PARALANCZOS GPU LIBRARY (VERSION 0.2)

BRIAN TUOMANEN

Abstract. This is a preliminary document describing the usage, implementation, and underlying theory of the Lanczos GPU Library. The author would appreciate any feedback on this document or the library; moreover, he is interested in possible collaboration and discussion with researchers in Compressive Sensing. He can be reached at `btuomanen@outlook.com` .

1. Introduction

The purpose of this software library is to perform fast massively parallel eigenvalue computations of large quantities of comparably small non-singular submatrices within a much larger positive-symmetric matrix in CUDA; notably, this library has been developed to assist in the study of Reduced Isometry Property (RIP) sensing matrices in the field of Compressive Sensing. To this end, several Matlab front-end examples are included, as well as standard command-line C examples.

As per the name, the main CUDA kernel function performs a Lanczos tridiagonalization on each submatrix, and then a QL decomposition to extract the eigenvalues of the resulting similar submatrix. The user indicates which submatrices of the larger matrix are to be analyzed by a parameter given to the kernel function, which is then run directly on the GPU.

Due to the extensive usage of CUDA "shuffle" commands to speed up communication between GPU threads and avoid race conditions, this library requires at least a Kepler-level NVidia GPU to run (2012-); note that ParaLanczos was written and tested on Maxwell-level (2014) devices. Compilation parameters and Matlab parameters might have to be tested and tweaked for some systems.

Note that the Matlab examples are reliant on the Mathworks Parallel Computing Toolbox to be installed; only Matlab 2015A has been tested. In the future, Python, R, and Octave front ends will also be included.

Also note this library has only been tested on Linux Mint 17.2, Linux Mint 18, and Windows 10. **Due to problems associated with the current NVidia drivers for Linux regarding the usage of the shuffle command in full-occupancy warps, this library is only fully supported under Windows. Some kernels in this library will not work under Linux!** Modifications may be duly needed for it to work properly on other operating systems.

2. Installation

Please ensure that you have CUDA 7.5 or later installed on your system, and the appropriate paths are set up before proceeding; in the case of Windows, compilation is performed under the Visual Studio command line by running the batch file "`make-windows.bat`", while under Linux, a `Makefile` is provided. (Compilation is done with the command "`make`".)

To compile the library for a default Maxwell level (or later) GPU, type `make`. To compile the library for a Kepler level GPU, modify the batch or makefile for the appropriate architecture.

A basic example showing usage for command-line C programming are given in the file `example1.cu` file; examples showing usage for Matlab are given in `MATLAB` directory.

3. Usage

The reader is thus encouraged to look at the library code (particularly `example1.cu`, `example2.cu` and `example3.cu`.) to infer usage from within C++, and in particular the Matlab wrapper functions specifically how the GPU kernel functions work. Thus, we only show usage in Matlab.

Currently, ParaLanczos only supports real, nonsingular submatrices of a given dyadic size (i.e., $2^x \times 2^x$), ranging from size 4×4 to size 128×128. Technical information regarding the use of these kernels follows in this section.

Currently, these are the two main kernels in the library. Both of these perform the same task: they are given a positive-symmetric matrix in the form of a pointer to a double (`* mat`), the size of the matrix (`int n`), a pointer to a string that corresponds to the indices denoting the submatrices whose eigendecompositions that will be made (`int * subIndex`) , the size of each individual subIndex (`int subn`) , and pointers to preallocated arrays for each of the outputers. (`double *Eigenvalues, double *determinants, double *errors.`). The eigenvalues for each submatrix given by `subIndex` is put accordingly in the `Eigenvalues` pointer, in the proper order.

The two main kernels are as follows.

- `POS_SYM_SUBMATRIX_KER` : This kernel can evaluage sub-matrices of dyadic (i.e., power-of-two) sizes ranging from 4 to 128. The launch parameters of the kernel should be "`<<< NumberOfSubmatricesToEvaluate,` log_2 `( SizeOfSubmatrix ) >>>`". (This kernel currently works under both Linux and Windows.) `example1.cu` shows the usage of this kernel.
- `POS_SYM_SUBMATRIX_KER2` : This kernel is much more efficient than the prior kernel, as it achieves full-warp occupancy; however, due to limitations on `__shared` memory in CUDA, only submatrices of the sizes 4 to 32 are supported. Additionally, due to this kernel achieving full warp occupancy, the block size must be of size 32, the length of `subIndex` must be divisible by 32. This kernel is launched with the parameter "`<<< subIndexLength`

`/ 32, 32 >>>`". `example2.cu` and `example3.cu` show the usage of this kernel. (This kernel currently only works under Windows.)

The ParaLanczos kernel works by performing eigendecompositions of a specified set of submatrices of a larger real matrix, stored in an array of doubles in the GPU memory.

3.1. **Matlab Wrappers Usage.** As stated, please make sure that you are using Matlab 2015A and have the Parallel Computing Library installed for proper usage. Note the outputs will all be -1 if there is an error. (i.e., the user tries to extract eigenvalues from a singular matrix or submatrix.)

To see a basic, concrete usage of the following functions, please see the Matlab example file `how2use.m`.

- `gpu_subsvd.m` : This function extracts the eigenvalues from one nonsingular dyadic sized submatrix of a given positive-semidefinite matrix G; the indices are given in the variable `subIndex`:
  `[Eigenvalues, Determinant, Error] = gpu_svd(G, subIndex)` . (This uses the first kernel as seen in the prior section.)
- `gpu_multi_subsvd.m` : This acts similarly to the prior function, only multiple submatrices can be analyzed; each index is stored in `subIndices`, one after another; and each submatrix must be of the same size. The size of the submatrices is given by minor, which should be a dyadic integer. The eigenvalues for each submatrix are then stored in `Eigenvalues`, each following another and corresponding to the matrices given in `subIndices`:
  `[Eigenvalues, Determinants, Errors] = gpu_multi_subsvd(G, subIndices, minor)` . (This also uses the first kernel.)
- `gpu_multi_subsvd2.m` : This is like the prior function, only it achieves full warp occupancy; each index is stored in `subIndices`, and **this must have length divisible by 32**. The size of the submatrices is given by minor, which should be a dyadic integer. The eigenvalues for each submatrix are then stored in `Eigenvalues`, each following another and corresponding to the matrices given in `subIndices`:
  `[Eigenvalues, Determinants, Errors] = gpu_multi_subsvd(G, subIndices, minor)` . **(This uses the second kernel, and only works within Windows.)**

Acknowledgments

This work was supported by the National Science Foundation. (NSF ATD 1321779) The author expresses his gratitude to Professor Michela Becchi of the University of Missouri Department of Electrical and Computer Engineering for her invaluable assistance and advice with CUDA and GPU programming, and also thanks Professor Stephen Montgomery-Smith of the Department of Mathematics for his feedback.

Department of Mathematics, University of Missouri, Columbia, MO 65211-4100
E-mail address: `btuomanen@outlook.com`