for publication in SIAM J. Sci. Comput.*, preprint: http://arxiv.org/abs/1408.5535, July 2014.

[3] ——, “Enhancing the PRIMME eigensolver for computing accurately singular triplets of large matrices,” Department of Computer Science, College of William and Mary, Tech. Rep. WM-CS-2014-03, 2014.

[4] ——, “PRIMME_SVDS: A preconditioned SVD solver for computing accurately singular triplets of large matrices based on the PRIMME eigensolver,” Department of Computer Science, College of William and Mary, Tech. Rep. WM-CS-2014-06, 2014.

[5] L. Wu, K. Wu, A. Sim, M. Churchill, J. Y. Choi, A. Stathopoulos, C. Chang, and S. Klasky, “Towards real-time detection and tracking of blob-filaments in fusion plasma big data,” *submitted to the IEEE Transaction on Big Data*, preprint: http://arxiv.org/abs/1505.03532, May 2015.