AMATH 483/583 Problem Set #4 (Part A) **Assigned:** 4/16/2017 **Due:** 4/24/2017

1 Preliminaries

1.1 C++ Classes

You are expected to have a surface level understanding of C++ classes for this assignment. If you are still unclear on C++ classes after lecture, there is a Panopto tutorial available on the "Annoucnements" tab of the course webpage.

1.2 Pre-Processing

Although not part of the C++ language per se, the pre-processor allows you to programmatically manipulate the text of the program itself. That is, you can make changes to the program text – in an automated fashion – before it is actually compiled. One example we touched on was in the use of assert () for defensive programming.

When assert () is enabled, it executes a check of the expression that is passed to it – that is, it executes a small amount of code. In performance critical applications, this can steal cycles as well as prevent certain compiler optimizations. For production (release) versions of a program, we need to be able to remove the assertions.

One way to remove the assertions, of course, would be to go to all of your files and either delete or comment out all of the calls to assert(). This is a anti-solution: during development we need to be able to switch between debug and release versions of the code. We need to be able to switch *all* of the calls to assert() on and off in one fell swoop each time. To support this, calls to assert() can be "turned off" by the pre-processor if the macro NDEBUG is defined.

The pre-processor works in the following basic way. It takes your program as input and provides program text as an output. How this output is produced is controlled by the pre-processors own macro language – which is also embedded in the program text. Some of the statements you have already seen in your programs, such as #include are pre-processor statements. In fact, all statements beginning with # are pre-processor language commands.

Beyond just filtering input and output, the pre-processor has some sophisticated features for text manipulation. However, we are just going to look at some basic capabilities for including or excluding specific parts of your program before being passed to the compiler.

The command you have already seen from the pre-processor is #include, which is used to pull the text of one file into another file. For example, if your file Vector.cpp includes the file Vector.hpp

```
#include "Vector.hpp"
int main() {
  return 0;
}
```

the text that is passed to the compiler is the concatenation of the two files. That is, what is compiled when you invoke the compiler on <code>Vector.cpp</code> is the complete text of <code>Vector.hpp</code> with <code>Vector.cpp</code>, with the text of <code>Vector.hpp</code> inserted at the location of the <code>#include 'Vector.hpp''</code>. If you would like to see all of the text that is passed to the compiler after pre-processing, you can add the option "-E" to your compilation command. Be careful with this though – if you have included anything from the standard library you will get an enormous amount of text back. The statements <code>#include <iostream></code> are not special – there is a file named <code>iostream</code> that is part of the standard library and its text is pulled in with that pre-processor directive.

Now, as we mentioned above, the pre-processor can be programmed by the user. To do this, we need to be able to do expected programming tasks, such as defining and testing variables and branching based on the tests. Since the pre-processing text and the program text are combined – and since the pre-processor manipulates the program text – it is important to distinguish between pre-processor commands and variables, and the program text and variables. The convention is to use <code>ALL_CAPS</code> for any pre-processor macros or variables.

Variables in the pre-processor are defined with the following syntax:

```
#define MY_VARIABLE sometext
```

This creates the macro (pre-processor variable) with the name MY_VARIABLE in the pre-processor namespace. Two things happen when a pre-processor macro is defined. First, whenever the pre-processor encounters the text MY_VARIABLE in the program text, it substitutes the defined text for that variable. In other words, the following transformation occurs.

```
#include <iostream>
using namespace std;

#define MY_GREETING "Hello World"
#define OP *

int main() {

   cout << MY_GREETING << endl;
   cout << "7 x 6 = " << 7 OP 6 << endl;
   return 0;
}

return 0;
}</pre>

using namespace std;

using namespace std;

int main() {

   cout << "Hello World" << endl;
   cout << "7 x 6 = " << 7 * 6 << endl;
   return 0;
}
</pre>
```

Note that the macros MY_GREETING and OP are replaced by *exactly* the text they are define to be. In this case, that even includes the quotation marks in MY_GREETING.

Branching in the pre-processor is controlled by a family of <code>#if</code> directives: <code>#if</code>, <code>#ifdef</code>, and <code>#ifndef</code>, which test the value of a compile-time expression, whether a macro is defined, or whether a macro is not defined, respectively. In response to evaluating an <code>#if</code> the pre-processor doesn't execute one branch of pre-processor code or another, rather it sends one stream of your program text or another to the compiler.

Defensive Programming One standard use of the branching capabiliites of pre-processor is to make sure that the text from any give header files is only placed once into the text stream to the compiler. This can easily happen when multiple headers include each other and/or include multiple headers from the standard library. In that case, even though the headers might be #included, we only insert their text once. The following is standard technique. For any header file, the following pre-processor directives are used (assume the header file is Matrix.hpp):

```
#ifndef MATRIX_HPP  // if the macro MATRIX_HPP is not defined, include the following text
#define MATRIX_HPP  // First, define the macro
```

```
#endif // The program text up to the matching #endif is what is included
```

You should make a habit of always protecting your header files in this way. As you might expect, there is an #else to go along with #if.

```
#include <iostream>
int main() {
```

```
#ifdef BAD_DAY
  std::cout << "Today is a bad day'' << std::endl;
#else
  std::cout << "Today is a good day" << std::endl;
#endif
  return 0;
}</pre>
```

In this example, the compiler will get the first branch of text if the macro BAD_DAY is defined, otherwise it will get the second branch. **NB**: With #ifdef, it does not matter what the value of the macro is. The test is only whether the macro exists or not. It is perfectly acceptable to #define a macro with no value.

In fact, when we want to disable assert (), we just need to #define the macro NDEBUG, we don't need to give it any particular value. But, since the pre-processor just processes your program text and sends it to the compiler, the NDEBUG macro must be defined before the #include <cassert> statement.

But this raises almost the same scalability issue we mentioned before. If we want to globally remove assert, we need to have the NDEBUG macro defined when processing our program files. One way to do this would be to edit each of the files and insert (or remove) #define NDEBUG in every one. This is impractical for all but the smallest of programs (and maybe even not then).

There is an essential feature of the C++ compiler that solves this problem, namely the -D option. This option passes a macro (with or without a defined value) to the prep-processor. In particular, you can pass NDEBUG to your programs this way (without ever having to change the program):

```
c++ -DNDEBUG main.cpp -o main.exe
You can give the macro a value by using =
c++ -DNDEBUG=1 main.cpp -o main.exe
but for NDEBUG the definition, not any value, is what turns on or turns off assert().
```

Automation In keeping with the course philosophy of "automate anything repetitive", the place to take advantage of incorporating (or not) NDEBUG definitions is in your Makefile. However, a problem with a familiar feature appears. Namely, if we want to globally turn on or turn off assert(), we have to make a sweep through the Makefile and add it or remove it from every production rule. If we truly want to be able to enable or disable assert() with the flip of a switch (as it were), we need to just be able to change it in *one place* but have the effect be global.

To programmatically control the automation that is introduced by make, the make program also has its own macro language. Fully using those capabilities can quickly turn into "Deep Magic" and we want to avoid that, but there are some basic features that enable the "edit once - change everywhere" behavior we are looking for.

In particular, make allows you to define macros that are expanded within the body of the Makefile. Two very common macros that you might use are to define which compiler you want to use, and to define which flags you want to pass to the compiler. For example, to make the programs dot5893 and vectorNorm we might have the following intermediate rules in the Makefile:

And one can easily imagine having many more production rules for any number of files. Now, suppose we wanted to introduce NDEBUG, or change the compiler we are using, or switch between release levels of C++? Manually, we would have to edit every line, doing it completely consistently, and not introducing any errors. Or, we can use a macro and do it once. Using a Makefile macro is straightforward. You define it with = and use it with \subsetential \subseteq \subseteq \text{.} If we used the standard approach in the above Makefile, it would look like this:

Now, if we wanted to add NDEBUG to disable assert (), we simply change the definition of CXXFLAGS:

```
CXXFLAGS = -Wall -g -std=c++11 -DNDEBUG
```

We can do the same with switching from debug mode to release mode

```
CXXFLAGS = -Wall -03 -std=c++11 -DNDEBUG
```

(Note the use of -03 rather than -g.) There are a number of conventions used in naming and use of Makefile macros - and a number are pre-defined for you. Those wanting to further leverage the automation capabilities of make are encouraged to consult the on-line references given on the course web site.

Just for Ninjas For the truly lazy (in a good way), you will quickly notice that there is *still* alot of repetition in the Makefile. All of the production rules have the same pattern

```
$(CXX) $(CXXFLAGS) -c <something>.cpp -o <something>.o
```

Compile "something.cpp" to create "something.o". In large programs with large Makefiles, keeping these all consistent could benefit from automation. To automate pattern-based production rules, make has some "magic" macros that pattern match for you.

A few things to note here before we get to our final Makefile. First, dependencies don't have to all be on one line. In the above, we express the dependencies for the object files in multiple lines, one dependency per line. The lines that then have production rules are executed when any of the dependencies are not met.

```
$(CXX) -c $(CXXFLAGS) $< -o $@
```

In this production rule, the macro \$< means the file that is the dependency – the .cpp file – and \$@ means the target.

There is one last thing to clean up here. We *still* have a repetetive pattern – all we did was substitute the magic macros in. But every production rule still has:

```
something.o : something.cpp
$(CXX) -c $(CXXFLAGS) $< -o $@</pre>
```

There is one more pattern matching mechanism that make can use – implicit rules. We express an implicit rule like this:

```
%.o : %.cpp
$(CXX) -c $(CXXFLAGS) $< -o $@</pre>
```

This basically the rule with "something" above – but we only need to write this rule *once* and it covers all cases where something.o depends on something.cpp. We still have the other dependencies (on headers) to account for – but this can also be automated – google for "makedepend".

At any rate, the Makefile we started with now looks like this:

This will handle *all* cases where something o depends on something cpp.

NB: Unless otherwise instructed, you do *not* have to use any of the advanced features of make in your assignments as long as your programs correctly build for the test script. However, keep in mind that these mechanisms were developed to save time and decrease mistakes while programming.

2 Warm Up

2.1 The make Command

To verify, and gain some familiarity with, the operation of Makefile and pre-processor macros, work through the following examples.

Create the following program and name it fail.cpp:

```
#include <iostream>
#include <cassert>

int main() {
   assert(1 == 0);
   std::cout << "Hello World" << std::endl;
   return 0;
}</pre>
```

Notice that the assertion (1 == 0) will always be false and so the assertion will fail. On failure, the program will abort, and will abort before it prints the hello message. Try compiling and running that program with no particular compiler options and verify that it aborts before printing the message.

Next compile the program with the -DNDEBUG option.

```
$ c++ -DNDEBUG fail.cpp
```

What happens when you run the resulting a.out?

Finally, create a test Makefile that looks like the following:

Preliminary question – when you type make, what gets created? What happens when you run that program? Finally, modify the above Makefile so that it compiles a program that ignores the assert(). Do that without changing the production rule (the compilation line after the dependency rule for fail).

2.2 Timing

As discussed in lecture – if we are going to achieve "high performance computing" we need to be able to measure performance. Performance is the ratio of how much work is done in how much time – and so we need to measure (or calculate) both to quantitatively characterize performance.

To measure time, in lecture we also introduced a Timer class (available on the "Schedule" tab of the course website).

```
class Timer {
private:
    typedef std::chrono::time_point<std::chrono::system_clock> time_t;

public:
    Timer() : startTime(), stopTime() {}

    time_t start() { return (startTime = std::chrono::system_clock::now()); }
    time_t stop() { return (stopTime = std::chrono::system_clock::now()); }
    double elapsed() { return
    std::chrono::duration_cast<std::chrono::milliseconds>(stopTime-startTime).count(); }

private:
    time_t startTime, stopTime;
};
```

To use this timer, you just need to #include "Timer.hpp". To start timing invoke the start() member, to stop the timer invoke the stop() member. The elapsed time between the start and stop is reported with elapsed().

To practice using the timer class, write and compile the following program.

```
#include <iostream>
using namespace std;
#include "Timer.hpp"
```

```
int main() {
  long loops = 1024L*1024L*1024L;

  Timer T;
  T.start();
  for (long i = 0; i < loops; ++i)
   ;
  T.stop();

  cout << loops << " loops took " << T.elapsed() << " milliseconds" << endl;
  return 0;
}</pre>
```

First, to get a baseline, compile it with no optimization at all. On my laptop, the 1G loops above took about 2 seconds. If your computer takes too long or too short, you can adjust the loop value (multiply it by 2 for example, or change one of the 1024 values into 512). What value does your computer give when timing this loop? How many milliseconds per loop? Note that the empty statement ";" in the loop body just means "do nothing."

Second, let's look at how much optimizing this program will help. Compile the same program as before, but this time use the -03 option. How much did your program speed up? Does this make sense? If you are unsure about the answer you are getting here, start a discussion on Piazza. Try to have this discussion sooner rather than later, as you will need some of the information gained for later in this assignment.

2.3 Abstraction Penalty and Efficiency

One question that arises as we continue to optimize, e.g., matrix multiply is: how much performance is available? The performance gains we saw in class were impressive, but are we doing well in an absolute sense? To flip the question around, and perhaps make it more specific: We are using a fairly deep set of abstractions to give ourselves notational convenience. That is rather than computing linear offsets from a pointer directly to access memory, we are invoking a member function of a class (recall <code>operator()()</code> is just a function. Then from that function we are invoking another function in the <code>vector<double></code> class <code>- operator[]()</code>. And there may even be more levels of indirection underneath that function. Calling a function involves a number of operations, saving return addresses on the stack, saving parameters on the stack, jumping to a new program location <code>-</code> and then unwinding all of that when the function call has completed. When we were analyzing matrix-matrix product in lecture, we were assuming that the inner loop just involved a small number of memory accesses and floating point operations. We didn't consider the cost we might pay for having all of those function calls <code>-</code> calls we could be making at every iteration of the multiply function. If we were making those <code>-</code> or doing anything extraneous <code>-</code> we would also be measuring those when timing the multiply function. And, obviously, we would be giving up performance. The performance loss due to the use of programing abstractions is called the <code>abstraction penalty</code>.

One can measure the difference between achieved performance vs maximum possible performance as a ratio – as *efficiency*. Efficiency is simply

Achieved performance Maximum performance

Let's write a short program to measure maximum performance – or at least measure performance without abstractions in the way.

```
long N = 1024L;
double a = 3.14, b = 3.14159, c = 0.0;
Timer T;
T.start();
for (long i = 0; i < N*N*N; ++i) {</pre>
```

```
c += a * b;
}
T.stop();
```

To save space, I am just including the timing portion of the program. In this loop we are multiplying two doubles and adding to doubles. And we are doing this N^3 times – exactly what we would do in a matrix-matrix multiply.

Time this loop with and without optimization. What happens? How can this be fixed? Again, this is a question I would like the class to discuss as a whole and arrive at a solution.

2.4 The Matrix Code

For this assignment you will be provided with a number of source code files, as well as a Makefile, in the Collect It Dropbox that you should put into a subdirectory ps4. The code for the Matrix class will be in Matrix.cpp and Matrix.hpp. Note that the files will contain most of the examples that we showed in lecture, but you need to use the Collect It version of Matrix.hpp instead of the webpage version. There will also be a driver program bench.cpp that when built (see the Makefile) will create an executable that will run performance tests for a variety of matrix sizes and algorithms (as specified on the command line). You should be able to deduce how it functions based on inspection of the code as well as simply building it and running it. Try some different sizes and different algorithms. In general you should keep the -O3 flag turned on or you will be waiting a long time for the programs to complete.

Note that in the Matrix.cpp file we have provided a function for filling a matrix with random numbers as well as a function for zeroing out a matrix.

NB: I recommend that you take a quick tour through the code and familiarize yourself with some of the contents.

3 Exercises

As a reminder, you are at risk for losing points for the following:

- Having using namespace std; in any header file that is included into other code
- Having warnings appear when your code is compiled with the -Wall flag

3.1 Abstraction Penalty

During lecture, I made the claim that implementation of the Matrix class with a linear array (as a vector<double> would be more efficient than an implementation as vector<vector<double> > . Let's verify if that is true or not.

As we discussed in lecture, all of the algorithms in the provided Matrix.cpp (see Warm Up section) use the *external* interface of the Matrix class. That means we can change the implementation in Matrix.hpp without changing any of the code in Matrix.cpp. **NB**: We don't have to change the code as written, but we do have to recompile it if the implementation changes.

To change the implementation of the Matrix class from a one-d to a two-d representation, we need to change a couple of things in the file Matrix.hpp. First, we have to change the implementation of arrayData to be a vector<vector<double> >. Once that is changed, we have to change all of the member functions that use it (since it has changed). That is, we have to write a new constructor and new operator() (). Note that the interfaces of these will not change, only how they access the internal element storage.

To compare the performance of one implementation to another, I want you to be able to switch from one representation to another, just my recompiling with a compiler flag. To do this, use the <code>#ifdef</code> technique above. You have a couple of options in how to approach this. You can either put the <code>#ifdef</code> around the entire <code>Matrix</code> class definition, or you can do it for the constructor and separately for the accessors and for the implementation. I would recommend the first way, but the choice is completely up to you.

The macro to select one implementation or the other must be called ALTMATRIX.

Deliverable A modified version of the Matrix.hpp file. When the matrix benchmarking code is compiled with the macro ALTMATRIX defined, it should compile the implementation based on vector<vector<double> > . Otherwise it should compile the version of the file given to you.

Example Switching between versions should look essentially like this:

```
#ifndef ALTMATRIX
   // our code
#else
   // your code
#endif // ALTMATRIX
```

The compilations would then be (without any of the other usual flags, which you would need in real life):

```
$ c++ -c Matrix.cpp -o Matrix.o
$ c++ -DALTMATRIX -c Matrix.cpp -o Matrix.o
```

This will respectively compile our version and your version.

Evaluation We will compile a test program similar to bench.cpp that uses your modified Matrix.hpp and the provided Matrix.cpp. We will verify performance and correctness, testing with a variety of matrix sizes, using matrix-matrix products. We will switch between implementations by defining (or not) the macro ALTMATRIX.

3.2 Matrix Vector

In the source code directory for this assignment you will also see there are the files <code>Vector.hpp</code> and <code>Vector.cpp</code>. You will also notice they have skeletons of a <code>Vector</code> class and supporting functions, along with comments about the portions of the code you need to fill in. For this part of the assignment, you are asked to complete their implementation.

Accessor Based on what we discussed about const and references and so forth relative to the accessors in the Matrix class, implement the necessary member functions for operator() () for the Vector class.

Multiply Given a matrix $A \in \mathbb{R}^{M \times N}$ and a vector $x \in \mathbb{R}^N$, we define the product of the matrix with the vector as

$$y_i = \sum_j A_{ij} x_j$$

The translation of that to a doubly-nested loop in code should be obvious (just one loop less than matrix-matrix). In the file amath583.cpp, complete the implementation of matvec() and operator*() (the analogy to mat-mat should make this straightforward). Your implementation of matrix-vector multiply should only use the external interface of Vector.

Finally, test your implementation for performance and correctness.

Deliverable Modified version of the Vector.hpp and amath583.cpp files. The Vector.hpp file should complete the implementation of the Vector class as described above. The amath583.cpp file should complete the matrix-vector multiplication operation as described above.

Evaluation We will compile a test program similar to bench.cpp that uses your modified Vector.hpp and amath583.cpp files, along with the provided Matrix.cpp and Vector.cpp. We will verify correctness of your implementations.

3.3 Tuning Matrix Vector (AMATH 583 Only)

As with matrix-matrix product, we would like to get better performance from matrix vector product than with just the naive algorithm. This exercise will investigate some potential mechanisms.

First, based on your matrix-vector function, create two matrix-vector functions matvec_inner and matvec_outer in Vector.cpp. The difference between these is that the first should iterate over the vector in the *inner* loop, while the second should iterate over the vector in the *outer* loop. One of these should be identical to your existing matvec in amath583.cpp. Make sure to include their prototypes in Vector.hpp.

Modify bench.cpp to time your matrix-vector functions with the command line arguments multMVinner and multMVouter. This will likely require adding new runBenchmark and benchmark functions with the following prototypes:

Next, create a function matvec_student in Vector.cpp, with the appropriate prototype in Vector.hpp. The only requirements on in it are that it produce a correct answer, that it not be identical to either matvec_inner or matvec_outer, and that it be faster than either matvec_inner or matvec_outer. You can provide conditions on this requirement in your written responses. Modify bench.cpp to time this new function with the command line argument multMVstudent.

Deliverables Modified versions of Vector.cpp, Vector.hpp, and bench.cpp as described above.

Evaluation We will compile both your modified bench.cpp and a test program similiar to your modified bench.cpp that use your modified Vector.hpp and Vector.cpp. We will verify performance and correctness, testing with a variety of matrix sizes, using matrix-vector products.

3.4 Written Exercises

Recall the code and your result from Section 2.3. The provided code needed to be modified to obtain a meaningful result for the Maximum Performance denominator in computing efficiency. Once you have successfully obtained a value for your computer, provide the following:

- 1. Description of what changes you made to the timing code
- 2. Explanation of why you made those changes
- 3. Clock rate of your computer
- 4. Max achieved floating point rate of your timed code
- 5. (AMATH 583 only) Under what circumstances, if any, is your matvec_student faster than matvec_inner and matvec_outer?

Place the above prompts/questions and your responses into a text file ex4.txt.

4 Turning in The Exercises

All your exercises for both part A and part B of PS4 will be submitted as one tarball file. As such, a testing python script will be provided with PS4 part B that tests all the exercises in the single tarball file.

5 Learning Outcomes

At the conclusion of week 3 students will be able to

- 1. Describe the interface and implementation of the Matrix class, including the operator() () member function.
- 2. Implement basic matrix-vector product and matrix-matrix product functions using the external interface of the Matrix class.
- 3. Include or exclude code in a source code file using statements from the #if family.
- 4. Explain the high-level functionality of the -03 compiler flag.
- 5. Derive the mapping from two Matrix indices to one std::vector<double>index.
- 6. Write a simple C++ class.
- 7. Correctly use initializer in the constructor for a C++ class.
- 8. Write a simple test harness for measuring performance of matrix-vector and matrix-matrix multiply routines.
- 9. Explain the hardware mechanisms that are leveraged by the hoisting, tiling, blocking, and copytranspose matrix-matrix optimizations.
- 10. Explain the difference between column-major and row-major ordering.
- 11. Derive the (basic) computational complexity of matrix-matrix product and matrix-vector product.